THE USV ANNALS OF ECONOMICS AND PUBLIC ADMINISTRATION | VOLUME 20, ISSUE 2(32), 2020

QUANTITATIVE ANALYSIS OF BUSINESS COMMUNICATION

Oana COSMAN
"Ştefan cel Mare" University of Suceava, Romania
oana.cosman@usm.ro

Received 28 September 2020; Accepted 15 December 2020

Abstract:
Considering that the analysis of any specialized language should have a representative set of terms, this paper describes the selection of specific Romanian business terms, extracted and processed from a self-compiled corpus, to carry out a quantitative and qualitative analysis of Romanian business terminology in order to build an updated Romanian-English glossary of business terms and collocations. The analysis of these linguistic structures is based on the use of analytical techniques that belong to corpus linguistics as it would be impossible to detect them intuitively. The study focuses on the design of a real corpus extracted from the Romanian business environment. The objective of this paper is to describe the methodology followed in order to compile a corpus for Romanian business language and propose a list of 100 most frequent business words in Romanian. This proposal is made after the compilation and analysis of a study corpus of approx. 1 million words extracted from 8 genres belonging to written business communication. The present paper is part of a larger study that carried out a contrastive analysis of the Romanian and English business languages in order to find similarities and differences between the two target languages.

Key words: business communication; specialized terminology; corpus analysis; Wordsmith Tools 6.0.

JEL classification: C18, C49, C60, C80, Y80, Z11, Z13, Z19, M29

1. INTRODUCTION

It is essential to posses the terminological keys to decode the message of a given professional domain, that is to know the terms of that specific domain. Thus, terms become vehicles of those concepts that determine a certain field of knowledge. The researchers that have used corpus linguistics have demonstrated the analytical potential of the techniques applied in the analysis of specialized language terminology (Stubbs, 1996, McEnery & Wilson, 1996) in comparison with the traditional methods for linguistic analysis. The concept of 'terminology' has several definitions: "specialized language [...] which uses terminology and other linguistic or non-linguistic means to achieve unambiguous specialized communication" (DSL, 2005:535); "the totality of the specialized terms used in a discipline or in a branch of activity" (DEX, 2009); "vocabulary used in a professional field" (Petit Robert, 1995)[1]; "the technical or special terms used in a business, art, science, or special subject" (Merriam-Webster Dictionary); "special words or expressions used in relation to a particular subject or activity" (Cambridge Dictionary); "the body of terms used with a particular technical application in a subject of study, theory, profession, etc." (Oxford Dictionary).

The international standard ISO 1087:2019 defines terminology as "a set of designations and concepts belonging to one domain or subject". M. T. Cabré (1999) defines terminology as ″an interdisciplinary field of enquiry whose prime object of study are the specialized words occurring in natural languages which belong to specific domains of usage″ (1999: 32). For the specialists in the business domain, terminology can thus be perceived as ″a necessary medium of expression and professional communication″ (1999: 11). Another definition of the concept of terminology is that of a specialized language, ″a linguistic system which uses a terminology and other linguistic means which target communication non-ambiguity in a particular domain" (Lerat, 1995:32).

Nowadays, corpora play an essential role in a wide range of linguistic investigations and a very important trend in academic research takes into account the connection between Artificial Intelligence and terminologies, generated by the use of corpus linguistics. Accordingly, we assert that a *corpus*, defined as a collection of machine-readable authentic texts, sampled to be representative of a particular natural language/language variety (McEnery & Wilson, 1996:5), may

become remarkably important in business communication research due to the properties that it acquires if it is well-designed and carefully-constructed. Hence, our research focuses on the investigation of business terminology extracted from a self-made corpus of Romanian business texts. The benefits of using corpus linguistics techniques rely on the possibility of thoroughly interpreting various features of existent elements and the opportunity to detect new patterns (i.e. collocations) in written business communication, based on electronic data analysis.

The most important principle in our study resides in the hypothesis of an equivalency between word frequency and its importance in the study corpus. Consequently, the analysis carried out emphasizes the most frequent realizations of the dominant semantic relations, which have been identified in a list of keywords and collocations extracted from our study corpus. We believe that the corpus analysis approach based on frequency avoids the subjective choice and introduces an objective criterion to written business communication research. Thus, the specialized language of business in Romanian is analyzed in terms of 'keywords', that is those words that are unusually more frequent in a study corpus than in a general corpus.

The paper was inspired by A. Coxhead's Academic Word List (AWL)[2] and Mike Nelson's research on a corpus-based study of Business English as we wanted to get the most significant words of Business Romanian to find out if there is a common core business vocabulary in Romanian and English. To carry out our research, we created a 1,000,000-word Business Romanian Corpus (RBC) to retrieve word lists based on this study corpus. We made comparisons to general Romanian using the ROMBAC corpus as a reference. To sum up, we consider that the use of a computer-based corpus analysis may provide a solid empirical foundation for Romanian business language tools and descriptions and enable research of a scope not otherwise possible.

2. METHODOLOGY

Since a corpus must be 'representative' (well-designed and carefully-constructed) to be appropriately used as the basis for generalizations concerning a language as a whole (Biber, 1993b), in building our study corpus we strived to make it as 'representative' as possible of the language from which it was chosen – business Romanian, taking into account the target notions of *balance, sample,* and *representativeness to guide the design of the corpus and the selection of its components. Since for a corpus to be pronounced balanced, 'the proportions of different kinds of text it contains should correspond with informed and intuitive judgements' (Sinclair, 2004), we tried to roughly align the written components of the corpus (subdivided into newspapers, magazines, books etc.) so that there is not too much very formal or very informal language in the corpus as a whole. In terms of sampling, we tried to incorporate samples of language for our corpus that consisted of entire documents wherever possible or as close to this target as possible (Sinclair, 2004), meaning that the samples differ substantially in size. To establish representativeness, we took into account that the business language varieties should be proportional with their importance and tried to cover a wide variety of large business documents to obtain the proportional representation and diversity elements needed for the validity of our study corpus, so we included texts from both the private sector and the institutional one.*

