Corpus-Driven Analysis of Multi-Word Terms Including the Word *'Risk'* in English, French and Lithuanian

OKSANA SMIRNOVA
Mykolas Romeris University, Lithuania

SIGITA RACKEVIČIENĖ
Mykolas Romeris University, Lithuania

KEYWORDS: descriptive terminology, corpus driven analysis, financial terms, term extraction, term formation patterns

1. INTRODUCTION

Computational technologies have opened new possibilities to linguistic research: digital corpora and corpus analysis software facilitated access to the deepest language structures and relations among different language units and to reveal peculiarities of their usage in different domains over different periods of time. Therefore, corpus-driven analysis of natural language is applied nowadays in a broad range of linguistic areas: lexicography, language teaching, development of machine translation, compilation of linguistic databases, etc.

Terminology research has also become to a large extent corpus-driven. Digital corpora allow terminologists to work with a big amount of documents, observe particular features of a specialized language, collect information about real usage of terms and their evolution, capture new terms which could not be intuitively felt or predicted as well as to carry out contrastive analysis of data of several languages. Corpus analysis software, automatic and semiautomatic term extractions tools also contribute to terminology standardisation, for example, the frequency count of synonyms can provide useful distributional evidence indicating statistically preferred terms (Khurshid, Rogers 1992: 36). Thus, digital corpora enable terminologists to revise terminographical information about terms in existing databases and constantly update them.

According to M. Teresa Cabré, M. Amor Montané and Rogelio Nazar, the goal of modern terminology is to produce formal, semantic and functional descriptions of lexical units having terminological value as well as to explain their relation with the rest of the units of the linguistic system. The object of terminology research is "the *living* terminology" (terminology that naturally occurs in specialized texts) and the communicative aspect of the use of terms which is best instantiated in a corpus (Cabré, Montané, Nazar 2012).

The aim, object and objectives of the research. The aim of the research is to apply the methodology of corpus linguistics for extraction and formal structure analysis of financial multi-word terms including the word '*risk*' as the head noun in English, French and Lithuanian.

In order to achieve this aim, the following objectives were set:

1) to analyse the principles of descriptive corpus driven terminology including the methods of collocational-colligational analyses;
2) to compile corpora of the EU legal acts of financial domain in three languages (English, French and Lithuanian) and select the software appropriate for the corpus-driven research;
3) to extract the most frequent words from the corpora in the investigated languages and select the most frequent keyword (noun) for the further analysis;
4) to carry out collocational analysis of the selected keyword in the English corpus and extract multi-word terms including the selected keyword as the semantic and syntactic head of terms from the English corpus material;
5) to establish French and Lithuanian equivalents of the selected English terms in the parallel English-French-Lithuanian corpus;
6) to perform formal structure quantitative analysis of the selected multi-word terms and determine which modification patterns and syntactic structures of the terms are predominant in the investigated languages.

Data and scope of the research. For the purposes of the research, four corpora of the EU documents of financial domain were compiled: three monolingual corpora (English, French and Lithuanian) and one parallel corpus (EN-FR-LT). The sizes of the corpora are as follows: EN 802 933 words, FR 940 655 words, LT 639 279 words. In total, 210 financial terms including the word *'risk'* as the head noun were extracted from the corpora: 70 English terms and their equivalents in French and Lithuanian. The choice of word *'risk'* was determined by the corpus data which revealed that this word was the most frequent in the selected EU documents.

2. THEORETICAL PRINCIPLES OF THE RESEARCH

2.1. Prescriptive and descriptive terminology management

Evolution of computational technologies for natural language research clearly differentiated the traditional terminology from the modern one. Terminologists' attitude towards the conception of a term, term sources, standardisation of terms and other terminology issues has changed considerably. Two directions of terminology research and management have been distinguished: *prescriptive terminology* and *descriptive terminology* (Zeller 2005; Bielinskienė et al. 2015).

Advocates of prescriptive terminology focus on terminology standardisation based on terminology dictionaries, databases, lists of approved terminology and documents on standardisation principles. In prescriptive terminology, conception of a term is based on the place of the concept, described by it, in the conceptual hierarchical system of the domain and its relationship with other concepts (Pearson 1998: 10; Marcinkevičienė 2000: 6). According to the principles of prescriptive terminology, one term should be used to describe one concept as such reciprocity reduces probability of ambiguity, facilitates communication, and, simultaneously, development of conceptual hierarchical system of a domain.

Descriptive terminologists distance themselves from the strict attitude towards term conception, term unambiguity, development of hierarchical conceptual system and standardisation. In descriptive terminology, contrary to prescriptive one, a context plays a vital role in analysis of a term, and a term is assumed as a lexical unit dependent on its context. Their goal is not to form a term according to certain principles, but to record its usage, variety of its forms in different texts and describe its peculiarities thus forming a basis for its standardisation (Bielinskienė et al. 2015: 10–11).

Development of technologies gave rise to huge flow of information presented in various types of texts, and, simultaneously, constantly growing number of terms used in them. In this ever-changing reality, terminologists are no longer able to control rapid development of terminology as terms change faster than they are processed. Therefore, management of terminology requires more flexible approach based on description rather than on prescription methodology. Focus has gone from standardisation of terminology to the analysis of terminology in corpora (Bielinskienė et al. 2015: 11).

Descriptive terminology led to development of the *Communicative Theory of Terminology* which sets focus on the living terminology that is used in specialized discourses and puts the emphasis on the communicative aspect of the use of terms (Cabré, Montané, Nazar 2012). According to this theory, the main role of the terms is to communicate expert knowledge; thus they are conceived "as a three-fold polyhedron having a cognitive component (the concept), a linguistic component (the term) and a communicative one (the situation)" (Cabré, Montané, Nazar 2012: 1).

