English for Specific Purposes World Online Journal for Teachers

ESP WORLD

ISSN 1682-3257
Indexed in
Genamics JournalSeek

Testing the validity of small corpus information

Sylvana Krausse
Languages Centre of Fachhochschule Nordhausen (University of Applied Sciences)
krausse@fh-nordhausen.de

Abstract

The present paper is based within the wider framework of the research that I am undertaking for my PhD-thesis. It will be a corpus-driven investigation into the lexical patterning of the suggestion: specialist language of Environmental Engineering English. For the sake of this investigation, I designed ENGICOR, a corpus of around 2 million words, consisting of "expert performances" (Tribble 1997:109) in the field of Environmental Engineering. During the process of planning the analysis of the corpus material I came across some interesting questions concerning the representativity of such a small corpus of specialist language and the validity of keyword analysis results. These interim findings will be subject of this paper.

The context of the study

More and more study programmes of German universities and universities of applied sciences reflect the necessity to provide students with specific language courses that are tailor-made to fit the needs of their particular course of study. These courses build on the English taught in second-level schools and should provide the subject-related language knowledge and skills the students need during their studies and in their future professions.

Specific teaching material for such courses is scarce or non-existent and one additional role that the teachers have to take on a large scale is that of the material designer. In order to fulfil this task the teachers need to apply certain methods and analyses to make sure the material they choose is valid and suitable. The search for teaching material should be guided by principles and not by chance. The above-mentioned PhD-study is planned as a contribution to solving this problem, where an analysis is demonstrated by means of the specific language Environmental Engineering English, which serves as a perfect example. Already in the 1950s Firth spoke of the need to create mini-glossaries and mini-grammars for restricted languages (Firth in Palmer 1968:106). The starting point of investigations into language from Firth's point of view were the "key-words, pivotal words, leading words, by presenting them in the company they usually keep" (Firth in Palmer 1968:106). This tenet that has become one of the main fields of exploration for modern corpus linguists will form the guiding principles for the dissertation and the main thoughts of this paper will also reflect this.

The found language data in form of the "pivotal words" and their surrounding company, in other words the subject-specific multiword items (hereafter referred to as mwi) will be the basis for syllabus design for the English courses and will thereby ensure that the students are presented with the most relevant language material which is a prerequisite for learning the "right" Environmental Engineering English and thus enable students to achieve more naturalness in language use (cf. Sinclair 1984).

The corpus

The foremost aim of my dissertation will be to establish a basis of lexical information of the most relevant mwi of Environmental Engineering English to serve as a valid source for syllabus design of specialist English courses at University level.

In order to decide on the topics and the text types to include in the corpus I sent a questionnaire to all the companies that had ever employed a graduate or a work placement for students from the course of study for which the English classes are designed. From the feedback of this survey it was clear that getting hold of samples of every mentioned text type was an impossible task given the time restrictions and availability of the material. For the same reasons, spoken texts were not included. However, to cover the recurring topics seemed to be a worthwhile endeavour. To name but a few, a small number of topics will be mentioned at this stage in order to provide an idea as to what Environmental Engineering deals with: *wastewater treatment, treatment of organic waste, revegetation of mining waste dumps, recycling of metals, plastics, debris, remediation of contaminated soil*, etc..

Taking the notion of topic as a starting point is an approach criticised by Sinclair, who holds that "it is clearly an internal matter, because the topic of the text is defined by the way in which language is chosen and used in the text;….The problem is twofold. One is of circularity, mentioned above. … The other problem is that there are as many classifications as there are researchers." (2003:172) Sinclair warns that corpus designer should be aware of subjectivity. He furthermore includes another useful idea for the corpus compiler; the idea that texts should be chosen because of the "social role" they play within the context of communication.

With these notions in mind, the research for texts to be included in the corpus started at the website of the Environment Protection Agency, a US government body that provides information sheets, technology fact sheets and reports of various kinds to the Environmental Engineer but also the interested American citizen, a social role which goes hand in hand with the future communicative situations with which the students have to be able to deal. The topical keywords from the survey were located on this website and suitable texts were extracted for inclusion in the corpus. Where the information of this website did not suffice, the few hints from the website and the topical keywords from the survey served as useful pointers as to where to start Google searches for more material.

