Corpus-based Linguistic Analysis of Business English Report Writing Papers by L2 English Language Speakers

Michaella Duruttya

Abstract: The author compiled a machine analyzable corpus from the Business English tests of writing from the academic years 2020/2021 and 2021/2022. The referred texts were uploaded into Sketch Engine, an online tool used for the analysis of authentic texts with large amounts of words (text corpora) to identify patterns, keywords, and terms in language often used by L2 speakers of the English language. This text-processing tool can demonstrate how a particular language works by presenting typical combinations, synonyms, phrases, examples of use, and context, it can extract keywords and terms, and it can look up translations, among other of its features.

The investigation focused on analyzing the most frequently used words and phrases by students in the first to the second years of business and economics studies by examining the selected words' collocational and grammatical behavior. The investigated texts were written by students who were all non-native English speakers. The results of this analysis compared their output with similar examples found in a larger reference corpus, thus seeking similarities and differences, ultimately leading to ways of improving the students' production and writing style of business English texts.

Key words: corpus linguistics, business English, business reports, sketch engine, keywords, terms, n-grams

Introduction

Language users never choose words randomly, and language is essentially non-random. When we look at linguistic phenomena in corpora, where there is enough data, we can discover relationships between two or more phenomena, which are nonrandom, and therefore we do not find arbitrary associations. Language is not random because we speak or write with a clear purpose. (Kilgariff, Language is never ever random, 2005)

This study seeks to analyze non-native students' English for economics and business writing styles; the findings of the study are based on the analysis of a report writing examination written by students at the Prague University of Economics and Business. The students attend a business English seminar every week where they learn suitable communicative skills, vocabulary, and topic-related grammar focused on terms connected to business and trade. The written tests in question here consist of a report writing task, wherein the students are presented with

a selection of charts in which they are asked to interpret, describe the data, and suggest the subsequent courses of action.

Sketch Engine is the tool used to compile the corpus of written examination outputs.

The following table shows the ten most frequently used words in the Business Reports corpus.

Tab. 1: *The most frequently used words*

	Item	Frequency
1.	the	12,576
2.	in	6,129
3.	of	5,797
4.	be	5,744
5.	to	5,292
6.	and	3,704
7.	a	2,667
8.	year	2,066
9.	from	1,741
10.	number	1,675

The wordlist shown in Table 1 is a frequency list generated by the corpus query tool, presenting nouns, verbs, adjectives, and other parts of speech. It is also possible to obtain information about frequency, frequency per million, and average reduced frequency (a modified frequency that prevents the result to be influenced by a certain part of the corpus, e.g., one or more documents containing higher concentrations of a certain token).

The general wordlist can give us an overview of the most frequently used words; however, looking at Table 1, it is clear the most frequently occurring items are understandably articles, prepositions, conjunctions, etc., which are quite common in any type of text or corpus, thus not characterizing the corpus itself. To obtain a clearer picture of the character of the texts wherein the typical words are used, it is more practical to narrow the wordlist down to gain insight into the frequently used individual parts of speech.

The most frequently occurring nouns in the focus corpus

After narrowing the wordlist down with the primary focus on the occurrence of nouns, the following result is shown as demonstrated in the next Table 2 displaying the ten most frequently used nouns.

Tab. 2: *The most frequently used nouns*

	Item	Frequency	Relative frequency
1.	year	2066	11941.98945
2.	number	1673	9670.35254
3.	sale	1520	8785.97481
4.	month	1341	7751.31067
5.	rate	1221	7057.68108
6.	people	1073	6202.20459
7.	unemployment	1022	5907.41201
8.	book	995	5751.34535
9.	tourist	994	5745.56511
10.	report	728	4208.01951

The frequency (or absolute frequency) refers to the number of occurrences or hits of a particular item and presents an absolute figure.

The relative frequency, or frequency per million, is the number of occurrences of an item per million tokens, i.e., the smallest unit that a corpus consists of, which are words and nonwords. Due to the fact that we are investigating a "Business Reports" corpus consisting of texts written by students of economy and business, it is reasonable to focus on vocabulary linked to the related topic.

Selected business vocabulary words from the above-mentioned frequency lists in an expanded context:

- These *sales* are illustrated in the line graph above on a monthly basis.
- ... the book sales *rate* fluctuated from January to October.
- In addition, probably due to Christmas, we reached a *peak* at sales in December.
- Our sales manager requested this *report* to get an overview and analysis of the sales situation.

The most frequently occurring verbs in the focus corpus

After narrowing the wordlist down with the primary focus on the occurrence of verbs, the following result is shown as demonstrated in the next table 3 displaying the ten most frequently used verbs.

Selected business vocabulary words from the above-mentioned frequency lists in an expanded context:

Tab. 3: *The most frequently used verbs*

	Item	Frequency	Relative frequency
1.	be	5744	33201.73639
2.	have	864	4994.13305
3.	reach	783	4525.93308
4.	see	611	3531.73066
5.	start	550	3179.13562
6.	increase	512	2959.48625
7.	show	495	2861.22206
8.	rise	419	2421.92332
9.	sell	358	2069.32828
10.	decrease	306	1768.75546

- After this moderate fall, the number of sold books *increased* sharply during May and June and continued to rise steadily.
- After that, the trend was volatile and has been *declining* again since 2014.
- Overall, book sales were *fluctuating* during the year.
- The purpose of this report is to *analyze* the changes of sales in different parts of the year.

