Lexicographic studies in medicine: Academic Word List for clinical case histories

Philippa Mungra and Tatiana Canziani
University of Rome La Sapienza1 & University of Palermo (Italy)
philippa.mungra@tiscali.it & tatiana.canziani@unipa.it

Abstract

Medical texts are often thought to pertain to a closed community, but how far the language used by that community overlaps with general academic lexis is unknown. We examined a corpus of clinical case histories using the software RANGE to characterise the lexis of clinical case histories quantitatively and present a wordlist for clinical medicine. Only 58% of the general academic wordlists are found in clinical texts but the 85% overlap with an important academic wordlist, the Coxhead 570, furnishes evidence for the academic nature of the lexis used for clinical cases in Medicine. There was little overlap (38%) between our clinical case wordlist and other wordlists of medical research articles but such overlap might constitute a core of medical lexis.

Keywords: lexis, Medical English, Medical Academic Word List (MAWL), clinical case histories, subject-specific wordlists.

Resumen

Un estudio lexicográfico del inglés en medicina: El listado de términos académicos para los casos clínicos

Se suele pensar que los textos de medicina son principalmente territorio de una comunidad restringida de expertos del sector, pero se conoce poco sobre cómo y en qué medida el lenguaje utilizado por la comunidad médica puede coincidir con el léxico académico general. Con el objetivo de presentar un listado de términos específicos de los casos clínicos, estudiando el léxico desde un punto de vista cuantitativo, se ha examinado un corpus de casos clínicos utilizando el software RANGE. Según el estudio realizado, los resultados han revelado la existencia de una coincidencia del 85% con un importante listado de términos académicos (el denominado Coxhead 570), y una coincidencia del 58% con el

listado de términos generales. Este hallazgo confirma la naturaleza académica del léxico que se recoge en los casos clínicos. Si bien la coincidencia entre el listado de términos de los casos clínicos y de los listados restantes de artículos científicos en el ámbito médico sea baja (38%), es posible concluir, a nuestro juicio, que esa doble coincidencia conforma, en realidad, un léxico esencial o común en medicina.

Palabras clave: léxico, inglés médico, listado académico de términos médicos, casos clínicos, listados de términos específicos.

1. Introduction

Medical texts are often thought to pertain to a closed community, but how far the language used by that community overlaps with general academic lexis is unknown. In a fascinating account of the development of the medical journals, Booth (1982) affirms that medical learning up to the turn of the last century was almost entirely based on the reading of patient cases, known variously as cases, case reports or case histories. Few studies have addressed clinical case histories. One of the reasons for this might be that the macro-structure of case histories has not changed radically over many years from its beginnings in the *Proceedings of the Royal Society* in the 1700s. The evolution of the case report has been studied in terms of structure, content, stylistic features and variations in a diachronic study of medical writing between 1735 and 1985 (Atkinson, 1992) and through vernacularisation of medical treatises between 1850-1900 and 1965-1995 (Taavitsainen & Pahta, 2000).

In recent years however, Medical Education has centred heavily on clinical research to find valid justification for clinical questions and decision-making, as part of a school of thought known as Evidence-Based Medicine (EBM). EBM, though important methodologically and didactically, has privileged the study of medical research articles or RAs. Thus in Linguistics, there is a plethora of studies of RAs such as those identifying moves (Nwogu, 1997), moves in introductions (Hyland, 2000; Lorés, 2004; Samraj, 2005), in the methods section (Hopkins & Dudley-Evans, 1988; Berkenkotter & Huckin, 1999) in the discussion (Skelton & Edwards, 2000; Peacock, 2002), in abstracts like the IMRaD (Introduction, Methods, Results and Discussion) model (Swales, 1990; Paltridge, 1997), in biochemistry RAs (Kanoksilapatham, 2005) and in the evolution of the genre of medical RAs (Swales, 1990; Paltridge, 1997). Other medical genres studied are popularizations (Myers, 2003) dissertations and other academic papers

(Hyon & Chen, 2004) and Consensus Conference Statements (Mungra, 2007). In RAs and RA abstracts, pragmatic features such as hedges, (Salager-Meyer, 1994), modality (Salager-Meyer, Defives & Hamelinsck, 1998), scholarly gratitude, acknowledgements and scholarly criticisms (Salager-Meyer, Alcaraz Ariza & Zambrano, 2003; Salager-Meyer, Alcaraz-Ariza & Pabón-Berbesí, 2009; Salager-Meyer et al., 2011) have been studied.

There have been fewer linguistic studies of case reports. The macro-structure of case histories (DeBakey & DeBakey, 1984) and content checklists (McCarthy & Reilly, 2000; Cohen, 2006) have been published by medical journals and editors. Anspach (1988) identified four rhetorical devices of case presentations - depersonalisation, passivation or omission of the agent, use of medical technology as an agent and verbal account markers such as "states", "report" or "denies". The depersonalized tone with passive construction is now considered a conventional characteristic of modern case reports (Taavitsainen & Pahta, 2000). Hunter (1996) compared cases to a subjective, personalised narrative. According to Donnelly (1997), when such narratives are generalized, they become a teaching-learning tool because their predictive and diagnostic power gives importance to the genre. Berkenkotter (2008 & 2009) analysed the linguistic and historical evolution of psychiatric case reports emphasizing its role as an invaluable tool in diagnosis of psychiatric disease.

1.1. Problems when teaching the medical literature and NNS

Because of a return to a teaching methodology based on a clinical scenario or patient case called Problem-Based Learning or PBL, introduced at McMaster University (Barrows, 1990; Frederiksen, 1999), reading case reports have become increasingly important because it is student-centred and uses the analysis patient data by expert physician-teachers to lead students to the discovery/identification of the underlying pathology and the development of problem-solving skills (Schmidt, 1993). Clinical case histories furnish ample teaching material for physician-training (McCarthy & Reilly, 2000; Taavitsainen & Pahta, 2000; Cohen, 2006), and are widely read by medical students and by both junior and senior professionals since their purpose is instructive and their goal is not explanation but understanding for diagnosis (Taavitsainen & Pahta, 2000; Vandenbroucke, 2001; Berkenkotter, 2008) and to give clinicians "better insight into the unusual riddles which specialists usually encounter in their everyday practice" (Yitschaky, Yitschaky & Zadik, 2011: 180).