The two key elements, diversity and proportional representation, led us to the size of the corpus ('running words'), so we tried to cover a very large area and included texts from university courses, textbooks, brochures, internal rules and regulations, guides, legal documents, press releases, job advertisements, annual reports, administrative reports, and business correspondence. Subsequently, we compiled a corpus of about 1,000,000 running words to meet all the target conditions, which includes a sufficient number of words for conducting research on Romanian business language. However, we are aware that the study corpus may not capture all the patterns of contemporary Romanian business language, nor represent them in precisely the right proportions, as no corpus, no matter how large, how carefully designed, can have exactly the same characteristics as the language itself. We should also mention that when compiling the corpus, we disregarded the diverse characteristics of language used for communication (speaker age, gender, level of

education, and socioeconomic background; place and time of a communicative event; relationship between interlocutors) as these features were not part of our research. As regards the type of language, our research has focused on written business communication for practical purposes.

3. MATERIALS

When compiling the study corpus for our research, we took into account Picket's (1988) view, who states that there is an essential difference between '*knowing'* about something and '*acting'* (1988:90); to extrapolate, we could affirm that there is a difference between the specialized language needed for knowing about a topic in business communication and the language needed for actually being able to perform in the business world.

Thus, the materials were divided into two main categories: '*writing about business'* and '*writing to do business'*.The '*writing about business'* category totalizes 486,000 running words, which includes business texts from the mass media (articles taken from online daily, weekly and monthly newspapers and magazines) and the academic field (academic papers, textbooks, teacher guides and other support materials). The '*writing to do business'* category comprises business texts such as annual reports, administrative reports, job postings, press releases, e-mails, flyers, legal documents (i.e. commercial contracts, articles of incorporation, articles of association etc.), totalizing 551,000 running words. We decided not to divide these texts into further categories (i.e. management, marketing, finances etc.), but to take into account the most prevalent terms found in Romanian written business communication.

As a result, the compiled *Romanian Business Corpus* (henceforth RBC) comprises a total of 1,037,000 words from approx. 1,129 texts, which was saved on a CD because of its extremely large dimensions. The tables below summarize the materials used to compile the RBC study corpus:

Table 1. Corpus of business language in Romanian. Texts about business

Component elements of the RBC Study Corpus	**No. of words**	**Contents**
Education (textbooks, courses, etc.)	270,000	11 extracts from various materials
Mass-media	216,000	130 articles from: Ziarul Financiar, Wall Street, BizCity, BloomBiz, Smart Financial, Banii noștri, Business Standard, Bani și Afaceri, Daily Business, Ziarul Economic, Money Express, Săptămâna Financiară, Business Magazin, Tribuna Economică, Top Business, Bucharest Business Week, Capital MarkMedia, Business Romania, Biz, Business Adviser, Euroinvest, Idei de afaceri, Cariere, Financiarul, Bilanț, Ghidul de Bani, Banker-ul, E-Finance.
TOTAL	**486,000**	

Table 2. Corpus of business language in Romanian. Texts about doing business

Component elements of the RBC Study Corpus	**No. of words**	**Contents**
Annual reports and reports of administrators	123,000	10 reports
Job postings	25,000	370 job postings from various companies
Legal documents	105,000	47 decisions of annual *general m*eetings (AGM) of shareholders, specifications, business contracts, articles of incorporation, constitutive acts, additional acts
Press releases	103,000	193 press releases
Business e-mails	86,000	350 business e-mails
Brochures, guides, regulations, manuals	109,000	18 documents
TOTAL	**551,000**	

The corpus used as the source of the keywords had to be compared to a general corpus so that keywords and terms can be identified correctly. Hence, to be able to extract these keywords and terms, we used the *Romanian Balanced Annotated Corpus* (henceforth ROMBAC) as a reference corpus. ROMBAC is a large balanced corpus for Romanian in XML format, constructed at the Research Institute for Artificial Intelligence of the Romanian Academy. This is the first large and richly annotated corpus for Romanian, which was intended to be the foundation of a linguistic environment containing a reference corpus for contemporary Romanian (Ion et al., 2012). The reference corpus contains about 36,000,000 words evenly distributed into five genres: journalistic (news and editorials), pharmaceutical and medical short texts, legalese, biographies of the major Romanian writers and critical reviews of their works, and fiction (both original and translated novels and poetry). Therefore, we consider it a representative and balanced corpus for our research. The table below presents the corpora used in the analysis of the Romanian business language:

Table 3. Corpora used in comparative analysis of Romanian and English business language

Corpora used in the study	No. of running words
Self-compiled Romanian Business Corpus – RBC study corpus	1,037,000
Romanian Balanced Annotated Corpus – ROMBAC reference corpus	60,000,000
Business English Corpus – BEC study corpus	575,703
British National Corpus[3]– BNC sampler	100,000,000

3.1. CORPUS DESIGN

Developing a study corpus presupposes defining its structure, its linguistic coverage, collecting texts according to the established structure, solving problems of copyright, processing text with linguistic technologies (segmentation, lemmatization, tagging etc.), extracting statistical data etc. (Ion et al., 2012). One of the first considerations in constructing the RBC study corpus concerned its overall design: the number and kind of texts to include; the selection of these particular texts; the length of text samples etc. (Biber, 1993b). One mention should be made: we chose criteria that are easy to establish to avoid a lot of labour at the selection stage; also, they are of a fairly simple kind so that the margin of error is likely to be small. We rejected the criteria that are difficult to establish, complex or overlapping as errors in classification 'can invalidate even large research projects and important findings' (Sinclair, 2004). Taking into account Sinclair's (2004) views on corpus creation, this study has used the following criteria to compile the corpus:

1. *the mode of the text – i.e.* whether the language originates in speech or writing, or in electronic mode: all texts were acquired in electronic form and the format of the files and their encoding was MS DOC;

2. *the type of text* – i.e. whether a book, a journal, or a letter: the types of texts included in the corpus have been extracted from 8 different discursive genres (i.e. books, university textbooks, articles taken from business newspapers, magazines and journals, business e-mails, press releases, job advertisements etc.);

3. *the language or language varieties of the corpus*: the texts were collected from contemporary written Romanian language;

4. *the domain of the text – i.e.* whether academic or popular: both the RBC study corpus and the ROMBAC reference corpus contain texts from various domains (academic, journalism, business correspondence etc.) which include formal and informal written language, representing various social and situational layers;

5. *the date of the texts*: the textual collection of the RBC study corpus is made up of publications covering the period from 2011 to 2014;

6. *the location of the texts*: the texts were collected from standard Romanian.