In the end, the corpus comprised 2 million words which were presented in 6 subcorpora. These subcorpora on the one hand represent the main topics the students of Environmental Engineering have to deal with at our university: Mining Reclamation, Soil Reclamation and Waste Disposal and Recycling. On the other hand, the subcorpora reflect various text types, like the subcorpus of web-based journals, the one of technology fact sheets and the one consisting of an introductory textbook. Within the text-type based subcorpora, effort was made to ensure that at the same time the texts reflected the above mentioned topics.

The subcorpora comprise one section of mining reclamation material (approx. 200,000 running words), one of soil remediation material (approx. 300,000 running words), approx. 200,000 words of technology fact sheets, 120,000 words from an introductory Environmental Engineering textbook, 400,000 words on waste treatment and recycling and approx. half a million words from web-based journals, all in all a corpus of 2 million running words.

Representativity

In the last paragraph the need to justify the choices made in corpus design has already been made obvious in the light of the claim that corpora should be representative in order to make valid statements about language. Representativity has become a much debated concept not only in small corpus studies (cf. Williams 2002, Teubert 2003). It is felt to be an ideal which should be aimed at by comprising as wide a range of samples as possible within certain limits. A truly representative corpus of Environmental Engineering English would consist of every utterance ever made and every text ever written in this professional field, an unachievable task.

Atkins et al comment on this dilemma : «… in our ten years of analysing corpora for lexicographical purposes, we have found any corpus - however unbalanced – to be a source of information and indeed inspiration." (1992:6) What this quotation says about the work of lexicographers holds true for teachers as well. As no suitable teaching material exists to cater for the needs of specialist language courses there is a strong need for small corpora of these sublanguages. If representativity is not the foremost corpus design criterion, other criteria have to take its place. Other small corpus studies like Hanchen (2002) mention diversity of addressees, topics and text types, Curado (2002) brings topic relevance and updatedness, course syllabi and availability into play. In the case of ENGICOR all these ideas have been considered to a certain degree when choosing the websites. However, when making claims on the basis of small corpus analyses, the shortcomings of every corpus have to be kept in mind. And this is what my PhD-thesis aims to show in the field of Environmental Engineering English. On my way to understanding the lexical level of Environmental Engineering English, the question as to what degree ENGICOR is "representative" yielded the following food for thought: working with small corpora is definitely valid and a step in the right direction to devising new ways in which specialist language can be learnt.

Corpus comparison

The cooperation with the companies did not end with the initial survey. After the corpus compilation they were contacted again and asked to provide some written English texts that they came across during their daily work routines. Compared to the initial survey where more than half of the companies filled in the questionnaire, the reply to this request was quite weak but a mini-corpus of eleven texts with altogether 45,600 words could be compiled. The topics ranged from brownfields and mining reclamation through biogas collection to river basin management, topic-wise clearly Environmental Engineering themes.

This mini-corpus of texts provided by the companies was used to make comparisons with ENGICOR. First, the wordlist of the two corpora were compared with the help of detailed consistency in WordSmith 4. The investigation revealed that the mini-corpus comprised 195 words that did not occur in ENGICOR (text-specific abbreviations, proper names, genitives have been disregarded). Then the frequency of these words was determined in order to see how much text coverage these words would yield and it was found that this figure amounted to 0.85%. After having a closer look at the 195 words, they were divided into certain categories of which geographical names (*Poland*, *Colombia*), American English/British English variation, easy words (*dear*, *umbrella*, *handout*) and obvious language mistakes (*shutted*, *criterions*) were eliminated as they were not regarded to count for words that could have occurred in ENGICOR if it was better designed. This left the number of words in the mini-corpus which did not feature in ENGICOR at 95 and their frequencies revealed that they only account for 0.25% text coverage.

At this point it has to be mentioned that I am not suggesting that a student is supposed to know all word-forms occurring in ENGICOR (which amount to 33,105). Text coverage is meant here as a measure that does not include a reader but only the two corpora. Furthermore, starting the whole journey from the wordlists does not mean the unit of investigation is a single word but rather the

most common mwi and therefore words are only the crutches by which the investigation is supported.

0.25% divergence is regarded to be insignificant by this study and shows that claims on corpus grounds about lexical patterning in Environmental Engineering English, for which the words are only a starting point, will be valid.