Terminological studies based on the communicative theory are not only interested in terminology established by standards or found in official databases, but also (and particularly) in those terms which are actually used in texts of language for specific purposes. Thus, the communicative theory "not only adopts an in vitro approach, but is also interested in terms in vivo" (Cabré, Montané, Nazar 2012: 1–2). Research of the living terminology discloses that the traditional notion of term univocity (one-to-one correspondence between a concept and particular terms in different languages) is not well-founded in real language. Variation (synonymy, ambiguity, periphrases, redundancy) is characteristic not only of general language units, but also of terminology; thus concepts and terms have to be studied "in their dynamic interplay" (Cabré, Montané, Nazar 2012: 2–3).

The focus on the usage of terms has made corpora the main workspace of modern terminology analysis. Digital corpora and corpus analysis software have enhanced terminology with rich resources and tools which enable terminologists to extract terms from a big amount of documents, capture the newest changes in terminology of a specific domain, analyse its evolution as well as establishing conceptual interconnection among terms of the same domain. Corpora allow to get reliable, statistically based, previously totally unavailable information about term usage in natural language. Therefore, *corpus-driven methodology* has become the prevailing methodology providing indispensable toolkit for terminology research and management (Zeller 2005; Cabré, Montané, Nazar 2012; Bielinskienė et al. 2015).

2.2. Corpus driven analysis principles: collocational and colligational methods

Corpus driven terminology extraction and analysis are performed using statistical and linguistic methods. In the given research collocational and colligational methods are applied.

Collocations refer to syntagmatic attraction between lexical units. According to Tomas Lehecka (2015), “the concept of collocation is based on the notion that each word in a language prefers certain lexical contexts over others, i.e. that any given word tends to co-occur with certain words more often than it does with others” (Lehecka 2015: 2). Statistical analysis of corpus data is used to measure the degree of attraction between words, its results enable to determine which word combinations appear together significantly more often than it would be expected by chance given the words’ total frequency in the corpus (Lehecka 2015: 2). In this way the most significant collocations of the chosen words in the analysed corpus are established.

This method is used mostly for contextual semantic analysis of lexical units as it enables to observe the whole variety of co-occurring words of the investigated words, establish the predominant co-occurrence patterns and thus to describe their meanings based on their contextual environment (Atkins, Rundell, Sato 2003: 340–341).

Collocational analysis has become an indispensable tool for lexicographers, it is also extensively applied in computational linguistics for the purposes of machine translation, natural language processing and other areas (Lehecka 2015).

Collocational method is often used in combination with colligational method. The concept of colligation refers to attraction of a lexical unit and a grammatical pattern and is based on the notion that words prefer to be used in certain grammatical patterns and avoid other grammatical patterns (Sinclair 1998; Lehecka 2015). Colligational analysis may been extended and encompass the relationship between the lexical unit and the position in a phrase, clause, sentence, text or discourse where the lexical unit can be used (Hoey 2005: 49–52).

Colligational analysis has been extensively employed in combination with collocational analysis to study the meanings of near-synonyms as corpus linguistics studies have revealed that they are often used in different lexical and grammatical contexts (Lehecka 2015). Corpus linguistics investigations have shown that even different senses of polysemous words may have different collocational and colligational patterns (Aston, Burnard 1998).

Thus collocational and colligational analyses have proved that semantics and grammar should not be treated as independent, but rather as closely interconnected systems (Lehecka 2015). Pragmatic aspect should also be taken into consideration in the analyses as collocational and colligational preferences of a lexical unit vary significantly between different domains and different types of texts (Butler 2004: 157; Newman, Rice 2006).

2.3. Application of collocational-colligational method in terminology extraction and analysis

Collocational-colligational method may also be used for extraction of multi-word terms; it is especially appropriate for extraction of terminology including pre-chosen keywords – words of potential terminological relevance characteristic of the domain that the corpus belongs to. This methodology is based on the assumption that complex terms are made of existing simple terms (Nakagawa 2001).

Corpus analysis tools extract co-occurring words of the pre-chosen keywords and enable to establish the dominant collocations. A part of these collocations is noun phrases which have to be selected from collocation lists using linguistic methods. The selected noun phrases are used for further analysis of the corpus data and extraction of multi-word terms consisting of the selected noun phrases and potentially of additional collocates.

In the given research, the extracted terms are analysed further using colligational analysis principles seeking to reveal formal structure models of terminology in the investigated languages and to contribute to contrastive multilingual studies of term formation patterns (cf. Janulevičienė, Rackevičienė 2014; Mockienė 2016).

3. DEVELOPMENT OF CORPORA AND SELECTION OF THE SOFTWARE FOR THE ANALYSIS

According to M. Teresa Cabré, M. Amor Montané and Rogelio Nazar, the purpose of compiling a corpus of documents of a specific domain is threefold. Firstly, terminologists need to become familiar with the type of language of the domain; secondly, a corpus is needed to perform terminology extraction and conduct statistical analyses of the terms; and finally, texts are used to obtain complementary information about the terms such as semantic, syntactic or collocational clues (Cabré, Montané, Nazar 2012: 3–4).

The compiled corpus should be of sufficient size and quality to be considered representative of the chosen domain. There are no precise requirements for the size of a corpus; but it should contain as many documents as possible because the bigger it is, the more reliable conclusions about terminology usage in the domain can be made (Cabré, Montané, Nazar 2012: 4).

Taking into consideration all these aspects, the corpora of the EU legal acts in English, French and Lithuanian were compiled. The legal acts were downloaded from the Official Journal of the European Union which is a freely viewed online source. 101 legal acts, regarding financial issues of the EU and enacted in the period 2014–2017, were collected. The documents were transformed into plain text and aligned for extraction of terms and their analysis.

Three programs were used to compile, align the texts and extract the necessary data from them: *AntConc* (2014), *AntPConc* (2017) created and certified by Laurence Anthony and *NOVA Text Aligner* (2014).