As such, this paper identified linguistic structures that encode textual specialized functions and meanings which are intuitively imperceptible. The present analysis has focused on the inventory of specialized terms extracted from business texts that are part of our RBC study corpus. Consequently, the benefits of using analytical techniques belonging to corpus linguistics in the study of specialized texts, compared to traditional approaches (which do not use computer programs for linguistic analysis) reside in the possibility to detect new meanings in the specialized lexicon by analyzing new linguistic structures and in the opportunity to perform an in-depth interpretation of the features of objectively existing elements. Both advantages are based on electronic data analysis and the most important principle lies in the hypothesis of an equivalence between *frequency* and *importance* in the study corpus. Hence, we assume that certain features that appear very frequently in the RBC study corpus are fundamental to the structure of written business communication and data analysis. Thus, an increased frequency of certain features can be explained as a way of emphasizing their importance in the text, because, due to their frequency, these traits have a significance, either for the content of the text or for its structure. Therefore, the analyzes in our thesis are based on the most frequent or dominant achievements of the discussed characteristics, i.e. on the dominant semantic relations identified in a list of *keywords* and *collocations* in the RBC study corpus. Moreover, the validity for the interpretation of the results is ensured by the 'concordance lines' for each linguistic unit analyzed.

The lack of electronic tools for exploring and exploiting a corpus of texts dedicated to business language, or solutions for incorporating it into other existing work tools, may reveal current trends in Romanian language research on specialized business discourse. The lack of such tools may explain the lack of linguistic sources in the Romanian linguistic landscape, such as: general dictionaries, specialized dictionaries, thesaurus dictionaries, glossaries of terms, automatic proofreaders, etc. In conclusion, we believe that improving this aspect of language research by using corpus-based analysis and retrieval tools can result in better management of linguistic information and, consequently, better quality in terms of developing linguistic resources and tools that describe the specialized language of business in Romanian. In addition, to undertake a comparative analysis of the business language in Romanian and in English, the present study has also used Nelson's (2000) corpus-based research on Business English whenever we considered that a comparison between the two languages is edifying to highlight the specific features of business language in general and Romanian business language in particular. The texts from the British English Corpus (BEC) belong to various macro-genres such as university textbooks, media articles, various types of business documents, business correspondence.

Next, following Scott's (2012) approach, we used the Wordsmith Tools 6.0 software, the ROMBAC reference corpus, and the RBC study corpus to investigate and interpret the list of keywords (the frequency of the keywords and collocations4) found in the self made corpus of Romanian written business communication.

3.2. KEYWORDS

Keywords represent an essential feature for business communication analysis since they can provide a concise and precise high-level summarization of this specialized language. Corpus linguistics studies have focused on the concept of *frequency,* that is those words that are used most frequently are, in essence, the most important. Our approach is based on the concept of *keyword* defined by Scott as "a word that appears with an unusual frequency in a given text" (1997:236). Therefore, the concept of *keywords* has been applied to a wide range of discursive business genres. We view *keywords* are salient words in a corpus whose frequency is unusually high (*positive keywords*) or low (*negative keywords*) in comparison with a reference corpus.

In this regard, we used the Wordsmith Tools 6.0 program to statistically compare a smaller study corpus with a larger reference corpus, and calculate keywords, that is, those words that appear in the study corpus more often, or less frequently, than would be expected based on evidence from the reference corpus. Therefore, Wordsmith Tools 6.0 gave us the opportunity to explore the RBC study corpus based on user-customizable word lists, created from the inventory of the study corpus. There is also an option that allows the user to display all the shapes of a lexical unit present in the text, this can allow the identification of collocations, along with the frequency of their appearance in the text. The resulting keyword list represents a lexicon that is specific to a particular corpus. Specifically, our hypothesis states that the analysis of the list of keywords will highlight a lexical image of the business world.

The purpose of using *Wordsmith Tools 6.0* was to locate and identify keywords in a given text and, to do so, it compared the words in the text with a reference set of words taken from a large corpus of texts; any word which is found to be outstanding in its frequency in the text is considered 'key' and the 'keywords' are presented in order of 'outstandingness' (Scott, 2012:6). Thus, the *Wordsmith Tools 6.0* application compared statistically the RBC study corpus with the ROMBAC reference corpus and calculated *the keywords*, that is those words which appear in the study corpus more frequently, or more rarely, than it would be expected, taking into account the samples from the reference corpus. Consequently, *Wordsmith Tools 6.0 has* allowed us to explore the RBC study corpus on the basis of *word lists* (which can be personalized by the user). These lists of words are created starting from the corpus inventory used. There is also another option that allowed us to display all the forms of a lexical unit present in the text, thus allowing us to identify the collocations as well as their frequency in the text.