For non-native speakers of English such as our case, the lack of language skills may create a barrier in the patient-doctor setting or in problem-based group activity among peer-physicians (Mpofu et al., 1998; Dyke, Jamrozik & Plant, 2001; Koh et al., 2008).

For strategies to help professionals prepare publications of clinical cases, there are published guidelines (DeBakey & DeBakey, 1984) and content checklists (McCarthy & Reilly, 2000; Cohen, 2006). For RAs too, a consistent body of literature devoted to identifying strategies exists (Belcher, 1995; Medical Journals Editing, 2002; Okamura, 2006; Divasson-Cilvetti & León-Pérez, 2006; Belcher, 2007). One of the strategies that have received attention in recent years is that of wordlists for vocabulary development based on data from specialized corpora. To the best of our knowledge, no wordlists based on lexis have been used in clinical cases.

1.2. Wordlists: general and specialist wordlists

One of the earliest wordlists is that of the General Service List (GSL) by West (1953), corresponding to the most basic level of fluency in general English and consists of roughly the first most frequently used 2,000 words in English, acquired in primary school. Later, Campion and Elley (1971) compiled a list of the 500 most common words and 3,200 most frequently used words. Lexis used in an academic setting was first studied by Praninskas (1972) who compiled the American University Wordlist, based on a corpus of 272,466 words from ten university-level textbooks covering ten academic disciplines. Later, Lynn (1973) and Ghadessy (1979) worked with wordlists most frequently underlined by students who are speakers of other languages, while Xue & Nation (1984) published the University Word List (UWL) containing the 800 words most frequently used in the Humanities. Another general wordlist was compiled from written works in English called the British National Corpus or BNC. Subsequently, Coxhead (2000) developed the Academic Word List (AWL), using a corpus of 3.5 million running words from a variety of academic texts. The AWL consists of 570 word families which cover roughly 10% of running words and is now perhaps the most common reference academic wordlist (Coxhead, 2011).

Among specialist domains, wordlists for different purposes – either pedagogical or linguistic – have been built in the fields of Computer Science (Lam, 2001) and in Business (Hsu, 2009). In the Engineering field, Mudraya (2006) was the first to build the SEEC (Student Engineering English

Corpus), a corpus-driven lexical wordlist. Later, Ward (2009) created the BEL (Basic Engineering List) based on a range of topics in various major engineering fields and taken from 3rd and 4th year undergraduate textbooks. The aim was to create a wordlist that could be used in all engineering disciplines by learners with a low level of English. In Agriculture, Martínez, Beck & Panza (2009) created the Agrocorpus, a wordlist based on research articles in the agricultural sciences focusing the importance of creating wordlists based on pragmatic criteria more than on frequency. Two linguistic studies of the lexis of medical RAs have been published (Chen & Ge, 2007; Wang, Liang & Ge, 2008) providing a wordlist of the most frequently used medical academic words in this genre known as the Medical Academic Word List or MAWL but to the best of our knowledge, the lexis of case histories had not been studied despite the fact that they are important in developing and honing clinical decision-making skills (Skinner, 1956; Moran-Campbell, 1976; Barrows, 1990). Given the paucity of lexical studies of case histories, we present here a study of the lexis commonly used by medical and surgical case histories in order to:

1. Determine what differences, if any, there are between the lexis of case histories and other academic publications.
2. Compare our findings of the lexis of case histories with that of other published wordlists in medicine.

2. Procedure

2.1. Collection of the corpus

From URL: http://sciencedirect.com, a public science database, we randomly retrieved 3 journals from each of the 24 medical and surgical topics as seen in Figure 1, and we selected randomly two or three (when available) case histories from each of these 72 journals. All the sample case reports included in our corpus were kept at their original length, and published between the years 1997 and 2011. We used the strict criteria by Wood (2001): published by native speakers (NSs) only according to surname or with affiliations to North American, British, Canadian, or Australian hospitals and universities, and in English-language European journals. The resulting corpus consisted of 200 case histories. The total number of running words was 246,907 with a range of 151 to 2,883 words per case. The

average number of running words in the case histories of the corpus was 1,235.

1.	Anesthesiology and Pain Medicine	13.	Oncology
2.	Cardiology and Cardiovascular Medicine	14.	Ophthalmology
3.	Critical Care and Intensive Care Medicine	15.	Orthopedics, Sports Medicine and Rehabilitation
4.	Emergency Medicine	16.	Otorhinolaryngology and Facial Plastic Surgery
5.	Endocrinology, Diabetes and Metabolism	17.	Pathology and Medical Technology
6.	Forensic Medicine	18.	Perinatology, Pediatrics and Child Health
7.	Gastroenterology	19.	Psychiatry and Mental Health
8.	Health Informatics	20.	Pulmonary and Respiratory Medicine
9.	Infectious Diseases	21.	Radiology and Imaging
10.	Medicine and Dentistry (General)	22.	Surgery
11.	Nephrology	23.	Transplantation
12.	Obstetrics, Gynecology and Women's Health	24.	Urology

Figure 1. List of medical and surgical disciplines having clinical case histories.

The publications selected were identified as "cases" or "case reports" by the journal itself and all of them had the following elements/moves:

- abstract/introduction/presentation;
- diagnostic procedure;
- management of the patient and outcome;
- discussion/conclusion.

The files were saved in electronic form, exported and cleaned for unreadable characters, and then saved as text files (extension .txt). See Figure 2, which shows a preview of a case containing the elements/moves circled.

2.2. Text normalisation and use of the software

In the automatic creation of a text file from a pdf file, transcription mistakes or misspellings were corrected manually and corrections for names of drugs, abbreviations, symbols or units, were done in order to make the text readable by the Software used called RANGE, retrieved from Heatley, Nation & Coxhead (2002). As reference programs, we used three wordlists, namely, two wordlists from the GSL by West (1953), given as the most commonly used 2,000 words in English (also known as GSL-1 & GSL-2) and the third is the Academic Word List (AWL), consisting of 570 families compiled by Coxhead (2000). In addition to these three reference wordlists, we added the Medical Academic WordList (MAWL) of families published by Wang, Liang

and Ge, (2008). By using these four reference wordlists against our corpus of 200 clinical cases, we could identify the 2,000 most common words, the most commonly used words in academic domains and the most commonly used words in medical RAs. Any words outside these four lists would be considered peculiar to clinical cases, and thus a candidate for our clinical wordlist, which we called MAWLcc (Medical Academic Word List for Clinical Cases). With an occurrence of over 30 times per 200,000 tokens, this would constitute the MAWLcc wordlist, as indicated by the criteria in section 2.3 below.