For lack of a better word, ENGICOR seems to be "representative" of a high percentage of Environmental Engineering English and as a small corpus it promises useful information about linguistic characteristics of this specialist language.

Justifying the work with keywords

When planning the analytical procedures of corpus investigation the first question was which word-forms to choose and how to delimit the scope of this study. As the corpus comprises 33,105 types (single word-forms) of which 10,749 alone are hapax legomena (word-forms that occur only once in the corpus), some form of restriction was deemed necessary.

Numerous attempts to create wordlists of most appropriate or most common words for language learning or testing, reflect the striving to explain the relationship between language learning effort and profit. The following quotation by West, the designer of the General Service List, underlines the necessity of such groundwork for language teaching: " A language is so complex that selection from it is always one of the first and most difficult problems of anyone who wishes to teach it systematically." (West 1953:V). His General Service List includes the most frequent 2,000 general words of the English language. Another more recent study, this time on the basis of modern corpus linguistics, is the Academic Wordlist of Coxhead (2000), which comprises the 570 most frequent word families occurring in academic texts no matter which specific subject they represent.

Following these examples, the first step in the search for the most salient words of Environmental Engineering English was to make a frequency list of the word-forms occurring in ENGICOR. This was done by the help of the Wordsmith 4 software. The resulting frequency list looked at the beginning very much like any other corpus-based frequency list, starting with the most common function words like *the, of, and, to, in, a, is, for, that, be*. These words only have a support role in the specific language and should be known from school. Therefore they are not considered worth being the starting point in the investigation of mwi in this study. Still, with *water* as the first content word to come up in position 22, *waste* in 28, *system* in 31 and *site* in 42, ENGICOR starts showing its profile in this way.

The sheer frequency list as basis for the selection of word-forms or analysis was disregarded at this stage. The next procedure chosen was to create a frequency list that would not display the most common words that were in most cases known from school. In order to achieve this, a new frequency list was created, but this time without all the words featuring in the General Service List and the Academic Word List (by means of the stop list function of WordSmith 4). In doing this, the hope was to come up with a word list which would be the filtrate of Environmental Engineering English and some less frequent words that could have been eliminated. Again, the outcome was not satisfactory for the purposes of this study as words like *mine, waste* and *water* occur in the General Service List and the Academic Word List and have consequently been taken out. Although it can be argued that the meaning of these words should be known by students who went through the German schooling system, at this stage the reader should be reminded of the point that it is not the single word-form the study is after but the longer language patterns, the mwi. As long as the format and applicability of lists of mwi is not established, lists of single word-forms remain the aiding device.

The procedure that proved to yield the best results in the end was a keyword-comparison. Again

WordSmith4 was used, this time to create a keyword list. For such a comparison the frequency list of my own corpus has to be compared with a reference corpus. In this study, the British National Corpus (BNC), a general language corpus of 100 million words was chosen to serve as a reference corpus. The reason for comparing ENGICOR with a general language corpus was that in this way the divergence from the general English could best be displayed and would show all word-forms that occur significantly more frequently in Environmental Engineering English as opposed to in general English.

Here are the first 20 positive keywords from this comparison which can be said to already smell and taste of Environmental Engineering English: *waste, water, tailings, wastewater, site, treatment, soil, landfill, groundwater, material, contaminant, concentrations, percent, flow, contaminants, effluent, recycling, chlorine, system, solids*.