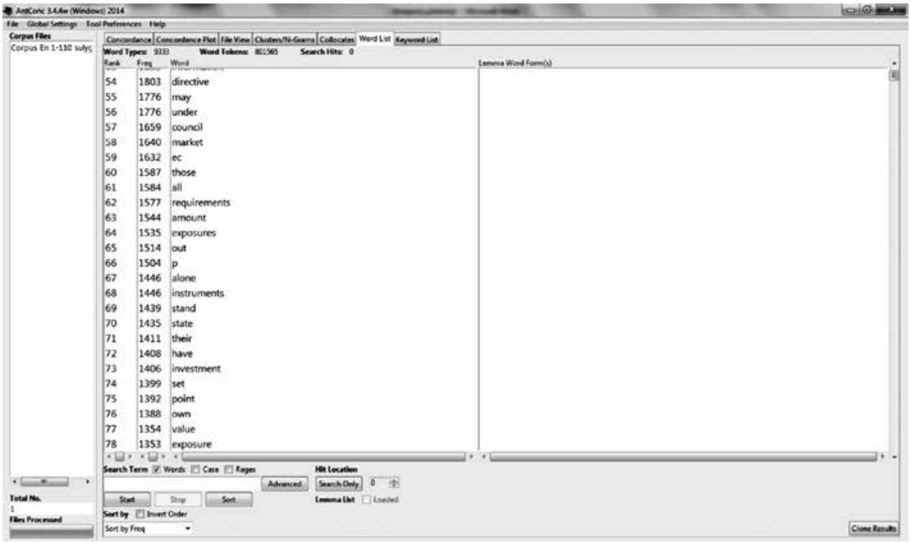
AntConc is a freeware multiplatform toolkit for carrying out corpus linguistic research and data-driven learning, while *AntPConc* is intended for creating a parallel corpus of several languages. Both programs enable the researcher to deal with a big amount of data and carry out a comprehensive multilingual linguistic analysis.

The tools which are provided for users by *AntConc* program are the following: *Word list, Collocates, Clusters, Keywords, Concordance, Concordance plot, File view tool.* In the given research four tools of *AntConc* were applied:

- *Wordlist* enabled to determine the most frequent words in the corpora;
- *Collocates* enabled to determine the dominant collocates of the word *'risk'* and measure the degree of their syntagmatic attraction to the word *'risk'*;
- *Clusters* allowed to disclose the most dominant multi-word terms including the word *'risk'* as the head noun;
- *Concordances* gave a wide variety of samples illustrating the usage of the terms in the texts.

AntConc program's snapshot is presented in Figure 1.

Figure 1. AntConc program



AntConc 3.4.4w (Windows) 2014
File Global Settings Tool Preferences Help
Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Corpus Files: Corpus En 1-110 sulyc

Word Types: 9333 Word Tokens: 801565 Search Hits: 0

Rank	Freq	Word
54	1803	directive
55	1776	may
56	1776	under
57	1659	council
58	1640	market
59	1632	ec
60	1587	those
61	1584	all
62	1577	requirements
63	1544	amount
64	1535	exposures
65	1514	out
66	1504	p
67	1446	alone
68	1446	instruments
69	1439	stand
70	1435	state
71	1411	their
72	1408	have
73	1406	investment
74	1399	set
75	1392	point
76	1388	own
77	1354	value
78	1353	exposure

Lemma Word Form(s)

Search Term ☑ Words □ Case □ Regex

Hit Location: 0

Lemma List □ Loaded

Total No. 1

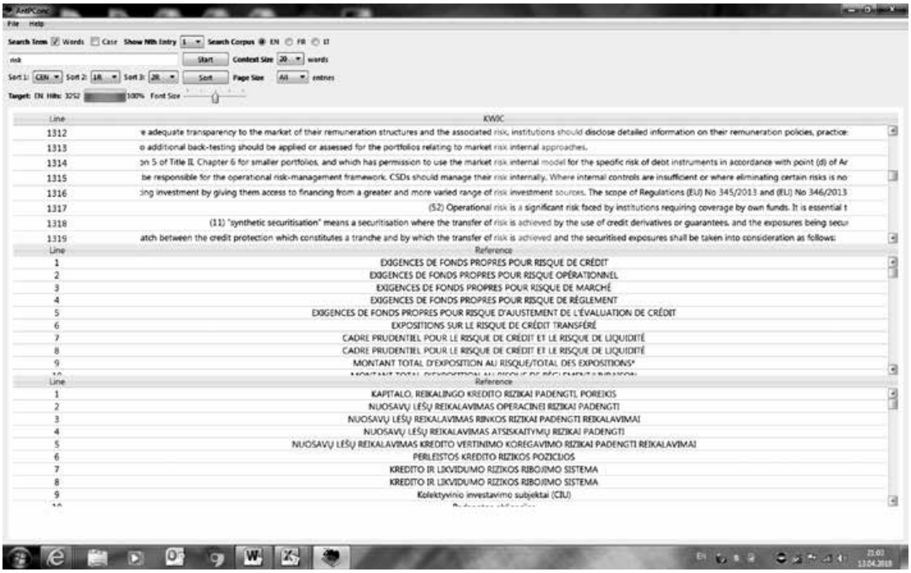
Files Processed

Sort by □ Invert Order

Sort by Freq

Before compiling a parallel corpus, the downloaded texts in three languages were aligned manually with the help of the program NOVA Text Aligner. After alignment of the texts, a parallel corpus was built up using the AntPConc program (see Figure 2).

Figure 2. AntPConc program



Search Term ☑ Words □ Case Show Nth Entry 1 Search Corpus ● EN ○ FR ○ LT

risk

Context Size 20 words

Sort 1: CEN Sort 2: 1R Sort 3: 2R Page Size All entries

Target: EN Hits: 3252 100% Font Size

Line	KWIC
1312	e adequate transparency to the market of their remuneration structures and the associated risk, institutions should disclose detailed information on their remuneration policies, practice
1313	o additional back-testing should be applied or assessed for the portfolios relating to market risk internal approaches.
1314	on 5 of Title II, Chapter 6 for smaller portfolios, and which has permission to use the market risk internal model for the specific risk of debt instruments in accordance with point (d) of Ar
1315	be responsible for the operational risk-management framework. CSDs should manage their risk internally. Where internal controls are insufficient or where eliminating certain risks is no
1316	ing investment by giving them access to financing from a greater and more varied range of risk investment sources. The scope of Regulations (EU) No 345/2013 and (EU) No 346/2013
1317	(52) Operational risk is a significant risk faced by institutions requiring coverage by own funds. It is essential t
1318	(11) "synthetic securitisation" means a securitisation where the transfer of risk is achieved by the use of credit derivatives or guarantees, and the exposures being secu
1319	atch between the credit protection which constitutes a tranche and by which the transfer of risk is achieved and the securitised exposures shall be taken into consideration as follows:

Line	Reference
1	EXIGENCES DE FONDS PROPRES POUR RISQUE DE CRÉDIT
2	EXIGENCES DE FONDS PROPRES POUR RISQUE OPÉRATIONNEL
3	EXIGENCES DE FONDS PROPRES POUR RISQUE DE MARCHÉ
4	EXIGENCES DE FONDS PROPRES POUR RISQUE DE RÈGLEMENT
5	EXIGENCES DE FONDS PROPRES POUR RISQUE D'AJUSTEMENT DE L'ÉVALUATION DE CRÉDIT
6	EXPOSITIONS SUR LE RISQUE DE CRÉDIT TRANSFÉRÉ
7	CADRE PRUDENTIEL POUR LE RISQUE DE CRÉDIT ET LE RISQUE DE LIQUIDITÉ
8	CADRE PRUDENTIEL POUR LE RISQUE DE CRÉDIT ET LE RISQUE DE LIQUIDITÉ
9	MONTANT TOTAL D'EXPOSITION AU RISQUE/TOTAL DES EXPOSITIONS*

Line	Reference
1	KAPITALO, REIKALINGO KREDITO RIZIKAI PADENGTI, POREIKIS
2	NUOSAVŲ LĖŠŲ REIKALAVIMAS OPERACINEI RIZIKAI PADENGTI
3	NUOSAVŲ LĖŠŲ REIKALAVIMAS RINKOS RIZIKAI PADENGTI REIKALAVIMAI
4	NUOSAVŲ LĖŠŲ REIKALAVIMAS ATSISKAITYMŲ RIZIKAI PADENGTI
5	NUOSAVŲ LĖŠŲ REIKALAVIMAS KREDITO VERTINIMO KOREGAVIMO RIZIKAI PADENGTI REIKALAVIMAI
6	PERLEISTOS KREDITO RIZIKOS POZICIJOS
7	KREDITO IR LIKVIDUMO RIZIKOS RIBOJIMO SISTEMA
8	KREDITO IR LIKVIDUMO RIZIKOS RIBOJIMO SISTEMA
9	Kolektyvinio investavimo subjektai (CIU)

The AntPConc program allowed to expose the examples from three languages at the same time: one could click on a chosen sentence to see the equivalents in other languages. The programme also enabled to distinguish the keywords and their left or right collocates by use of different colours. However, it was not possible to see the document from which the example was extracted. Using the parallel corpus, French and Lithuanian equivalents of 70 English terms were extracted.

4. CORPUS ANALYSIS AND TERM EXTRACTION USING THE TOOLS OF THE PROGRAMS AntConc AND AntPConc

4.1. Establishment of the most frequent words in the corpora

In order to determine the most frequent words in the corpus of the financial documents, the tool *Word list* of the program AntConc was used. It provided the word frequency results in the English, French and Lithuanian corpora which enabled to compare the word frequencies in the investigated languages and develop a trilingual list of TOP 10 most frequent words (see Table 1). *Word list* also provided information on how many word tokens and word types there were in the corpora and gave the access to the context in which the terms were used (see Table 1).

Table 1. TOP 10 most frequent words in the corpora

	EN (corpus 802 933 word tokens, 9333 word types)	**FR** (corpus 940 655 word tokens, 12365 word types)	**LT** (corpus 639 279, 24727 word types)
1.	*risk* (3252), *risks* (592)	*risque* (2855), *risques* (997)	*rizikos* (2549), *riziką* (1063), *rizika* (536), *rizikai* (355), *rizikas* (4)
2.	*credit* (3183), *credits* (18)	*credit* (3307), *credits* (222)	*kredito* (3103), *kreditų* (52), *kreditu* (29), *kreditą* (28), *kreditas* (8), *kreditai* (6), *kreditais* (4), *kreditus* (3), *kreditams* (2)
3.	*market* (1642), *markets* (424)	*marché* (1631), *marches* (484)	*rinkos* (1355), *rinkų* (310), *rinka* (229), *rinkoje* (226), *rinką* (76), *rinkose* (75), *rinkai* (51), *rinkoms* (18), *rinkas* (7), *rinkomis* (2)

	EN (corpus 802 933 word tokens, 9333 word types)	**FR** (corpus 940 655 word tokens, 12365 word types)	**LT** (corpus 639 279, 24727 word types)
4.	*amount* (1545)	*montant* (1672)	*suma* (1133), *sumos* (684)
5.	*instruments* (1448)	*instruments* (1489)	*priemonės* (1542), *priemonių* (1393), *priemones* (649)
6.	*value* (1356)	*valeur* (1754)	*vertės* (664), *vertė* (599)
7.	*securities* (1275)	*titres* (1229), *titrisation* (503)	*vertybinių popierių* (1095), *vertybinio poprieriaus* (691)
8.	*services* (1275)	*services* (1778)	*paslaugų* (1225), *paslaugas* (485)
9.	*capital* (1093)	*capital* (660)	*kapitalo* (1143)
10.	*funds* (1087)	*fonds* (2773)	*lėšų* (877)

The findings of this analysis reveal that the most frequent words in the corpora of all three investigated languages are *'risk'* and *'credit'* while the word *'market'* takes the third position in the frequency lists. All three words, and in particular the words *'risk'* and *'credit'* often go together in the financial documents. This could be explained by the tight semantic relation between the words: needless to say that any money transaction implies danger, in other words, *'credit'* generates *'risk'*. The most usual grammatical number of the words is singular though plural is also used in the investigated languages. In Lithuanian, which is a synthetic language with a rich inflectional system, the words are used in different grammatical cases, the dominant of which is the Genitive singular.