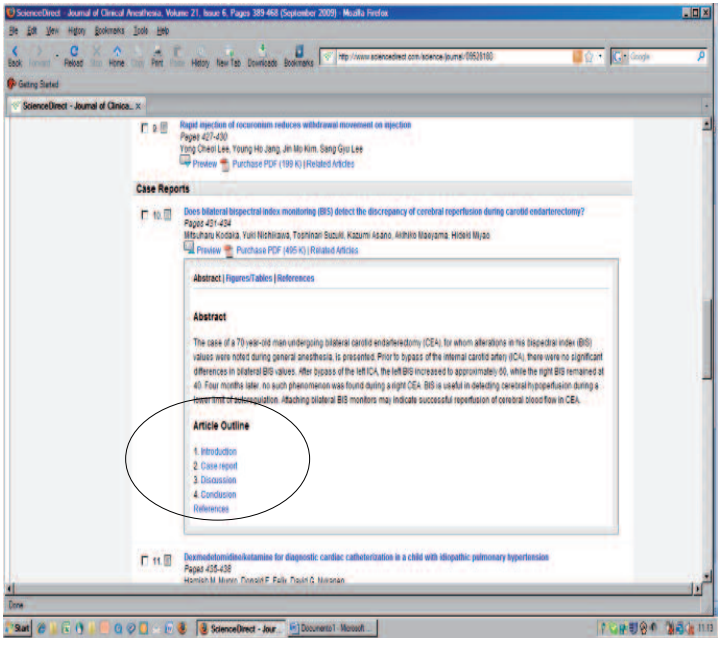


Figure 2. Article outline of a clinical case history in the Table of Contents from *Journal of Clinical Anesthesia* (2009), 21(6): 431-434. The four parts of the classic clinical case history are encircled.

Texts were normalised whenever possible, according to standards common in spoken clinical science for units, symbols and names of drugs or equipment. A complete list of corrections and normalisations may be found in Appendix 1. Cleaning and scoring of the texts were done by both researchers which might have created text differences but internal agreement was monitored at the start [case 10], midway [case 103] and at the end [cases 137, 184 & 188] of the entire scoring process. The Pearson coefficient was

found to show a strong positive correlation at r = 0.99 and the Student *t*-test was 0.013 ($p < 0.95$, 14 df).

Included with the software was a stoplist which included scientific notation and units ignored by the software which allowed for the possibility to ignore technical words such as the names of drugs placed within square brackets. The program returned data as a percentage of the tokens in each of the reference wordlists. It also returned lists consisting of both tokens/types and word families as well as a final list of tokens not present in any of the reference lists. A family of words (Bauer & Nation, 1993) consisted of the baseword plus its inflected version. Thus a word family would be counted as a single item or baseword such as "surgery" together with all the occurrences of the inflected versions of the baseword, for instance, "surgical", "surgeon", "surgeries", "surgeons" and "surgically". We extracted only the basewords to create our list of families.

2.3. Criteria for creation of the list

The principles used were the same as those by Wang, Liang and Ge (2008) for the creation of a MAWL from research articles (RAs). Our final list consisted of families having the following criteria:

a) Specialized occurrence: we collected all the word families outside the first 2,000 most frequently occurring in the West (1953);

b) The baseword families must have a range of at least 50% of all the medical fields, that is, 12 out of the 24 areas shown in Figure 1;

c) A frequency of at least 30 occurrences.

We then manually examined our list and eliminated all technical words such as the names of drugs, eponyms and procedures, which would be familiar to physicians. Our final list, MAWLcc was then compared to the MAWL from RAs published by Wang, Liang and Ge (2008) to compare overlap and ranking of the two MAWLs.

3. Results

3.1. Corpus of clinical case histories – characteristics

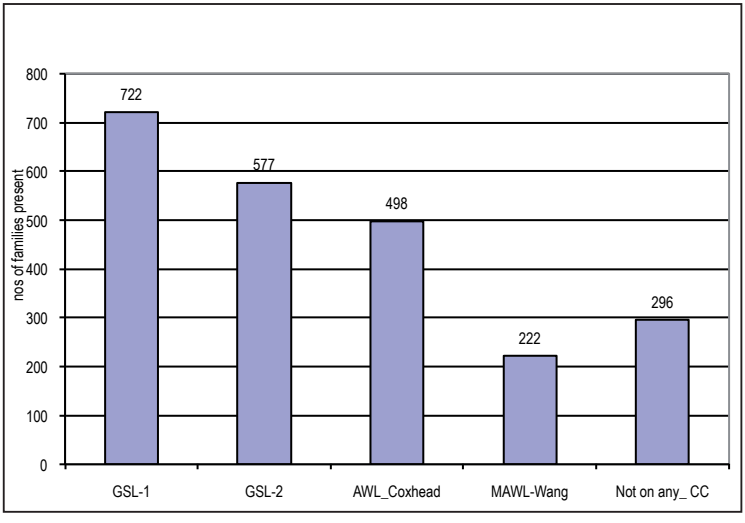
The types of texts included not only single patients but also groups of patients which resulted in a mixed text type having the characteristics of a case with statistical features of research articles in the results and discussion section [case 9] [case 14] [case 79]. In selecting cases for this corpus, we found that surgical disciplines often contain lexis describing procedures or describing reasons for an operation and that such lexis often demonstrated a metaphoric use of nouns and verbs as highlighted in italics in example (1):

> (1) An *interposition graft of reversed* <*saphenous*> vein was used for vascular reconstruction, and the <thrombus> was sent for culture. The external <carotid> artery stump was oversewn. There was a strong pulse and <Doppler> signal in the distal internal <carotid> upon completion of the repair, and *the wound was closed over a* <*Jackson-Pratt*> *drain.* The cultures were ultimately negative, and the patient was neurologically intact. <PTA> is commonly used for peripheral vascular atherosclerotic disease and was first applied to <FMD> by <Yoshida et al> on the theory that the smooth <neointima> created after *dilation could relieve the corrugation* of the vessel lumen in <FMD>. The *decorrugation* of the vessel then reduces the turbulent flow within the vessel, halting the *showering of* <*microthrombi*> to the distal *arterial tree.*

Although this is a small corpus, we believe it is fully representative of case reports in that it was culled from over 50% of medical fields of the corpus. According to Sinclair (2004), representativeness requires that typically 40-50% of all word types occur only once in a given corpus. In our corpus, 55.4% of the types occurred at least once.