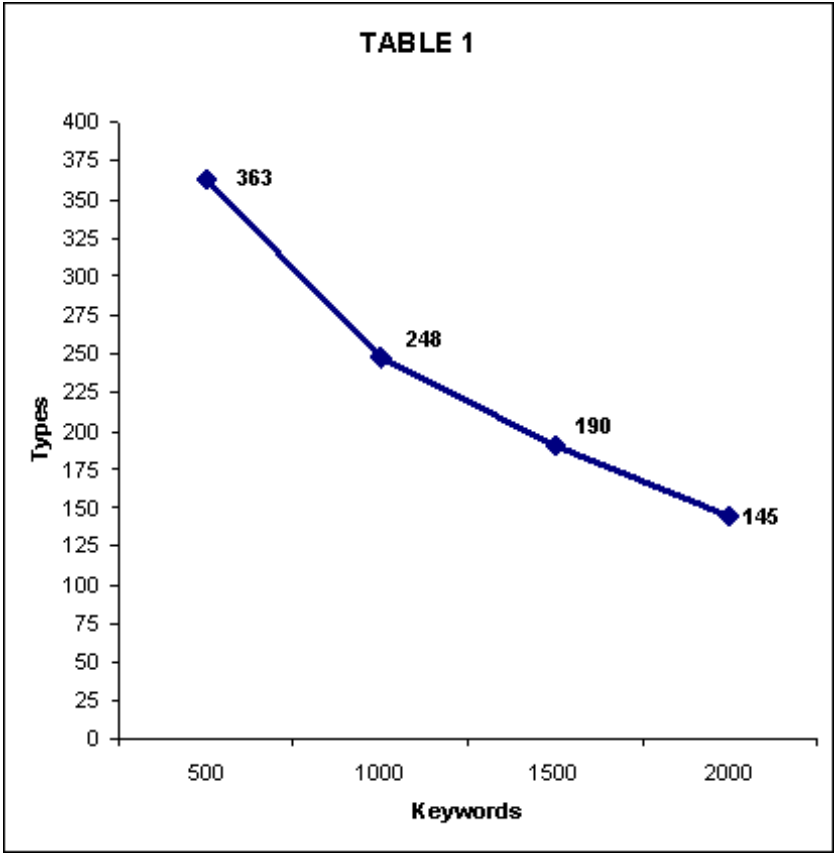
From the point of view of this study it is not considered as a drawback that the list consists of technical words and subtechnical words alike (for differentiation of these terms see Curado 2002:Chapter 1/V). It is felt that distinguishing between technical and subtechnical words is not a sensible thing to do in the light of teaching the subject-specific language to students who either have no or just a very minor level of tuition in English during their study programme. Anyway, decisions on which words will be more the centre of interest will be based on the relationships with other words with which they will function more closely. Furthermore, some very prominent function words can be also found in the keyword list like *is* which is ranked at 237thand *are* ranked at 115th place. This is a strong indicator to demonstrate that the passive voice is used a lot in scientific and technological texts.

For the sake of the PhD-thesis, the first 500 word-forms of the keyword comparison between ENGICOR and the BNC were chosen with a view to making them the subject of a detailed analysis on the lexical (collocation), lexico-grammatical (colligation) and morphological level and in certain cases also to investigate the semantic and pragmatic level (semantic preference and semantic prosody) (cf. Sinclair 1996). As this is the main ground to be covered by the further analysis in the course of the PhD-thesis, the question that remains to be answered by this paper is to what extent the choice of the first 500 words of the keyword comparison is justified in terms of learning effort and extent of the text coverage of these 500 words. How reasonable is it to concentrate on words from a keyword comparison between a specialist language corpus and a general language corpus?