Keywords may thus be identified by comparing the words frequencies taken from a text or a corpus of texts with the words frequencies from a reference corpus, which has to be at least 5 times bigger (Berber Sardinha, 2004). For instance, in our analysis *COMPANIE* (engl. 'company') has an occurrence of 1.948 hits; even though it does not have a high frequency (being placed on the 40th position), it can be encountered on the first position in the keywords list as the RBC percent frequency (0.19%) is very high in comparison with that found in the ROMBAC reference corpus (0.01%). To identify a keyword, the program calculates both the word frequency in the RBC study corpus, as against the total number of words from this corpus, and the frequency of the same word in the reference corpus, contrasted with the total number of words from the ROMBAC corpus. Next, the program classifies the data obtained: applying one of the two statistical tests[5], the two corpora are compared and the *positive keywords* are retrieved, that is those words that occur in the RBC study corpus with an unusually high frequency, compared with the frequency from the ROMBAC reference corpus. The software also identifies the *negative keywords*, that is those words that occur with an unusually low frequency in the RBC study corpus. A mention needs to be made – the study focused on the analysis of the negative keywords only when they could provide more information about the positive keywords, which we consider as specific business terminology. Hence, the notion of keywords in this paper takes into account the positive keywords only.

Moreover, we fully agree with Scott's (2012) view that keywords can reveal the 'aboutness[6]' of a text. Thus, WordSmith Tools 6.0 allowed us to carry out the statistical analysis necessary to generate our own keyword list. According to M. Scott (2012), a word will be part of the list if this is unusually frequent, compared with what it is expected. In this context, an essential element is represented by the nature of the reference corpus that needs to be used for comparison reasons. Hence, we followed the procedure frequently adopted and supported by other analysts (Tribble, 2000; Scott, 2000; 2001b, 2002; Johnson et al, 2003) and we used the ROMBAC reference corpus for the Romanian language. As a result, we assert that the *keyword list* extracted from the RBC study corpus represents the lexicon characteristic for the specialized language of business in Romanian and its analysis would emphasize a lexical image of the Romanian business world.

3.3. COLLOCATIONS

This study has also examined written business communication using the notion of 'collocation' as we believe that we can master a specialized language if we can identify its specific collocations. Since in natural language words are not combined randomly into phrases and sentences, constrained only by the rules of syntax, we can claim that the ways in which they go together may be a significant source of information for business communication research. Collocations are defined as frequently recurrent combinations of commonly two linguistic elements which have a direct syntactic relationship, but whose co-occurrence in texts cannot be explained only by grammatical rules. In Scott's (2012) view, collocations are "words which occur in the neighborhood of your search word" and their analysis is important in working out characteristic lexical patterns "[…] by finding out which 'friends' words typically hang out with. It can be hard to see overall trends in your concordance lines, especially if there are lots of them. By examining collocations in this way you can see common lexical and grammatical patterns of co-occurrence." (2012:179).

We consider collocations co-occurrences of words (not necessarily adjacent) which follow two statistical criteria: a) the distance between words is relatively constant; b) they occur in the same contexts in a statistically significant high number. These criteria are evaluated using the Log-Likelihood score (Scott, 2012:Help Menu), which calculates the probability ratio of two statistical hypotheses which may be used to describe the text data collected through observation. The distribution to the open class words has also been restricted, i.e. nouns, adjectives, and verbs (excluding auxiliary verbs) and extracted the following types of collocations: noun-noun, noun-adjective/adjective-noun, and noun-verb/verb-noun from the RBC study corpus. Moreover, words were reduced to their lemmas (their canonical lexical form, not to their stem or root) to compute the term distribution in the corpus. Finally, to make computations more efficient, we scaled down the set of lemmas to those occurring at least 5 times in the corpus.

To extract the collocations, the source text has been first lemmatized; then, to set collocation horizons, a '*collocate window*' of 11 words (this is the context horizon in which co-occurrences may be considered) checks each sentence from the source text, so that each word becomes at a certain point the center of the 'collocate window'; the words introduced in this 'collocate window' are nouns or verbs only, while the other parts of speech have been ignored. The context horizons determine how far the program must look to left and right of the search word when checking whether the search criteria have been met. The default is 5.5 (5 to left and 5 to right of the search word) but we decided that a distance of 4 (left/right) is adequate to identify interesting pairs, in which only nouns and verbs can be found (for other types of collocations, we have taken into account nouns and adjectives or nouns solely).

All pairs of words (lemmata) that are formed between the center of the 'collocate window' and the other words, together with the distance between the words that form these pairs, have been introduced in the database. After searching the whole text, the Wordsmith Tools 6.0 program calculates the *mean* and the *dispersion* for each pair of words from the database, taking into account the occurrences at various distances. Dispersion represents the degree to which occurrences of a word are distributed throughout a corpus and a 'dispersion value' is the degree to which a set of values is uniformly spread (Scott, 2013:19). Since our study includes a quantitative analysis, it involves the frequency with which a word occurs in a corpus, representing a register or variety of Romanian business language.

4. RESULTS

Firstly, all the files from the RBC study corpus were introduced in Wordsmith Tools 6.0 to compile an extended word frequency list, which was saved as annex on a CD due to its large size. Similarly, we created a word frequency list of the ROMBAC reference corpus. After having

lemmatized both lists, we created a list of the first 500 lemmatized words from the RBC study corpus. To facilitate a further qualitative analysis, we extracted the first 100 words from this list as seen in annex 1 of this paper. To draw a first conclusion, in the list of the most frequent 100 words found in the RBC study corpus, only 17 words can be considered terms belonging to business language[7] (see table 4). The frequency of these 17 words extracted from the RBC study corpus totals 30.618 occurrences, representing only 2,95% of the entire corpus.