3.2. Wordlists retrieved after analysis by the software – MAWLcc

Following the workings of the software, we retrieved our medical wordlist, MAWLcc using four wordlists as a reference. These are the 2 GSL wordlists, the Coxhead AWL and finally the MAWL from RAs published by Wang, Liang and Ge (2008). The data are plotted in Figure 3 and show the occurrence of each reference list as a percentage; the frequencies of families are given in Table 1, which is the raw software output.



Baseword List	Total no. of families	% of total running words in the corpus
GSL-1	722	18
GSL-2	577	10.6
AWL	498	13.6
MAWL	222	2.1
Not in the lists therefore peculiar to clinical cases	296	

Figure 3. Plot of numbers of families in medical case histories using the GSL-1 and 2, Coxhead-570 and the MAWL published by Wang, Liang and Ge (2008) as reference wordlists.

Wordlist	Tokens / %	Types / %	Families
GSL-1	60,645 / 38.77	1,893 / 18.01	722
GSL-2	19,453 / 12.44	1,115 / 10.61	577
AWL	22,547 / 14.41	1,425 / 13.55	498
MAWL	9,862 / 6.30	222 / 2.11	222
Not in the lists	43,923 / 28.08	5,858 / 55.72	?????
Total	156,430	10,513	2,019

Table 1. RANGE output (raw data) from 200 case histories.

In the raw data (see Table 1) we retrieved 2,019 families recognized by the reference lists. The string "?????" given by the RANGE software refers to the fact that these words in the corpus were not present in any of the reference wordlists. With the application of the first criterion (elimination of the GSL-1 and GSL-2), 720 word families remained. The text coverage of the AWL word families in our clinical corpus was 13.55% (rounded to 13.6% in Figure 3), somewhat higher than the 10.1%, of the AWL words in medical RAs texts reported by Chen & Ge (2007).

We found that 498 families of Coxhead-570 list of families (corresponding to 85%), were present in our corpus and may be considered strong evidence for the academic nature of our corpus of clinical cases. By way of contrast, Chen & Ge (2007) found that only 292 families were present in medical RAs which corresponded to 51.2% of the Coxhead-570.

In our corpus, 222 families were found in common with the 623 word families published by Wang, Liang and Ge (2008), but that accounted for only 2% of the running corpus. This might be interpreted as a result of the clinical rather than research nature of our corpus.

When we applied the last two criteria to the remaining 296 families, only 241 families were found to occur with a frequency of >30 times in the corpus. A complete list of MAWLcc containing these 241 word families can be found in Appendix 2, together with the frequencies of occurrence in our corpus. We identified word families among the tokens returned by the software and created families as seen in Table 2, with a breakdown of the individual elements of the families and their frequencies.

Headword = Cell	No. of occurrences	Headword = Clinic	No. of occurrences
Cell-mediated	10	Clinical	445
Cells	203	Clinically	70
Cell	208	Clinician	4
Cellular	22	Clinicians	29
Cellularity	10	Clinico	20
		Clinicopathological	17
		Clinics	20
Total	453	Total	605

Table 2. Two examples of converting tokens into families and breakdown of occurrences in the corpus.

In our list of clinical cases, the most frequent word was "patient" which occurred 1,155 times in the corpus, and was registered in all 24 clinical areas, followed by the word "diagnose", both concerning the sick person. The least frequent words were "anatomy", "bone", "dysplasia", "ingestion" and "ureter", used 30 times in the corpus and present in only 18 clinical areas.

3.3. Comparison of our MAWL families with other studies

It is difficult to compare our data with that of other corpora because our family wordlists were compiled using the GSL with 2-wordlists or the Coxhead-570 AWL, compiled from academic texts written in general English

classes. Almost all the word families in GSL-1 and GSL-2 and 498 out of the 570 word-families from Coxhead's AWL were present in our corpus.

Table 3 shows a comparison with a selection of the most common corpora in the hard sciences. If we remove GSL-1 and GSL-2, corresponding to the first 2,000 most frequent words, our resulting list of 395 words shared 153 families in common with the wordlist published by Wang, Liang and Ge (2008), corresponding to 38.25%. This shared list may be considered a core wordlist for medicine, in that they have been found in both clinical cases and medical research articles. The remaining 241 families of our list (see Appendix 2) might be considered typical of clinical cases.

In a comparison between other hard sciences, such as Engineering or Agriculture, and with our MAWLcc there was very little overlap, as seen in Table 3, which supports the highly specific nature of specialist corpora (Hyland & Tse, 2007).

		Compiled from	Wordlist families	Overlap with our $MAWL_{CC}$
Academic Wordlists				
	Coxhead (2000) AWL	From written academic texts	570	85%
Specialist Wordlists				
1	Chen & Ge (2007)	Medical RAs (used Coxhead's 570 AWL as a reference)	292	No comparison is possible since wordlists were not published
2	Wang, Liang and Ge (2008) (Medicine) MAWL	Medical RAs (used GSL as reference)	623	38.25 %
3	Mudraya (2006) (Engineering) SEEC	Textbooks (BNC as reference)	1,200	1.02%
4	Martínez, Beck & Panza (2009) (Agriculture) AgroCorpus	RAs / Coxhead's AWL	92	2.9%
5	Ward (2009) (Engineering Corpus) BEL	Textbooks, computing work token frequency	299 word types (not families)	No comparison is possible since types were used

Table 3. Comparison with other lexical corpora in the hard sciences to show overlap with MAWLcc.