The findings of the analysis allow to state that the words *'risk'*, *'credit'* and *'market'* are to be assessed as the lexical items denoting the fundamental concepts of the financial domain. The word *'risk'*, which is the most frequent in the corpora, was chosen for further term extraction and analysis.

Since the object of the research is English multi-word terms including the selected keyword (word 'risk') and their equivalents in French and Lithuanian, the subsequent work of term extraction was organised in the following stages: establishment of the dominant English collocations of the word *'risk'*; extraction of the noun phrases with the word *'risk'* as the head noun, selection of the phrases having terminological value and establishment of their French and Lithuanian equivalents.

4.2. Establishment of the collocates of the word *'risk'* in the English corpus

In the second stage of term extraction, the dominant collocations of the word *'risk'* in the English corpus were established. This objective was pursued with the help of the AntConc tool *Collocates.* This tool provided left and right collocates of the chosen keyword and allowed to analyse the non-sequential patterns in the languages.

The tool provided total frequencies of collocates as well as their frequencies on the left and on right of the word *'risk'*. It also provided the values of statistical measures (mutual information (MI) scores) which showed the degree of syntagmatic attraction between the keyword and its collocates. The results of the collocational analysis are presented in the Table 2 which provides the exhaustive data about the collocates with the highest MI scores (top 10).

Table 2. TOP 10 collocates of the word *'risk'* in the English corpus

	Total frequency	Freq.(L)	Freq.(R)	MI score
1. dilution	51	37	14	8.1
2. mitigation	100	9	91	8.01
3. weights	65	4	61	7.7
4. systemic	120	119	1	7.7
5. profile	87	1	86	7.48
6. low	68	62	6	7.2
7. assigned	127	95	32	6.89
8. exposure	491	71	420	6.4
9. specific	177	148	29	6.3
10. operational	129	114	15	5.99

The findings reveal the closest lexical context of the word *'risk'* in the English language and allow to envisage the dominant two-word nucleus of multi-word terms including the word *'risk'* in the investigated corpus.

4.3. Extraction of noun phrases including the word *'risk'* as the head noun in the English corpus

Further corpus analysis and term extraction focused on collocations of certain formal structure – the noun phrases including the word *'risk'* as the head noun. They were extracted from the English corpus using the tools *Clusters* and *Concordance.*

The tool *Clusters* enabled to search for the clusters (word combinations) including the word *'risk'* with its left and right collocates. It allowed to select the minimum and the maximum length (number of words) of the clusters and the minimum frequency of the clusters displayed. The ordered clusters could be displayed either according to their frequency or to the number of files in which the clusters appeared in the corpora.

The following parameters were set for extraction of noun phrases including the word *'risk'* as the head noun: cluster size from 2 to 5, sorting by frequency. The position of the keyword also had to be selected: firstly clusters with the keyword on the right and secondly clusters with the keyword on the left were ordered. Under the parameter 'search keyword on right', the tool displayed 4814 cluster tokens and 181 cluster types; under the parameter 'search keyword on left' 4284 cluster tokens and 137 cluster types.

Out of the displayed cluster lists, noun phrases including the word *'risk'* as the head noun were extracted manually. Their wider context was analysed using the tool *Concordance* which enabled to determine the boundaries of the noun phrases and the noun phrases including both left and right collocates of the word *'risk'*. The Table 3 presents the most frequent noun phrases.

Table 3. TOP 20 noun phrases including the word *'risk'* as the head noun

Noun phrases ('risk' on the right)	Freq.	Noun phrases ('risk' on the left)	Freq.
1. credit risk	471	*1. risk of excessive leverage*	18
2. systemic risk	113	*2. risk of debt instrument*	12
3. operational risk	102	*3. risk of loss*	12
4. liquidity risk	97	*4. risk of capacity withholding*	10

Noun phrases ('risk' on the right)	Freq.	Noun phrases ('risk' on the left)	Freq.
5. specific risk	86	*5. risk to third parties*	8
6. market risk	76	*6. risk to purchased receivable*	7
7. counterparty credit risk	41	*7. risk of a price change*	7
8. dilution risk	33	*8. risk of disruption*	6
9. interest rate risk	32	*9. risk to financial stability*	2
10. business risk	26	*10. risk of a borrower*	2

The results reveal that the noun phrases including the word *'risk'* on the most right position are much more frequent than the noun phrases including the word *'risk'* on the most left position. Only some examples of the noun phrases including both left and right collocates of the word *'risk'* were found in the corpus: *credit risk of repurchased transactions, credit risk of securization position, credit risk to third parties.*

A list of TOP 70 English noun phrases including the word *'risk'* as the head noun was made for the further research. Only those noun phrases which had terminological value (denoted abstract concepts of financial domain) were included in the list. In the further analysis, they are called multi-word terms.

4.4. Establishment of French and Lithuanian equivalents of the selected English terms

In the final stage of term extraction, the French and Lithuanian equivalents of the selected English terms were established. This objective was pursued using AntPConc software developed for analysis of parallel corpora. The corpora of three investigated languages were uploaded to the program. As the selected English terms were searched in the English corpus, the program displayed the parallel strings of the French and Lithuanian texts and thus enabled to select manually the necessary equivalents. In total, 70 French and 70 Lithuanian equivalents of the selected English terms were established. Table 4 presents TOP 10 English terms and their French and Lithuanian equivalents.

The findings of this analysis revealed the formal differences of the terms in the investigated languages which are investigated and discussed below.

Table 4. TOP 10 English terms and their equivalents in French and Lithuanian

EN	FR	LT
credit risk	*risque de crédit*	*kredito rizika*
systemic risk	*risque systémique*	*sisteminė rizika*
operational risk	*risque opérationel*	*operacinė rizika*
liquidity risk	*risque de liquidité*	*likvidumo rizika*
specific risk	*risque spécifique*	*specifinė rizika*
market risk	*risque de marché*	*rinkos rizika*
counterparty credit risk	*risque de crédit de conterpartie*	*sandorio šalies kredito rizika*
dilution risk	*risque de dilution*	*gautinų sumų rizika*
interest rate risk	*risque de taux d'intérêt*	*palūkanų normos rizika*
business risk	*risque économique*	*verslo rizika*

5. FORMAL STRUCTURE ANALYSIS OF THE EXTRACTED TERMS

The extracted terms were analysed further using colligational analysis principles seeking to reveal formal structure models of terminology in different languages.