Table 4. Specific business terms from the 100 most common words list - RBC Study Corpus

No.	Term	Frequency in the RBC Study Corpus	Percentage in the RBC Study Corpus
28	PRODUS (engl. product)	2945	0.28%
31	SERVICIU (engl. service)	2451	0.24%
33	PIAȚĂ (engl. market)	2379	0.23%
34	FINANCIAR (engl. financial)	2332	0.23%
37	CONTRACT (engl. contract)	2123	0.20%
40	COMPANIE (engl. company)	1948	0.19%
43	VÂNZARE (engl. sale)	1801	0.17%
46	ECONOMIC (engl. economic)	1766	0.17%
50	AFACERE (engl. business)	1696	0.16%
51	CLIENT (engl. customer)	1680	0.16%
63	MUNCĂ (engl. work)	1463	0,14%
70	FIRMĂ (engl. firm)	1407	0.14%
72	ÎNTREPRINDERE (engl. enterprise)	1385	0.13%
75	COMERCIAL (engl. commercial)	1357	0.13%
77	PREȚ (engl. price)	1329	0.13%
82	COMERȚ (engl. commerce)	1283	0.12%
83	BANCĂ (engl. bank)	1273	0.12%

To highlight the most frequent terms used in written business communication in English and Romanian, we compared the list of words extracted from the RBC study corpus (see table 4) with the list of the most frequent English terms in the BEC study corpus. (see table 5). Thus, we found that all 7 terms in English appear in the list of terms specific to the business language in Romanian. More importantly, the RBC study corpus revealed a great number of terms specific to Romanian business language, displaying the following words on the first entries in the word list:

- 'financiar' (engl. financial), 'contract' (engl. contract), 'vânzare' (engl. sale), 'economic' (engl. economic), 'client' (engl. customer), 'comercial' (engl. commercial), 'comerț' (engl. commerce), 'bancă' (engl. bank).

As a particularity, the term 'întreprindere' (engl. entreprise) occurs near its hyponyms, i.e. 'companie' (engl. company) and 'firmă' (engl. firm). Also, for the 'întreprindere' (engl. entreprise) paradigm, the common sememe is 'unitate economică' (engl. economic unit), to which the following distinctive sememes are added '± superordinate' and '± size'. Other terms of the same paradigm may be 'concern' or 'organization'.

Table 5. Specific business terms from the most common 100 words list - BEC Corpus

No.	Term	Frequency in the BEC Corpus	Percentage in the BEC Corpus	Term lemmata
38	COMPANY	2934	0.29%	companies [1092]
41	BUSINESS	2837	0.28%	businesses [287]
54	MARKET	2336	0.23%	markets [469], marketing [469], marketed [10]
56	WORK	2234	0.22%	works [226], worked [134], working [680]
84	SERVICE	1461	0.14%	services [641], servicing [43], serviced [5]
89	PRODUCT	1385	0.14%	products [644]
94	PRICE	1302	0.13%	Prices [417], pricing [69], priced [20]

Taking these results into account, we can draw a partial conclusion as regards the frequency of words usage that the word frequency which belong to a specialized language is a very useful instrument/tool of analysis, but it does not represent the only criterion to identify a variety of any language. Thus, a more exact description of the terminology belonging to business language may be obtained by analyzing the 'keywords' which appear with a much higher frequency in written business communication in contrast with their occurrence in the general language. (Scott, 1999). Consequently, to extract the keywords, we used the two frequency lists: the list of the most frequent words from the RBC study corpus (named *study corpus wordlist)* and the list of the most frequent words from the ROMBAC reference corpus (named *reference corpus wordlist*), obtained by using the *WordList function* of the Wordsmith Tools 6.0 program. By using the *KeyWord function*, the program processed the keywords as mentioned below:

1. it calculated the ratio of the word frequency in the RBC study corpus;
2. it calculated the same ratio for the same word in the ROMBAC reference corpus;
3. it made these calculations for each word in the frequency list of the RBC study corpus and made concordances using the *LOG Likelihood* (Dunning, 1993) statistical criterion, which is considered to have a higher relevance, especially when comparing corpora of big dimensions (a mention should be made: the minimum number of occurrences of a keyword has been set to 23 occurrences).

The final result is a *keyword list* which is used in Romanian business communication statistically more frequently than in the general language. The Wordsmith Tools 6.0 program arranged the keywords identified in a certain order, according to their *keyness*. Thus, a term enters the list of keywords if it is unusually frequent (positive keyword) or unusually unfrequent (negative keyword) in the RBC study corpus, in comparison with the keyword list from the ROMBAC reference corpus (Scott, 2012, KeyWords Help File). Consequently, *positive keywords* are those words which have a degree of co-occurrence significantly higher than in the reference corpus, while *negative keywords* are those words with a degree of co-occurrence significantly lower than in the reference corpus.

In our analysis, we used the *positive keywords* category to highlight the specialized terms used in Romanian written business communication. Subsequently, the raw list was processed manually to eliminate grammatical words, proper nouns, numerals and numeral adjectives. The full list of keywords has been stored on an optical media-CD, and the list of the first 500 keywords in an appendix. In the diagram below, we summarized the process of extracting the keywords for written Romanian business communication:

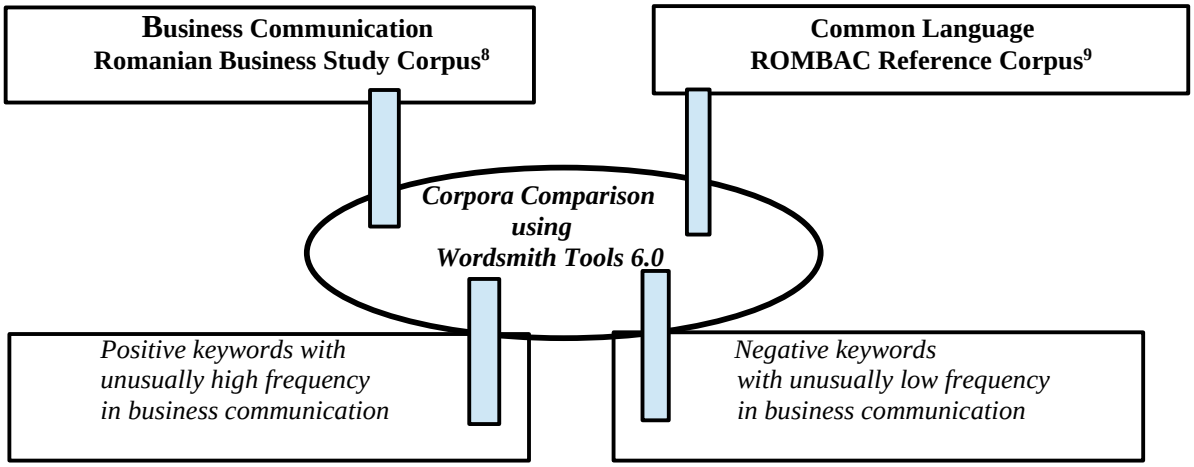

Diagram 1. Keywords for Romanian business language

To perform an in-depth quantitative analysis, the first 100 keywords specific to Romanian business language were extracted from the whole keywords list. After comparing the list of the first 500 keywords with the list of the most frequent 500 words from the RBC study corpus, we can draw the conclusion that these two lists are very different regarding the number of words and their contents. In the list of the most frequent 100 words, we identified only 17 terms which belong strictly to business language, while in the list of the first 100 keywords, we found 55 business terms. Table 6 shows a compressed keyword list, from which the non-specialized words have been taken out to highlight the business terms:

Table 6. Specific business terms extracted from the first 100 keyword list in Romanian

No	Keyword	Frequency in RBC	Percentage in RBC	Genres	Frequency in ROMBAC	Percentage in ROMBAC	Keyness
1	COMPANIE (engl. company)	1948,00	0.19%	8,00	7650,00	0.01%	6553,15
2	FINANCIAR (engl. financials)	2332,00	0.23%	8,00	17757,00	0.03%	5299,19
3	AFACERE (engl. business)	1696,00	0.16%	6,00	9112,00	0.01%	4825,20
4	MANAGER (engl. manager)	934,00	0.09%	8,00	1324,00		4643,62
5	CONTRACT (engl. contract)	2123,00	0.20%	8,00	17065,00	0.03%	4637,56
6	CLIENT (engl. customer)	1680,00	0.16%	8,00	12199,00	0.02%	3946,75
8	VÂNZARE (engl. sale)	1801,00	0.17%	8,00	15870,00	0.03%	3667,45
9	MANAGEMENT (engl. management)	1049,00	0.10%	8,00	4314,00		3448,16
10	SERVICIU (engl. job)	2451,00	0.24%	8,00	32701,00	0.05%	3417,55
11	ÎNTREPRINDERE (engl. company)	1385,00	0.13%	4,00	9670,00	0.02%	3342,21
13	ECONOMIC (engl. economic)	1766,00	0.17%	8,00	17431,00	0.03%	3277,83
14	COMERȚ (engl. trade)	1283,00	0.12%	7,00	8330,00	0.01%	3248,88
15	OFERTĂ engl. offer)	1165,00	0.11%	8,00	7450,00	0.01%	2979,03
16	MARKETING (engl. marketing)	626,00	0.06%	8,00	1074,00		2933,77

18	ACȚIONAR (engl. shareholder)	635,00	0.06%	5,00	1448,00		2695,57
19	PROFIT (engl. profit)	814,00	0.08%	8,00	3386,00		2660,17
20	PIAȚĂ (engl. market)	2379,00	0.23%	8,00	40004,00	0.06%	2526,74
21	FIRMĂ (engl. company)	1407,00	0.14%	8,00	14479,00	0.02%	2518,99
22	ANGAJAT (engl. employee)	918,00	0.09%	8,00	5246,00		2517,51
23	ECONOMIE (engl. economy)	876,00	0.08%	5,00	4807,00		2461,05
24	COMERCIAL (engl. commercial)	1357,00	0.13%	8,00	14468,00	0.02%	2354,40
26	PRODUS (engl. product)	2945,00	0.28%	8,00	62273,00	0.10%	2238,64
27	CHELTUIALĂ (engl. expenses)	1106,00	0.11%	5,00	10027,00	0.02%	2201,40
30	CAPITAL (engl. capital)	816,00	0.08%	8,00	4984,00		2148,09
31	BANCĂ (engl. bank)	1273,00	0.12%	8,00	14280,00	0.02%	2108,37
33	PLANIFICARE (engl. planning)	437,00	0.04%	7,00	873,00		1945,70
34	COST (engl. cost)	959,00	0.09%	8,00	8809,00	0.01%	1888,69
39	RESURSĂ (engl. resources)	729,00	0.07%	5,00	4966,00		1788,24
40	MUNCĂ (engl. work)	1463,00	0.14%	8,00	22888,00	0.04%	1701,01
43	RECESIUNE (engl. recession)	208,00	0.02%	4,00	19,00		1576,19
45	INVESTIȚIE (engl. investment)	899,00	0.09%	5,00	9590,00	0.02%	1558,76
48	BUSINESS (engl. business)	406,00	0.04%	7,00	1222,00		1542,41
50	DATORIE (engl. debt)	537,00	0.05%	4,00	2893,00		1524,87
51	NEVOIE (engl. need)	849,00	0.08%	8,00	9362,00	0,02%	1428,36
52	PROMOVARE (engl. promotion)	651,00	0.06%	8,00	5204,00		1427,21
53	BUGET (engl. budget)	744,00	0.07%	6,00	7315,00	0.01%	1384,98
54	ANTREPRENOR (engl. entrepreneur)	304,00	0.03%	3,00	577,00		1377,60
57	MANAGERIAL (engl. managerial)	257,00	0,02%	3,00	289,00		1362,73
59	MARFĂ (engl. goods)	696,00	0.07%	8,00	6951,00	0.01%	1278,45
60	PRODUCȚIE (engl. production)	952,00	0.09%	8,00	13166,00	0.02%	1276,11
63	VENIT (engl. income)	960,00	0.09%	7,00	14020,00	0.02%	1210,55
65	FURNIZOR (engl. provider)	510,00	0.05%	7,00	3829,00		1169,94
66	REVÂNZĂTOR (engl. reseller)	141,00	0,01%	2,00	0,00		1156,58
67	CONCURENȚĂ (engl. competition)	397,00	0.04%	7,00	2184,00		1113,54