4. Discussion

4.1. Validity of the wordlist

With respect to the collection of the corpus, we had some difficulty in selecting the publications because of paucity of linguistic publications identifying the rhetorical moves inherent in clinical case histories, unlike

RAs, which have the IMRaD structure (Swales, 1990). Because of this, we decided to use the basic structure given to those publications in medical journals characterised as "cases" (McCarthy & Reilly, 2000; Cohen, 2006) having the characteristics indicated in Figure 2, namely, abstract, introduction, description of the case, discussion. Secondly, the presence of mixed subgenres of case series and also of cases from surgical disciplines, may have affected our MAWLcc because of language used in an unorthodox or metaphoric manner as seen in the words in Example 1 in the results. Thirdly, for reasons of comprehension and clarity, we were forced to make a selection of clinical cases prepared by English native-speaking physicians. The reason for this can be explained by examining a text like the following example (2), written by a Brazilian researcher and published in an international journal after peer review. To the casual reader or even a competent medical student, the inter-sentencial reasoning is not clear:

> (2) When antibiotics are really needed, their efficacy may be enhanced by the immune system, but in the case of advanced, serious infections in immunocompromised patients, the risk of therapeutic failure is much greater. T*he first scenario conceals antibiotic failures*, the second ends in severe complications or death but does not call the attention of clinicians because it is the expected outcome.

In the highlighted sentence, it is not clear what the first scenario is. In fact, the author probably means that efficacy may be due not to the antibiotic, but rather to a very active immune system. Therefore, one can prescribe a generic poorly-produced antibiotic, with poor efficacy and it will still work. The second scenario is self-explanatory. A text of this sort is difficult to decodify for two reasons: first, because of the misleading use of the word "scenario" and secondly, because of the compressed contextual implication, that requires a pragmatic referential jump. In linguistic terms, the absence of "well-formedness" (Giora, 1997) and "relevance" (Sperber & Wilson, 1986) make such texts difficult to understand, thus requiring the maximum processing effort. Because of the high frequency of opacity in cryptic texts like this, written by native speakers of other languages, we agreed to select cases according to the conditions of Wood (2001). Despite these difficulties, we believe that our corpus of case histories, though small, is representative of case histories in most of the medical disciplines since the selection criteria follow the canons of Sinclair (2004) regarding representativeness and balance.

In a comparison with MAWL in Wang, Liang and Ge (2008) based on research articles frequencies and ranking, our MAWLcc is more characteristic of patient care rather than scientific research. This is clearly reflected in the rankings of specific lexical items: we found the verb "to analyze" ranked as #242 as compared to #6 in that of MAWL. Similarly, the word "symptom", #2 on our list, appeared at #81 in MAWL. Both corpora were culled from medical texts, but belong to different subgenres. In a comparison of these two subgenres, we observed an overlap of 250 lexical families, corresponding to 28.8% of our list of MAWLcc families. Data such as this reflect the highly specific nature of both corpora and furnishes some support for the idea put forward by Hyland and Tse (2007), and supported by Granger and Paquot (2009) that a core of academic lexis is difficult if not impossible to define. Although Hyland and Tse (2007) came to these conclusions by studying mixed corpora including professional and learner texts as well as dissertations and undergraduate theses and Granger & Paquot (2009) assembled their corpus from non-native English student writings, their ideas regarding the specific nature of academic lexis are essentially true; namely, that academic genres in the hard sciences and professional academe are highly specific in nature. Nevertheless, we believe that within a single professional field such as Medicine, an essential or core lexis may exist such as the 153 lexical families we observed in both our wordlist and that of Wang, Liang and Ge (2008).

This comparison is an attempt to answer our second query regarding whether the lexis of case histories is very different and how different it is when compared with other published wordlists in medicine. Here we observe that our data validates the academic nature of AWL-570 by Coxhead and that our corpus and the MAWLcc might be considered a complementary wordlist to that of Wang, Liang and Ge (2008). Although other research groups have built corpora of academic English, like MICASE, BAWE or the Louvain corpus, most of these corpora consist of writings by students of varying proficiencies and not professional peer-reviewed publications (Alsop & Nesi, 2009).

4.2. Pedagogical considerations

This wordlist presents important lexis used in clinical publications regarding patient care. This list could form a base for specific lexical exercises like word-building since they are field-specific and if implemented within the

medical curriculum it would help students to acquire appropriate professional language. This means that students must already have studied basic English and should be familiar with most of the lexis appearing in GSL-1 and GSL-2. The remaining field-specific wordlists – the Coxhead-570 AWL and the Medical wordlist from RAs (Wang, Liang & Ge 2008) and that of the present study – could be implemented in later years of the medical curriculum. Word-building exercises would help students expand their vocabulary, aid in reading with precision, so important in the hard sciences, and contribute to developing the professional *persona*. The present wordlist may also be useful in testing for language proficiency, not only for medical students but also for international medical trainees.

5. Conclusions for further research

Although we have gathered a small to mid-sized corpus, this study is by no means exhaustive. There are several areas for improvement such as: identifying polysemy, collocational strings or formulaic clusters and other linguistic comparisons such as that done by Méndez-Cendón (2009) in the field of phraseology. Besides this, it would be interesting to compare this wordlist with other concept-oriented glossaries and lexicographic collections such as MeSH words and UMLS (from the National Library of Medicine) and the wordbuilder DEFINDER (Muresan & Klavans, 2002). Although ours is a preliminary work, its main value lies in the didactic application of wordlists for both teaching and testing. We have found that these wordlists can be calibrated to sequential ESP courses in successive years of the medical curriculum as suggested by Baker (1988). This would furnish a good base for students whose first language is not English and who have to develop reading and comprehension skills. This is particularly significant when clarity of meaning and precision in discussing medical principles is important in a field which is rapidly burgeoning.

[Paper received 21 December 2011]
[Revised paper accepted 15 August 2012]

References

Alsop, S. & H. Nesi (2009). "Issues in the development of the British Academic Written English corpus". *Corpora* 4: 71-83.

Anspach, R.R. (1988). "Notes on the sociology of medical discourse: The language of case presentation". *Journal of Health and Social Behavior* 29: 357-375.

Atkinson, D. (1992). "The evolution of medical research writing from 1735 to 1985: The case of The Edinburgh Medical Journal". *Applied Linguistics* 13: 337-374.

Baker, M. (1988). "Sub-technical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles". *Reading in a Foreign Language* 4: 91-105.

Barrows, H.S. (1990). "The pedagogical importance of a skill central to clinical practice". *Medical Education* 24: 3-5.

Bauer, L. & P. Nation (1993). "Word families". *International Journal of Lexicography* 6: 253-279.

Belcher, D. (1995). "Writing critically across the curriculum" in D. Belcher & G. Braine (eds.), *Academic Writing in a Second Language*, 135-154. Norwood, NJ: Ablex.

Belcher, D.D. (2007). "Seeking acceptance in an English-only research world". *Journal of Second Language Writing* 1: 1-22.