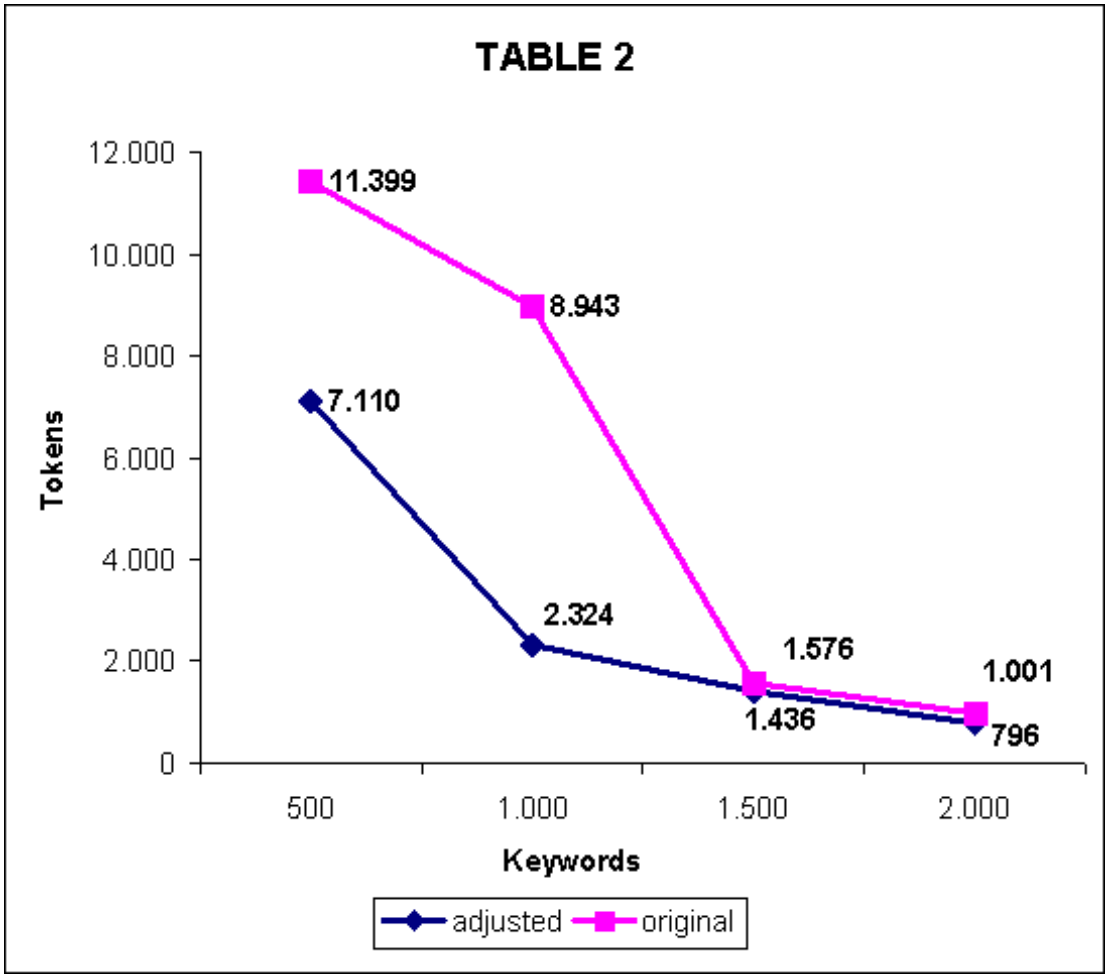
By the way, 500 words is an arbitrary number. 500 lexical items were deemed to be feasible to make analyses within the major study of the thesis.

Corpus comparison again

In order to find evidence for the suitability of the first 500 keywords the mini-corpus of texts provided by the companies came into play again. In batches of 500 words, the first, second, third and forth 500 words from the keyword comparison were located in the frequency list made up of the texts from the mini-corpus. Table 1 reflects the first findings which reflect that more items from the first 500 keywords occur in the mini-corpus compared to the second, third and fourth 500.