73	SALARIAT (engl. employee)	447,00	0.04%	3,00	3278,00		1042,52
75	PORTOFOLIU (engl. portfolio)	271,00	0.03%	6,00	939,00		967,14
77	COMANDĂ (engl. order)	488,00	0.05%	7,00	4583,00		943,74
79	CONSUMATOR (engl. consumer)	524,00	0.05%	4,00	5436,00		931,22
80	FRANCIZĂ (engl. franchise)	139,00	0.01%	6,00	50,00		923,45
85	PRESTATOR (engl. provider)	263,00	0.03%	2,00	1012,00		893,22
87	CREDIT (engl. credit)	808,00	0.08%	8,00	13194,00	0.02%	890,49
88	ACȚIUNE (engl. share)	1065,00	0.10%	8,00	21238,00	0.03%	887,89
89	VOUCHER (engl. voucher)	128,00	0.01%	5,00	35,00		881,54
91	PUBLICITATE (engl. advertising)	339,00	0.03%	8,00	2133,00		875,69
92	ACTIVE (engl. assets)	655,00	0.06%	8,00	9102,00	0.01%	873,28

We also noticed that the contents of the two lists is different: while the list of the most frequent terms contains numerous grammatical words (i.e., '*and', 'but', 'so'*) and an extremely reduced number of notional words ('company', 'year', 'business', 'market', 'produce', 'product', 'price', 'system'), the keyword list includes almost entirely notional words. The above mentioned differences between the two lists are synthesized in table 7:

Table 7. Frequent words – keywords comparison

Most Frequent 100 Word List	First 100 'Keywords' List
17 words belonging exclusively to the business language	55 words belonging exclusively to the business language
Higher number of grammatical words	Smaller number of grammatical words
Very small number of notional words	Extremely numerous notional words

Thus, the list of keywords generated by the RBC study corpus can be considered a set of specialized terms, in close relation to activities from the business world. Therefore, we claim that these terms, which have a very high frequency compared to the general language, are part of the core vocabulary of the Romanian business language. However, we need to mention that this set of terms for the specialized language of business represent the result of our research. Another study may highlight a slightly different set of terms compared to ours. However, we consider that the general semantic homogeneity of the list of keywords extracted from the RBC study corpus reflects the lexicon of the Romanian business language with a very high degree of accuracy. A mention should also be made: as regards the demarcation of the business lexis versus the non-business lexis, keywords are regarded in terms of *tendency* and not so much as *absolute*. Our analysis reveals that the keywords identified using the two corpora (study and reference) and the Wordsmith Tools 6.0 program, have a tendency to be used predominantly in the business environment, compared to other words.

The specific business terms from the list of the first 100 English keywords extracted from the BEC study corpus can be found in annex 3. The analysis of keywords reveals that the world of business is clearly marked by the lexicon used in this type of specialized language. Comparing the results of the quantitative analysis of the keywords in Romanian and in English, we notice a very close similarity between the terms in the two languages. More specifically, 32 of the 49 English language specific terms in the list of the first 100 keywords are found in the list of the first 100

keywords in Romanian. This demonstrates that there is a common business vocabulary specific to both languages analyzed, Romanian and English. In conclusion, in compiling an updated Romanian-English glossary of business terms and collocations, the terms introduced were selected by taking into account both the quantitative analysis undertaken in this study and the comparative analysis of Romanian and English keywords from the two study corpora, RBC and BEC.

5. CONCLUSION

The present study described the process of elaborating a corpus of specialized language from the Romanian business context in order to analyze Romanian business terminology both qualitatively and quantitatively. The paper has highlighted the advantages of using new analytical techniques, which consist in identifying linguistic models which encode textual meanings that cannot be detected intuitively. Thus, the analysis has focused on the lexicon with the highest frequency in business language and how they function and encode meanings in written business communication. The lexicon has been analyzed in terms of keywords, that is those words that are statistically significantly more numerous in the corpus under study than in the reference corpus. The paper showed that the keywords found using the Wordsmith Tools 6.0 program and two corpora (the RBC study corpus and the ROMBAC reference corpus, respectively) have a tendency to be mainly used in written business communication, compared to other words from the common language. The keyword list generated by the RBC study corpus may be regarded as a set of specialized terms, in close connection with the business world. Therefore, we can state that these terms, that have a very high frequency in comparison with the common language, may constitute the core vocabulary of Romanian written business communication, which is clearly marked by the lexis used in this type of specialized language. Also, the analysis of keywords has showed that the business world revolves around recurring semantic sets: people, institutions, places and money.

All in all, it can be stated that the lexicon is, to a large extent, made up of a limited number of lexico-semantic structures that create a 'world' of meanings in business communication. This 'world' is a world of practical actions in relation to concrete entities, concerned with various modes of communication, populated by business people, companies, institutions, hierarchy, money, events, businesses and marked by a dynamic and non-emotional lexicon. All in all, the ultimate goal of our research is to build an updated glossary of business terms and collocations, where we include a selection of terms extracted from the quantitative and qualitative analyses of the study corpora used.

FOOTNOTES

1. Terminologie' -Vocabulaire particulier utilisé dans un domaine de la connaissance ou un domaine professionnel; ensemble structuré en termes. La terminologie de la médecine. 2. Etude systématique des „termes" ou mots et syntagmes spéciaux servant à dénommer classes d'objets et concepts; principes généraux qui président à cette étude [1, p. 2234]., Le Nouveau Petit Robert, Dictionnaire alphabétique et analogique de la langue française, Paris, 1997.
2. The Academic Word List (AWL) contains 570 word families which were selected according to principles and was primarily made so that it could be used by teachers as part of a programme preparing learners for tertiary level study or used by students working alone to learn the words most needed to study at tertiary institutions. For details on the development and evaluation of the AWL, see Coxhead, Averil (2000) A New Academic Word List. TESOL Quarterly, 34(2): 213-238.
3. The British National Corpus (BNC) was originally created by Oxford University press in the 1980s - early 1990s, and it contains 100 million words of text texts from a wide range of genres (e.g. spoken, fiction, magazines, newspapers, and academic).
4. Collocation is defined as "the habitual juxtaposition of a particular word with another word or words with a frequency greater than chance" (Oxford Dictionary).
5. The "Chi-square" test or Ted Dunning's "Log Likelihood" test (see Scott 2013: Help Menu, Tribble, 2000: 79-80).
6. Notion attributed to Phillips (1989).
7. The 17 terms belonging to the business language are: produs (engl. product), serviciu (engl. service), piață (engl. market), financiar (engl. financiary), contract (engl. contract), companie (engl. company), vânzare (engl. sales), economic (engl. economic), afacere (engl. business), client (engl. customer), muncă (engl.

work), firmă (engl. company), întreprindere (engl. company), comercial (engl. commercial), preț (engl. price), comerț (engl. commerce), bancă (engl. bank).