The constituents of the terms are of two main categories: the head noun '*risk*' and its modifiers. According to positions of modifiers, two modification patterns were established – prenominal modification (in which modifiers take the place before the head noun) and postnominal modification (in which modifiers take the place after the head noun). No terms including both prenominal and postnominal modifiers were included in the TOP 70 list of the terms selected for the statistical analysis.

The analysis focused on the following aspects of formation of the terms: number of constituents of the terms, modification patterns and syntactic structures (position and word classes of modifiers) of the terms. Tables 5 and 6 present the summarised findings of the analysis.

The findings presented in the tables reveal that the terms have different modification patterns in the investigated languages. The quantitative modification analysis was performed to identify the dominant modification patterns across the investigated languages; its results are presented in Diagram 1.

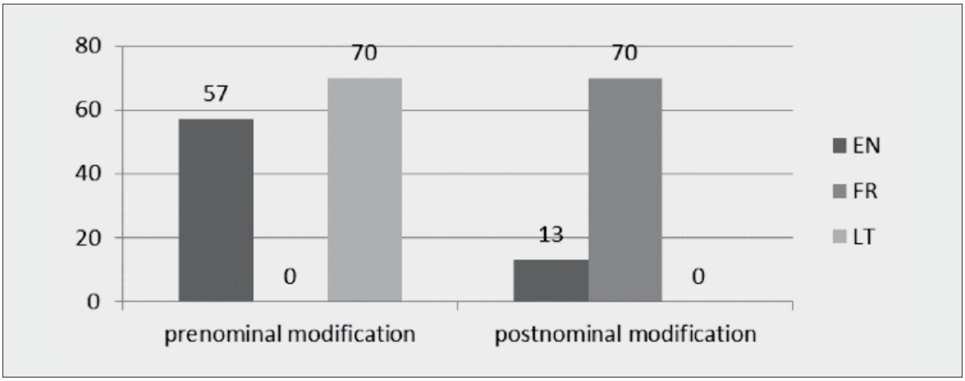
Table 5. Syntactic structures of the *prenominal modification* patterns of the terms

Syntactic structures	EN	FR	LT
N + *'risk'*	**19 terms, e.g.**: *credit risk,* *concentration risk,* *liquidity risk,* *business risk.*	–	**18 terms, e.g.**: *skolininko rizika,* *sandorių rizika,* *saugojimo rizika.*
A + *'risk'*	**18 terms, e.g.**: *entrepreneurial risk,* *internal risk,* *potential risk.*	–	**17 terms, e.g.**: *sisteminė rizika,* *operacinė rizika,* *investicinė rizika*
N + N + *'risk'*	**3 terms**: *credit and liquidity risk,* *interest rate risk,* *counterparty credit risk*	–	**9 terms, e.g.**: *dienos kredito rizika,* *palūkanų normos rizika,* *sandorio šalies rizika*
A + N + *'risk'*	**12 terms, e.g.**: *significant credit risk,* *specific credit risk,* *foreign exchange risk.*	–	**9 terms, e.g.**: *maža kredito rizika,* *specifinė kredito rizika,* *gautinų sumų rizika*
A/Num + N + + N + *'risk'*	**1 term**: *minimal credit and market risk*	–	**12 terms, e.g.**: *minimali kredito ir rinkos rizika,* *vienos nakties likvidumo rizika,* *vienos nakties kredito rizika*
A + A + N + + *'risk'*	**4 terms, e.g.**: *specific and general credit risk,* *intraday and overnight credit risk,* *intraday and overnight liquidity risk*	–	**5 terms, e.g.**: *specifinė klaidingų sprendimų rizika,* *įsigytų gautinų sumų rizika,* *bendra ir specifinė kredito rizika*

Table 6. Syntactic structures of the *postnominal modification* patterns of the terms

Syntatic structures	EN	FR	LT
‘*risk*’ + PP	**13 terms, e.g.**: *risk of the institution,* *risk to financial system,* *risk of loss*	**22 terms, e.g.:** *risque de crédit,* *risque de concentration,* *risque de modèle*	–
‘*risk*’ + A	–	**20 terms, e.g.:** *risque intrajournalier,* *risque interne,* *risque spécifique*	–
‘*risk*’ + A + PP	–	**4 terms, e.g.**: *risque spécifique de corrélation,* *risque significatif de corrélation ,* *risque générale de corrélation*	–
‘*risk*’ + PP + A	–	**15 terms, e.g.:** *risque de crédit spécifique* *risque de crédit quotidien,* *risque sur matière première*	–
‘*risk*’ + PP + A + + PP	–	**9 terms, e.g.:** *risque de crédit intrajournalier à 24h,* *risque de liquidité intrajournalier à 24h*	–

Diagram 1. Modification patterns of the terms



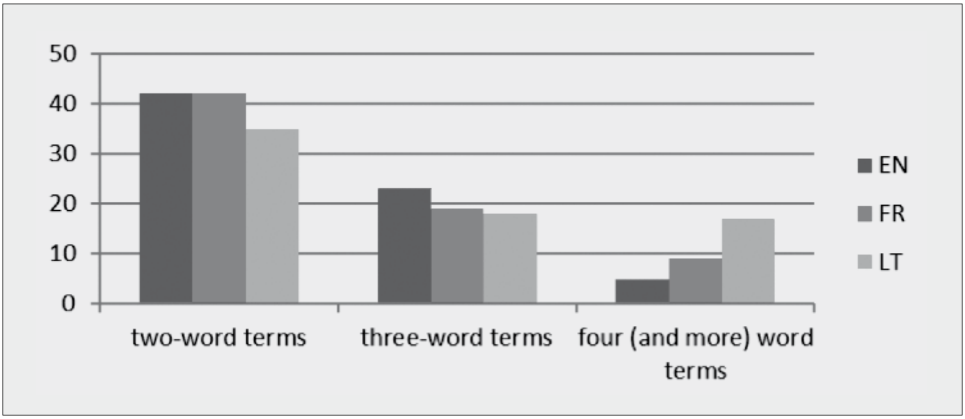
The results show that prenominal modification is dominant in the English and Lithuanian languages, while postnominal modification is characteristic of the French language. No terms of prenominal modification pattern were found in French, and no terms of postnominal modification patterns were detected in Lithuanian. Only English terms were of both types though the number of terms with postnominal modifiers is rather low.