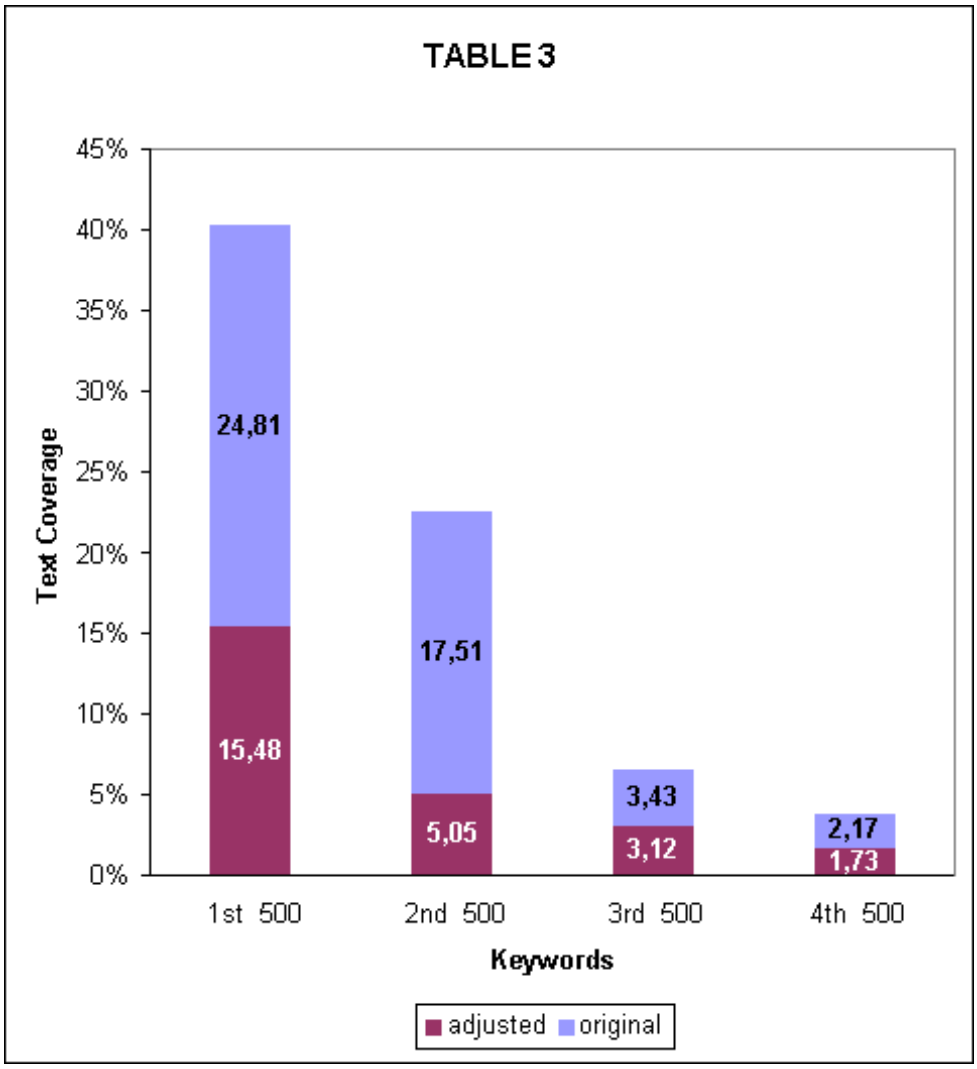


As a next step the frequencies of these items were established and their values can be seen in Table 2, showing as well that the first 500 keywords also yield more tokens (how often the items occur) compared to the second, third and forth 500 keywords.



It is important to mention that with this rather crude analysis the aim was to try and detect a trend rather than absolute numbers. It would be interesting to see a more detailed analysis which would include lemmatisation and the elimination of function words. If the lexical items of both lists, keyword list and mini-corpus word list were lemmatised, a tidier picture could have been drawn as the students are supposed to understand words from the same word family. The inclusion of function words were seen as a disadvantage particularly when looking at words like *the* which occurred 3611 times, and *which* occurred 1,406 or *of* which occurred 2,239 times in the frequency sums of the 500 keyword batches. Therefore in the second graph of Table 2 these very frequent items were eliminated which results in a slightly different graph which nevertheless follows the same trend.

The amount of times a lexical item occurs in a text, or in a corpus, as is the case here, is called the text coverage of this item. In other words, spending time and effort on the first 500 words is more fruitful than on the second 500, which in turn is more fruitful than on the third 500 and so on. Table 3 shows the percentage of text coverage that was achieved in the mini-corpus by the first, second, third and fourth 500 keywords.



Again, there is one version which uses the raw figures including also the most prominent function words and one version where these items were eliminated. The differences in the amount of text coverage by the first 500 keywords is significantly higher compared to the other batches of words. 15.48 % of text is covered by them, compared to 5.05% of the second batch, 3.12% of the third and 1.73% of the third (in the adjusted version).

Looking back at the type-token ratio (a type being a lexical item in a corpus and the quantity of its tokens is how often this item occurs in the corpus) compared to the percentage of text coverage, it can be said that the quotient of text coverage to type frequency proves to be higher for the first 500 keywords and decreases along the line (being 0.04% for the first batch, 0.02% for the second, 0.016% for the third and 0.011% for the fourth. This can be considered as further evidence of the fact that concentrating learning efforts on the first keywords is time well spent.

Conclusion and outlook

The effort spent in the analyses described by this paper are based on the belief that the specific

language patterns for English courses at university level should be carefully selected and reflect what students are required to achieve with language in their future professions. Working on intuition as regards which texts and teaching materials to choose within the subject specialism is not sufficient, and small corpus work can perfectly fill the gap that the lack of suitable material leaves.