Berkenkotter, C. & T. Huckin (1999). *Genre Knowledge in Disciplinary Communities.* Mahwah, NJ: Lawrence Erlbaum Associates.

Berkenkotter, C. (2008). *Patient tales. Case Histories and the Uses of Narrative in Psychiatry.* Columbia, South Carolina: University of South Carolina Press.

Berkenkotter, C. (2009). "A case for historical "wide-angle" genre analysis: A personal retrospective". *Ibérica* 18: 9-22.

Booth, C.C. (1982). "Medical communication: the old and new. The development of medical journals in Britain". *British Medical Journal (Clinical Research ed.)* 285, 8335: 105-108.

Brier, B. (2004). "Infectious diseases in ancient Egypt". *Infectious Disease Clinics of North America* 18: 17-27.

Campion, M.E. & W.B. Elley (1971). *An Academic Vocabulary List*. Wellington: New Zealand Council for Education Research.

Chen, Q. & C.C. Ge (2007). "A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs)". *English for Specific Purposes* 26: 502-514.

Cohen, H. (2006). "How to write a case report". *American Journal of Health - System Pharmacy* 63: 1888-1892.

Coxhead, A. (2000). "A new Academic Word List". *TESOL Quarterly* 4: 213-238.

Coxhead, A. (2011). "The Academic Wordlist 10 years on: Research and teaching implications". *TESOL Quarterly*, 45: 355-362.

DeBakey, L. & S. DeBakey (1984). "The case report. II. Style and form". *International Journal of Cardiology* 6: 247-254.

Divasson-Cilvetti, L. & I.K. León-Pérez (2006). "Textual and language flaws: Problems for Spanish doctors in producing abstracts in English". *Ibérica* 11: 61-79.

Donnelly, W. J. (1997). "The language of medical case histories". *Annals of Internal Medicine* 127: 1045-1048.

Dyke, P., K. Jamrozik & A.J. Plant (2001). "A randomized trial of a problem-based learning approach for teaching epidemiology". *Academic Medicine* 76: 373-379.

Frederiksen, C. (1999). "Learning to reason through discourse in a problem-based learning group". *Discourse Processes* 27: 135 -160.

Ghadessy, P. (1979). "Frequency counts, word lists, material preparation: A new approach". *English Teaching Forum* 17: 24-27.

Giora, R. (1997). "Understanding figurative language: The graded salience hypothesis". *Cognitive Linguistics* 8: 183-206.

Granger, S. & M. Paquot (2009). *In search of a General Academic Vocabulary: A corpus-driven study*. URL: http://sites.uclouvain.be/cecl/archives/ In_search_ of_a_ general_academic_ english.pdf [09/03/11]

Heatley, A., I.S.P. Nation & A. Coxhead (2002). *RANGE and FREQUENCY programs*. URL: http://www.vuw.ac.nz/lals/staff/Paul_Nation/ [15/09/11]

Hopkins, A. & T. Dudley-Evans (1988). "A genre based investigation of the discussion section in articles and dissertations in three disciplines". *English for Specific Purposes* 7: 113-121.

Hsu, W. (2009). "Measuring the vocabulary of college General English textbooks and English-

medium textbooks of Business Core Courses". *Electronic Journal of Foreign Language Teaching* 6: 126-149.

Hunter, K. M. (1996). "Narrative, literature and the clinical exercise of practical reason". *Journal of Medical Philosophy* 21: 303-320.

Hyland, K. (2000). *Disciplinary Discourses: Social Interactions in Academic Writing*. London: Longman.

Hyland, K. & P. Tse (2007). "Is there and 'Academic Vocabulary'?". *TESOL Quarterly,* 41: 235-253.

Hyon, S. & R. Chen (2004). "Beyond the research article: university faculty genres and EAP graduate preparation". *English for Specific Purposes* 23*:* 233-263.

Kanoksilapatham, B. (2005). "Rhetorical structure of biochemistry research articles". *English for Specific Purposes* 24*:* 269-292.

Koh, G.C.H., E.K. Hoon, M.L. Wong & D. Koh (2008). "The effects of problem-based learning during medical school on physician competency: a systematic review". *Canadian Medical Association Journal* 178*:* 34-41.

Lam, J. (2001). "A study of semi-technical vocabulary in computer science texts, with special reference to ESP teaching and lexicography". *Research reports 3*, 65-78. Hong Kong: Language Centre, Hong Kong University of Science and Technology.

Lorés, R. (2004). "On RA abstracts from rhetorical structure to thematic organization". *English for Specific Purposes* 23*:* 280-302.

Lynn, R.W. (1973). "Preparing word lists: a suggested method". *Regional Language Centre Journal* 4*:* 25-32.

Martínez, I.A., S.C. Beck & C.B. Panza (2009). "Academic vocabulary in agriculture research articles: A corpus-based study". *English for Specific Purposes* 28: 183-198.

McCarthy, L.H. & K.E.H. Reilly (2000). "How to write a case report". *Family Medicine* 32: 190-195.

Medical Journals Editing (2002). "IV International Peer Review Congress". *Journal of the American Medical Association* 287, 21: 2749-2871.

Méndez-Cendón, B. (2009). "Combinatorial patterns in medical case reports: an English-Spanish contrastive analysis". *The Journal of Specialized Translation* 11: 169-190.

Moran-Campbell, E. J. (1976). "Basic science, Science and Medical Education". *Lancet* 307*,* 7951: 134-136.

Mpofu, D.J.S., J. Lanphear, T. Stewart, M. Das, P. Ridding & E. Dunn (1998). "Facility with the English Language and problem-based learning group interaction: Findings from an Arabic setting". *Medical Education* 32: 479-485.

Mudraya, O. (2006). "Engineering English: A lexical frequency instructional model". *English for Specific Purposes* 25: 235-256.

Mungra, P. (2007). "A research and discussion note: The macrostructure of consensus statements". *English for Specific Purposes* 26: 79-89.

Muresan, S. & J. Klavans (2002). *A Method for Automatically Building and Evaluating Dictionary Resources*. URL: http://www.cs.columbia.edu/nlp/papers/2002/muresan_klavans_02.pdf [25/09/11]

Myers, G. (2003). "Discourse studies of scientific popularisation: questioning the boundaries". *Discourse Studies* 5: 265-279.