All investigated terms include one or several modifiers which are nouns, adjectives or prepositional phrases taking different positions in the terminological units. The results of the analysis reveal that the dominant modifiers of the English and the Lithuanian terms are nouns and adjectives while in the French terms the word *'risk'* is mostly modified by prepositional phrases. The terms including 3 and more words may be further classified according to their modification levels (terms including modifiers modifying the head noun and terms including modifiers modifying other modifiers), but, due to limited space, this analysis is not presented in this paper.

The summarised findings in the tables also reveal that the terms differ in the number of their constituents. The quantitative analysis of the term length was performed to establish the dominant number of term constituents in the investigated languages; its results are presented in the Diagram 2.

The diagram reveals two main tendencies. Firstly, the EU term developers respect the main requirement of language economy (brevity of terms): two-word terms are prevalent in all three languages. Secondly, only a few English and French terms have more than 2-3 words while in

Diagram 2. Number of term constituents



Lithuanian terms including 4 and more words constitute a significant part of the selected data (17 of 70).

The tendency of prevalence of two-word terms coincides with the tendencies established by other terminology researchers. The research on automatic extraction and definition of education and science terminology by Agnė Bielinskienė et al. revealed that most Lithuanian terms of this domain are two-word terms: they constituted more than two-thirds of the terms selected from the term candidates automatically extracted from a specialised corpus (Bielinskienė et al. 2015: 62). The contrastive research on constitutional law terminology by Liudmila Mockienė revealed the same tendency in three different languages. In the investigated English, Russian and Lithuanian legal acts of a constitutional nature, the majority of the extracted multi-word terms consisted of two constituents: they constituted 78.5% of the English multi-word terms, 62% of the Russian multi-word terms and 74.5% of the Lithuanian multi-word terms (Mockienė 2016: 43–45). Thus, developers of terms of different domains and different languages tend to adhere to the same principle of language economy and user-friendliness.

6. CONCLUSIONS

The software AntCont, used for monolingual corpora analysis, and AntPConc, used for parallel corpus analysis, allowed to perform extraction of English, French and Lithuanian multi-word terms including the word ‘*risk*’ as the head noun from the specialized corpora of the EU documents of financial domain compiled for the purposes of the research. The extraction was performed in the following stages:

- extracting the keywords of the English, French and Lithuanian corpora (the words of potential terminological relevance characteristic of the domain that the corpora represent) and choosing the most frequent keyword (the word ‘*risk*’) for further analysis;
- extracting collocations of the word ‘*risk*’ and noun phrases including the word ‘*risk*’ as the head noun with left and/or right collocates from the English corpus;
- selecting lexical units which have terminological value from the list of the extracted noun phrases;
- establishing French and Lithuanian equivalents of the selected English terms in the parallel English-French-Lithuanian corpus.

The applied methodology proved to be suitable for effective multilingual extraction of terminology which can be used for development/updating of termbases or scientific analysis of terms.

Formal structure analysis of the extracted terms revealed some major term formation tendencies in the investigated languages:

- prenominal modification is dominant in the English and Lithuanian languages, while postnominal modification is characteristic of the French language;
- the dominant modifiers of the English and the Lithuanian terms are nouns and adjectives while in the French terms the word *'risk'* is mostly modified by prepositional phrases;
- the prevalent term type according to the number of constituents is two-word terms – they constitute the biggest number of the terms in the TOP 70 lists in English, French and Lithuanian; that shows that the EU term developers respect the main requirement of language economy (brevity of terms);
- only a few English and French terms have more than 2-3 words while in the Lithuanian TOP 70 list terms including 4 and more words constitute a significant part of the selected data (17 of 70).

The findings of the formal structure analysis disclose term formation trends in the investigated languages and provide terminological information which might be useful for term developers and translators. Syntactic patterns, established in the research, may be used for development of automatic linguistic methods of term extraction without any pre-chosen keywords.

REFERENCES

Aston G., Burnard L. 1998: *The BNC Handbook. Exploring the British National Corpus with SARA*, Edinburgh: Edinburgh University Press.

Atkins S., Rundell M., Sato H. 2003: The contribution of FrameNet to practical lexicography. – *International Journal of Lexicography* 16(3), 333–357.

Bielinskienė A., Boizou L, Grigonytė G., Kovalevskaitė J., Rimkutė E., Utka A. 2015: *Lietuvių kalbos terminų automatinis atpažinimas ir apibrėžimas*. Monografija, Kaunas: Vytauto Didžiojo universitetas.

Butler C. S. 2004: Corpus studies and functional linguistic theories. – *Functions of Language* 11(2), 147–186.

Cabré M. T., Montané M. A., Nazar R. 2012: *Corpus-based Terminology Processing. TKE 2012 Tutorial.* Available at http://terminus.iula.upf.edu/tke2012/.

Hoey M. 2005: *Lexical Priming: A New Theory of Words and Language*, London: Routledge.

Janulevičienė V., Rackevičienė S. 2014: Formation of criminal law terms in English, Lithuanian and Norwegian. – *LSP journal – Language for special purposes, professional communication, knowledge management and cognition* 15(1), 4–20.

Lehecka T. 2015: Collocation and colligation. – *Handbook of Pragmatics Online.* J.-O. Östman, & J. Verschueren (Eds.), Amsterdam: John Benjamins Publishing Company, 1–23.

Khurshid A., Rogers M. 1992: Terminology management: a corpus-based approach. – *Translating and the Computer* 14, 33–44.