This study shows the attempts to secure a certain degree of objectivity in the process of building the corpus and establishing the first operational procedures. First a comparable corpus of Environmental Engineering texts was used to prove that the 2 million-word ENGICOR corpus is representative enough to draw reliable conclusions from its analysis. Only 0.25% of text of the mini-corpus was not covered by words in ENGICOR, a figure that is considered small enough to be optimistic that lexical work based on ENGICOR can be considered typical of Environmental Engineering English. Second the results from a keyword comparison between ENGICOR and the BNC were located in the comparable corpus and it can be stated that words that range high up in the keyword list cover more text in the comparable corpus than those occurring in lower positions in the list. On this basis, the decision to analyse the first 500 word-forms of this keyword list in the framework of the PhD-thesis was taken.

As this paper only represents the initial steps in the analytical planning, much work is left for the rest of this research journey. Corpora deliver information on various aspects of language in general and of specialist language. This corpus information on the lexico-grammatical and morphological levels will be collected first starting from the first 500 keywords. A format of storage of the found data will be established which will illustrate the collocational and colligational relations which come into play when certain words are placed with others and their corpus-attested word-formation patterns. The phenomena of semantic preference and semantic prosody will only be a subject of some case studies, as an exhaustive investigation of this field for all 500 word-forms would be beyond the scope of study. The established linguistic data is meant to serve as a basis for syllabus design for the Environmental Engineering English courses and in their extracted form will be a suitable material for work on the specialist language. All in all, the work on a corpus of one specialist language can serve as model for work on other specialist languages and guide the search of teachers for an empirically proven way to organise their courses.

Bibliography

Palmer, F.R. (1968): Selected Papers of J.R.Firth 1952-59. London Beccles: Longman

Sinclair, J. (1984): Naturalness in Language. *Corpus Linguistics* Volume 45(): 203-211

Sinclair, J. (2003): Corpora for dictionaries, in Sterkenburg, P. van (ed.) (2003) *A Practical Guide to Lexicography*. Amsterdam: John Benjamins Publishing Company

Atkins, S./ Clear, J. and Ostler, N. (1992): Corpus Design Criteria. *Literary and Linguistic Computing* Volume 7(1): 1-16

Curado, A. (2002) A Lexical Common Core in English for Information Science and Technology. Caceres: Servicio de Publicaciones de la Universidad de Extremadura

Hanchen, R. (2002) : Die Franzosische Marketingsprache : Eine diachrone Untersuchung ihrer Terminologie anhand der Revue Francaise du Marketing (1960-2000). (*Sprache im Kontext* 13). Frankfurt: Peter Lang

West, M. (1953): A General Service List of English Words. London Beccles Colchester: Longman

Coxhead, A. (2000) : A New Acedemic Word List. *TESOL Quarterly* Volume 34(2):213-239

Sinclair, J. (1996): The Search for Units of Meaning. *Textus* IX:75-106

Teubert, W. (2003): Corpus Linguistics – A Partisan View. *International Journal of Corpus Linguistics* 5(1), http://tractor.bham.ac.uk/ijcl/teubert_cl.html

Tribble, C. (1997): Improvising corpora for ELT: quick and dirty ways of developing corpora for language teaching. In B. Lewandowska-Tomaszczyk and P. Melia, *Practical applications in language corpora*. Lodz: Lodz University Press. 106-117

Sanchez, A. (2000): Language Teaching before and after "digitilized corpora". Three main issues. *Cuadernos de Filologia Inglesa* 9(1): 5-37. Corpus-based Research in English Language and Linguistics. Universidad de Murcia.

Williams, G. (2002): In Search of Representativity and Specialised Corpora – Categorisation through Collocation. *International Journal of Corpus Linguistics* 7(1):43-64

WordSmith 4 - information about the program can be found at
http://www.lexically.net/wordsmith/version4/

British National Corpus – information about the corpus can be found at
http://www.natcorp.ox.ac.uk/

Biostatement:

Sylvana Krausse works as lecturer in foreign languages (English/French) at Fachhochschule Nordhausen, Germany. She is currently head of the languages centre and is working on her PhD-thesis with the working title "A Corpus-Driven Study into Environmental Engineering English".