National Library of Medicine. URL: http://umlsks.nml.nih.gov/ [04/10/11]

Nwogu, K. N. (1997). "The medical research paper: structure and functions". *English for Special Purposes* 16: 119-137.

Okamura, A. (2006). "Two types of strategies used by Japanese scientists, when writing research articles in English". *System* 34*:* 68-79.

Paltridge, B. (1997). *Genres, Frames and Writing in Research Settings.* Amsterdam: John Benjamins.

Peacock, M. (2002). "Communicative moves in the discussion section of research articles". *System* 30: 479-497.

Praninskas, J. (1972). *American University Word List*. London: Longman.

Rosen, T. (2008). "Ode to the case report". *Dermatology Online Journal* 14: 1.

Salager-Meyer, F. (1994). "Hedges and textual communicative function in medical English written discourse". *English for Specific Purposes* 11: 93-115.

Salager-Meyer, F., G. Defives & H. Hamelinsck (1998). "Epistemic modality in 19th and 20th century medical English written discourse: a principal component analysis". *Interface. Journal of Applied Linguistics* 10: 163-199.

Salager-Meyer, F., M.A. Alcaraz Ariza & N. Zambrano (2003). "The scimitar, the dagger and the glove: Intercultural differences in the rhetoric of

criticism in Spanish, French and English medical discourse (1930-1995)". *English for Specific Purposes* 22: 223-247.

Salager-Meyer, F., M.A. Alcaraz Ariza & M. Pabón Berbesí (2009). "Backstage solidarity in Spanish- and English- written medical research papers: publication context and the acknowledgement paratext". *Journal of the American Association of Information Science and Technology* 60: 307-317.

Salager-Meyer, F., M.A. Alcaraz Ariza, N. Luzardo Fernandéz & G. Jabbour (2011). "Scholarly gratitude in five geographical contexts: a diachronic and cross-generic approach of the acknowledgement paratext in medical discourse (1950-2010)". *Scientometrics* 86: 763-784.

Samraj, B. (2005). "An exploration of a genre set: Research article abstracts and introductions in two disciplines". *English for Specific Purposes* 24*:* 141-156.

Schmidt, H.G. (1993). "Foundations of problem-based learning: some explanatory notes". *Medical Education* 27: 422-432.

Sinclair, J. (2004). *Developing Linguistic Corpora: a Guide to Good Practice: Corpus and Text- Basic Principles*. URL: http://ota.ahds.ac.uk/documents/creating/ dlc/chapter1.htm [02/09/11]

Skelton, J.R. & S.J.L. Edwards (2000). "The function of the discussion section in academic medical writing". *British Medical Journal* 320: 1269-1270.

Skinner, B.F. (1956). "A case history in scientific method". *American Psychologist* 11: 221-233.

Sperber, D. & D. Wilson (1986). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Swales, J.M. (1990). *Genre Analysis. English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Taavitsainen, I. & P. Pahta (2000). "Conventions of professional writing. The medical case report in a historical perspective". *Journal of English Linguistics* 28: 60-76.

Vandenbroucke, J.P. (2001). "In defense of case reports and case series". *Annals of Internal Medicine* 134: 330-334.

Wang, J., S.L. Liang & G.C. Ge (2008). "Establishment of a medical academic wordlist". *English for Specific Purposes* 27: 442-458.

Ward, J. (2009). "A basic engineering English word list for less proficient foundation engineering undergraduates". *English for Specific Purposes* 28: 170-182.

West, M. (1953). *A General Service List of English words*. London: Longman.

Wood, A. (2001). "International scientific English: The language of research scientists around the world" in J. Flowerdew & M. Peacock (eds.), *Research Perspectives on English for Academic Purposes*, 71-83. Cambridge: Cambridge University Press.

Xue, G. & I.S.P. Nation (1984). "A University word list". *Language and Learning Communication* 3: 215-219.

Yitschaky, O., M. Yitschaky & Y. Zadik (2011). "Case report on trial: Do you, Doctor, swear to tell the truth, the whole truth and nothing but the truth?" *Journal of Medical Case Reports* 5: 179-180.

Philippa Mungra has taught English at the 1st Medical School of the University of Rome "La Sapienza" for the past 20 years. Her current research priorities revolve around the structure and evolution of specialist medico-scientific publications from a communicative and textual point of view and Integration of Language and Content (CLIL). She holds a M.Sc in cytogenetics from the University of Western Ontario, Canada where she has carried out her own scientific research on aneuploidy due to aging. She also holds the RSA Certification in TESOL from Cambridge.

Tatiana Canziani is a graduate from the University of Palermo, where she is currently a researcher. She is responsible for English Language teaching and testing for the entire Health Sciences sector within the Medical School. She is a trained Psychotherapist with experience in the paediatric field. Her current research priorities revolve around the lexicography of medico-scientific publications especially referring to Cognitive Linguistics. She has recently published a review of neuropragmatics regarding metaphor and brain function.

Appendix 1: List of terms and abbreviations altered to make text readable by Software

1. UNITS

a) Units of volume or size: nml/L = nanomoles per liter were inserted into the stoplist files. 141 x 170 x 208 mm was changed to 141 "by" 170 "by" 208 mm. Abbreviations of measure such as *cm* or *mm* for millimetres were inserted into the stoplist file as were other common abbreviations used in clinical medicine to describe laboratory parameters such as *ng/ml* or *Kg/M*^{2} or *m/z*.

b) Magnitude: such as 10^6 /ml was changed to "million per millilitre" as in the sentence " (…) the titre was $0.4x10^6$ / ml"

c) Slash: / was transcribed as "over" or "per" as in "(…) blood pressure …up to 250 /140 mm Hg". Similarly, in "(…) upper and lower limb power grades were 4 /5 and 3/5" the slash was transcribed as "over", common in spoken clinical science.

2. SYMBOLS

a) Mathematical symbols: > was changed to "greater than" or "over" and < "less than" as in the sentence "(…) should be given a therapeutic dose of the drug (5 mg for body weight <50 kg, 7.5 mg for 50 -100 kg and 10 mg for >100 kg)"

b) Hyphens: As required by Nation's software , hyphens were given spaces before and after the hyphen as in "(…) well - defined, soft - tissue nodules"

c) Symbols such as ~ in the phrase "(…) in ~50% of patients" was changed to "approximately" and °C to "degrees centigrade".

d) Acronyms: Commonly used acronyms were spelled out as in "(…) increase in both HR and BP" meaning "heart rate" and "blood pressure".