8. The self-compiled Romanian Business Corpus (RBC study corpus) contains 1.036.341 running words.
9. The Romanian Balanced Annotated Corpus (ROMBAC reference corpus) is made of approx. 61,093,390 running words.

REFERENCES

[1] Baker, P. 2009. *Contemporary Corpus Linguistics*. London: Continuum.

[2] Berber Sardinha, T. 2004. *Linguística de corpus*. São Paulo: Manole.

[3] Biber, D. Representativeness in Corpus Design, http://otipl.philol.msu.ru/media/biber930.pdf

[4] Biber, D. (1993). Squibs and discussions. Co-occurrence patterns among collocations: a tool for corpus-based lexical knowledge acquisition. Computational Linguistics, 19:3, 531-538.

[5] Bidu-Vrănceanu, A. 2007. *Lexicul specializat în mişcare. De la dicţionare la texte.* Editura Universităţii din Bucureşti.

[6] Cabré, M. T. 1999. *Terminology: Theory, Methods, and Applications*. John Benjamins Publishing Company.

[7] Chung, Y.M. & Lee, J.Y. (2001). *A corpus-based approach to comparative evaluation of statistical term association measures*. Journal of the American Society for Information Science and Technology, Vol. 52, No. 4, pp. 283-296.

[8] Church, K.W. & Hanks, P. (1990). *Word Association Norms, Mutual Information and Lexicography*. Proceedings of 27th Association for Computational Linguistics (ACL), Vol. 16, No. 1, pp. 22-29.

[9] DSL – Dicţionar de ştiinţe ale limbii, Ed. a 2-a, Bidu-Vrănceanu, A., Călăraşu, C., Ionescu- Ruxăndoiu, L, , Mancaş, M., Pană Dindelegan, G, , Ed. Nemira, Bucureşti, 2001-2005.

[10] Evert, S. & Krenn, B. 2001. Methods for the qualitative evaluation of lexical association measures. Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pp. 188-195.

[11] Evert, S. 2005. *The statistics of Word Co-occurrences: Word Pairs and Collocations*. Ph.D. thesis. Stuttgart: University of Stuttgart.

[12] Evert, S. 2006. *How random is a corpus?* http://www.stefan-evert.de/PUB/Evert2006.pdf

[13] Evert, S. 2009. Corpora and collocations. In A. Lüdeling & M. Kyto (Eds.), Corpus Linguistics: An International Handbook, vol.2 (pp. 1212- 1248). Berlin/New York: Mouton de Gruyter.

[14] Gries, S. Th. 2008. *Phraseology and linguistic theory. A brief survey*. In S. Granger & F. Meunier (Eds.), Phraseology: An Interdisciplinary Perspective (pp. 3-25). Amsterdam: John Benjamins Publishing.

[15] Gries, S. Th. 2013. *50 something years of work on collocations. What is or should be next....* International Journal of Corpus Linguistics, 18:1, 137-165.

[16] Gries, S. Th. 2014. *Frequencies, probabilities, association measures in usage-/exemplar-based linguistics: some necessary clarifications*. In N. Gisborne & W. Hollman (Eds.), Theory and Data in Cognitive Linguistics (pp. 15-48). Amsterdam/Philadelphia: John Benjamins Publishing Company.

[17] Ion, R., Irimia, E., Ştefănescu, D., Tufiş, D. 2012. *ROMBAC: The Romanian Balanced Annotated Corpus.* Research Institute for Artificial Intelligence - Romanian Academy.

[18] ISO 704: 2009, *Terminology work - Principles and methods.*

[19] Lerat, P. 1995. *Les langues spécialisées*, Presses Universitaires de France, Paris.

[20] McEnery, T., Wilson, A. 1996. *Corpus Linguistics*. Edinburgh University Press.

[21] Nelson, M. 2000. A Corpus-based Study of the lexis of Business English and Business English Teaching Materials. PhD Thesis. Manchester: University of Manchester.
[22] Pickett, D. 1988. *English in Business: Knowing and Acting*. In Holden, S. (ed) Language and Literature. MEP.
[23] Pârlog, H., Teleagă, M. 2000. *Dicționar englez-român de colocații verbale*, Ed. Polirom, Iași.
[24] Scott, M. 2012. *WordSmith Tools version 6*. Liverpool: Lexical Analysis Software
[25] Sinclair, J. McH. 2004. *Trust the Text: Language Corpus and Discourse*. London: Routledge.
[26] Stubbs, M. 1993. British Traditions in Text Analysis: From Firth to Sinclair. In Baker, M., Francis, G. & Tognini-Bonelli, E. (eds) *Text and Technology. In Honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins.
[27] Stubbs, M. 1995. Corpus Evidence for Norms of Lexical Collocations. In Cook, G. & Seidlhofer, B. (eds) *Principle & Practice in Applied Linguistics. Studies in Honour of H.G. Widdowson*. Oxford: Oxford University Press.
[28] Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.
[29] Tutin, A. 2008. *For an extended definition of lexical collocations.* Proceedings of the XIII Euralex International Congress, Barcelona, Spain, pp. 1453-1460.
[30] Wartena, C., Brussee, R. 2010. *Keyword Extraction Using Word Co-occurrence*, Conference: Database and Expert Systems Applications (DEXA).