Marcinkevičienė R. 2000: Terminografija ir tekstynas. – *Terminologija* 6, 5–23.

Mockienė L. 2016: *Formation of terminology of constitutional law in English, Lithuanian and Russian.* Doctoral Thesis, Vilnius: Mykolas Romeris University.

Nakagawa H. 2001: Experimental evaluation of ranking and selection methods in term extraction. – *Recent Advances in Computational Terminology*. D. Bourigault, Ch. Jacquenim and M.-C. L'Homme (Eds.), Amsterdam/Philadelphia: John Benjamins Publishing Company, 303–325.

Newman J., Rice S. 2006: Transitivity schemas of English EAT and DRINK in the BNC. – *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. S.T. Gries and A. Stefanowitsch (Eds.), Berlin/New York: Mouton de Gruyter, 225–260.

Pearson J. 1998: *Terms in Context*, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Sinclair J. 1998: The lexical item. – *Contrastive Lexical Semantics*. E. Weigand (Ed.), Amsterdam: John Benjamins Publishing Company, 1–24.

Zeller I. 2005: *Automatinis terminų atpažinimas ir apdorojimas.* Daktaro disertacija, Kaunas: Vytauto Didžiojo universitetas, Vilnius: Lietuvių kalbos institutas.

SOURCES

Official Journal of European Union: https://eur-lex.europa.eu/oj/direct-access.html

COMPUTER SOTFWARE USED FOR THE ANALYSIS

Anthony L. 2014: AntConc (Version 3.4.4w) [Computer Software]. Tokyo, Japan: Waseda University. Available at http://www.antlab.sci.waseda.ac.jp

Anthony L. 2017: AntPConc (Version 1.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available at http://www.antlab.sci.waseda.ac.jp

Supernova-soft. NOVA Text Aligner. Available at https://nova-text-aligner.soft112.com/.

ANGLŲ, PRANCŪZŲ IR LIETUVIŲ KALBŲ DAUGIAŽODŽIŲ TERMINŲ SU ŽODŽIU *RIZIKA* ANALIZĖ TEKSTYNŲ LINGVISTIKOS METODAIS

Straipsnyje pristatomi deskriptyviosios terminologijos tyrimo principai bei empirinis daugiažodžių terminų su žodžiu *rizika* tyrimas, kurio tikslas – taikant tekstynų lingvistikos metodus, surinkti terminus iš ES finansų srities dokumentų tekstynų ir atlikti jų formaliosios sandaros analizę.

Tyrimo tikslams buvo sukaupti keturi tekstynai: finansų srities dokumentų anglų kalba (802 933 žodžiai), prancūzų kalba (940 655 žodžiai) ir lietuvių kalba (639 279 žodžiai) bei lygiagretusis anglų–prancūzų–lietuvių kalbų tekstynas. Iš tekstynų surinkta 210 terminų, kuriuose žodis *rizika* eina pagrindiniu dėmeniu: 70 angliškų terminų ir po tiek pat jų atitikmenų prancūzų ir lietuvių kalbomis. Žodžio *rizika* pasirinkimą lėmė tai, kad šis žodis buvo dažniausias visų trijų kalbų tekstynuose.

Terminų atpažinimui ir surinkimui buvo naudojamos dvi kompiuterinės programos – AntConc ir AntPConc. Dirbta tokiais etapais:

- dažniausių žodžių, galinčių būti terminų branduoliu, angliškame, prancūziškame ir lietuviškame tekstynuose nustatymas ir vieno iš jų (žodžio *rizika*) atrinkimas tolesnei analizei;

- žodžio *rizika* kolokacijų ir daiktavardinių junginių su pagrindiniu dėmeniu *rizika* ir jo kairiaisiais bei dešiniaisiais kolokatais nustatymas angliškame tekstyne;
- daiktavardinių junginių, laikytinų daugiažodžiais terminais, atrinkimas;
- atrinktų angliškų terminų prancūziškų ir lietuviškų atitikmenų nustatymas.

Pritaikyta metodologija leido rezultatyviai surinkti daugiažodžius terminus iš daugiakalbių tekstynų. Tai duoda pagrindą teigti, kad ji gali būti taikoma terminų kaupimui bei tyrimams.

Surinktų terminų formaliosios sandaros analizė atskleidė keletą svarbių terminų darybos tendencijų tiriamose kalbose:

- vyraujantis terminų tipas pagal dėmenų skaičių visose trijose tiriamose kalbose yra dvižodžiai terminai; tai rodo, kad ES terminų kūrėjai laikosi kalbos ekonomijos principo ir stengiasi kurti kuo trumpesnius daugiažodžius terminus;
- tik keletas angliškų ir prancūziškų terminų turi daugiau kaip 2–3 dėmenis; tuo tarpu lietuviški terminai, susidedantys iš 4 ir daugiau dėmenų, sudaro beveik ketvirtadalį surinktų terminų;
- anglų ir lietuvių kalbų terminų darybos modeliuose vyrauja prepozicinė ir postpozicinė modifikacija, o prancūzų kalbos – postpozicinė modifikacija;
- daugumos anglų ir lietuvių kalbų terminų priklausomieji dėmenys yra daiktavardžiai ir būdvardžiai, o prancūzų kalboje – prielinksninės konstrukcijos.

Formaliosios sandaros analizės rezultatai suteikia informacijos, kuri gali būti naudinga terminų kūrėjams ir vertėjams. Tyrimo metu nustatyti sintaksinių struktūrų modeliai gali būti taikomi, kuriant kompiuterinius lingvistinius metodus automatiniam terminų atpažinimui be iš anksto pasirinktų raktinių žodžių.

Gauta 2018-05-21

Oksana Smirnova
Mykolo Romerio universitetas
Ateities g. 20, LT-08303 Vilnius
E. paštas osmirnova@mruni.eu

Sigita Rackevičienė
Mykolo Romerio universitetas
Ateities g. 20, LT-08303 Vilnius
E. paštas sigita.rackeviciene@mruni.eu