3. OTHER TEXT

a) Listings of pharamaceutical companies, names of authors, procedures, equipment and drugs were enclosed by <> brackets as in the sentence "(…) <Warkentin et al.> (2001) reported the occurrence of <HIT> after <fondaparinux>"

Appendix 2: MAWLcc

	MAWLcc - sorted by freq	FREQ
1	patient	1155
2	diagnose diagnosis	687
3	symptom symptomatic asymptomatic	621
4	clinic clinically	520
5	infect infected infections infectious	431
6	artery arteritis arteries	422
7	surgery	271
8	cardiac cardiovascular endocarditis	254
9	pulmonary	243
10	antibody antibodies	234
11	imaging	225
12	cell cellular	245
13	venous intravenous	212
14	pregnant pregnancy	205
15	kidney	201
16	cutaneous subcutaneous	200
17	fracture fractures	184
18	system systemic	175
19	nerve neurologic neurology	160
20	radiograph radiographic radiographs radiological	158
21	Inflame inflammatory	152
22	pelvis pelvic	151
23	tumors	149
24	recurrent recurrence	148
25	cyst cystic cysts	146
26	hemorrhage hemorrhagic	144
27	urine urinary	144
28	fetal	139
29	extremity extremities	136
30	victim victims	132
31	vascular intravascular	127
	vein ventricle	97
32	venous	119
33	abnormal abnormality	117
34	bilateral	117
35	resonance	116
36	thrombus thrombosis	116
37	posterior	114
38	nephritis nephrogenic	107
39	hip	106
40	distal	105
41	tomography	104
42	ultrasound	103
43	bowel	102
44	hematoma	102
45	elevate	100
46	intubate	99
47	obstruct	99
48	lymphoma	98
49	medication medications	98
50	medial	97
51	muscle	94
52	postoperative postoperatively	90
53	anterior	89
54	fibrosis	87
55	anaesthesia	86
56	congenital	85
57	discharge	85
58	emergency	85
59	excision	84
60	abdomen	83
61	myocardial myocarditis	82
62	discontinue discontinuation	80
63	histologic histological	80
64	millimeters metastasis	79

65	metastatic	78
66	carcinoma	77
67	coronary	77
68	cuff	76
69	infusion	76
70	stab stabbing	75
71	immune	73
	autoimmune	35
72	resection	73
73	infant	71
74	marrow	71
75	specimen	71
76	tendon	71
77	chemotherapy	70
78	transplantation	69
79	fatal	68
80	anomaly	67
81	bladder	67
82	therapy therapeutic	67
83	airway	66
	bacteria bacteremia	
84	bacterial	66
85	aneurysm	65
86	benign	65
87	cerebral	65
88	disorder	65
89	lumen	65
90	resistant	65
91	edema	64
92	episode episodes	64
93	gestation	64
94	compress compression	63
95	antibiotics	62
96	autopsy	62
97	forensic	62
98	malignant	62
	pathology pathological	62
99	pathogenesis	30
100	abscess	61
101	differential	61
102	hyperplasia	61
103	thorax thoracic	61
104	dose	60
105	dialysis	59
106	abuse	58
107	cervix	58
	cervical	
108	coagulate coagulation	58
109	drug	58
110	temporal	58
111	administer	57
112	invade invasive	57
113	necrosis	57
114	place placement	57
115	spine	57
116	maternal	55
117	progress progression	55
118	prolong	55
119	diabetes	54
120	dissect	54
121	focal	54
122	manifest manifestation	54
123	shotgun	54
124	function dysfunction	53
125	prognosis	53
126	regimen	53
127	rupture	53
128	vomit vomiting	53
129	aortic	52
130	lumbar	52
131	headache	51
132	hernia	51
133	remark unremarkable	50
134	cavity	49
135	rectal	49
136	transverse	48
137	vaginal	48
138	concentrate concentrations	47
139	angiography	46
140	nausea	46
141	nodules	46
142	transfusion	46
143	alarm	45
144	typical atypical	45
145	drainage	45
146	follow-up	45
147	hepatic	45
148	refer reference	45

149	tachycardia	45
	clear	
	clearance	
150	unclear	45
151	uterine	45
152	anticoagulation	44
	suggest	
153	suggestive	44
154	suicide	44
155	suture	44
156	septic	43
157	debridement	42
158	ectopic	42
159	anal	41
160	cesarean	41
161	platelet	41
162	puncture	41
163	ulcer	41
164	uneventful	41
165	ventilation	41
166	vertebral	41
167	analgesia	40
168	duct	40
169	lymph	40
170	autonomic	39
171	nasal	39
172	rotator	39
173	worsen	39
	coalesce	
174	coalition	38
	fix	
175	fixation	38
176	granuloma	38
177	ischemic	38
178	prescribed	38
179	radiation	38
180	susceptibility	38
	base	
181	basal	37
182	defect	37
183	fragment	37
184	occlude	37
	sense	
185	sensory	37
186	vitamin	37
187	arthritis	36
	embolus	
188	embolization	36
189	intact	36
190	massive	36
191	pulse	36
192	stenosis	36
193	tissue	36
194	axial	35
195	genetic	35
196	infarction	35
197	osseous	35
198	pneumonia	35
199	ankle	34
200	cast	34
201	etiology	34
	femur	
202	femoral	34
203	fibrous	34
204	gastrointestinal	34
205	giant	34
206	neonatal	34
207	phantom	34
	remiss	
208	remission	34
209	technician	34
210	tubular	34
211	tubules	34
212	bypass	33
213	conservative	33
214	cranial	33
	series	
215	serial	33
216	viral	33
	biopsy	
217	biopsies	32
218	breast	32
219	disc	32
	hospital	
220	hospitalization	32
	local	
221	localized	32
222	nodes	32
223	atrial	31
224	basement	31
	catheter	
225	catheterization	31
226	deposition	31
227	generic	31
228	hysterectomy	31
229	intravascular	31
	laparoscopy	
230	laparoscopic	31
231	organism	31
232	pediatric	31
233	rash	31
234	saturation	31
235	superficial	31
236	valve	31

	anatomy		
237		anatomic	30
	bone		
238		bony	30
239	dysplasia		30
240	ingestion		30
241	ureter		30