The most frequently occurring adjectives in the focus corpus

After narrowing the wordlist down with the primary focus on the occurrence of adjectives, the following result is shown as demonstrated in the next Table 4 displaying the ten most frequently used adjectives.

Tab. 4: *The most frequently used adjectives*

	Item	Frequency	Relative frequency
1.	more	454	2624.23195
2.	high	408	2358.34061
3.	young	372	2150.25173
4.	low	342	1976.84433
5.	good	222	1283.21474
6.	foreign	218	1260.09376
7.	first	217	1254.31351
8.	other	217	1254.31351
9.	next	210	1213.85178
10.	significant	198	1144.48882

Selected frequently used adjectives from the corpus in context:

- It shows some significant inconsistency in our companies' *monthly* sales

- There was a steady *upward* trend in the sales rate from April to August
- The *average* number of people coming to the Czech Republic remains consistent until February
- The *sharp* increase also continued between April and May

The most frequently occurring adverbs in the focus corpus

After narrowing the wordlist down with the primary focus on the occurrence of adverbs, the following result is shown as demonstrated in the next Table 5 displaying the ten most frequently used adverbs.

Tab. 5: *The most frequently used adverbs*

	Item	Frequency	Relative frequency
1.	again	312	1803.43693
2.	however	275	1589.56781
3.	also	268	1549.10608
4.	slightly	237	1369.91844
5.	more	213	1231.19252
6.	almost	185	1069.34562
7.	so	184	1063.56537
8.	significantly	175	1011.54315
9.	only	174	1005.76291
10.	again	312	1803.43693

Some notable frequently occurring adverbs in the corpus as shown in context:

- The rate continued to fall *slightly,* dropping to around 10% in 2016.
- After this growth, sales fell *dramatically*
- Youth unemployment rate from 15 to 25 years of age is *relatively* high.
- In summer, before the school year starts, the situation gets *better*.

The most frequently occurring n-grams in the focus corpus

Another area of interest worth investigating is the exploration of the most frequently used n-grams within the corpus. N-grams are continuous sequences of items from a given sample of text or speech. N-grams can be used in probability, communication theory, and statistical natural language processing. N-grams can be also called multi-word expressions or lexical bundles, and they are composed of tokens. Investigating these can shed more light on how certain lexical bundles are preferentially used by the candidates whose written works are being explored.

After narrowing the wordlist down and filtering out the occurrence of n-grams, the following result is shown as demonstrated in the Table 6 displaying the thirty (to provide a wider overview) most frequently used n-grams.

Tab. 6: *The most frequently used n-grams*

	Item	Frequency
1.	the number of	623
2.	of the year	455
3.	the Czech Republic	449
4.	the unemployment rate	348
5.	the end of	339
6.	in the Czech	319
7.	there was a	306
8.	in the Czech Republic	285
9.	we can see	271
10.	the beginning of	266
11.	of this report	238
12.	this report is	220
13.	report is to	216
14.	of this report is	210
15.	this report is to	202
16.	sales of books	182
17.	number of asylum	174
18.	in the EU	143
19.	end of the	142
20.	number of tourists	140
21.	beginning of the	138
22.	of people from	136
23.	of asylum seekers	136
24.	the end of the	135
25.	years of age	134
26.	the beginning of the	131
27.	The purpose of	129
28.	aim of this	126
29.	aim of this report	123
30.	purpose of this	121

However, upon looking at the presented results in Table 6, it is clearly visible that some n-grams are parts of larger lexical bundles, e.g., “aim of this” and “aim of this report”; thus, it is advisable to nest the n-grams which are sub-n-grams of another longer n-gram which will be grouped together with the longer n-gram. This leads to further narrowing down the corpus into grouped n-grams, the results of which are shown below (in italics)

Tab. 7: *The most frequently used nested n-grams*

Item	Frequency
the number of	623
of the year	455
• *the Czech Republic*	449
• *in the Czech Republic*	285
• *in the Czech*	319
the unemployment rate	348
• *the end of*	339
• *the end of the*	135
• *end of the*	142
there was a	306
we can see	271
the beginning of	266
• *of this report*	238
• *this report is*	220
• *report is to*	216
• *of this report is*	210
sales of books	182
number of asylum	174
in the EU	143
number of tourists	140
beginning of the	138
of people from	136
of asylum seekers	136
years of age	134

One can contrast the Business Reports corpus with a reference corpus. The business subcorpus of the English Web 2020 (enTenTen20), hereinafter referred to as the "reference corpus", was chosen as the best appropriate reference corpus (among the recommended similar corpora) following the recommendation and careful consideration as well as comparison with various monolingual English corpora.

The task is to identify what (if anything) is unique. The key data identified are the following:

Keywords – individual words (any token can be included). The focus corpus uses keywords more often than the reference corpus does, and vice versa. Any token that appears more frequently in the focus corpus can be considered a keyword. the reference corpus follows. As a result of the similarity in the frequency of other parts of speech across all texts, the final product will actually consist mostly of nouns and adjectives.

Terms – key multi-word expressions in a format typical of terminology in the language of the corpus.

In the focus corpus, as opposed to the reference corpus, terms are multi-word statements that also adhere to the language's normal terminology pattern. Lemmas are used to present the word extraction results.

Term extraction, also known as terminology extraction, is a way of automatically analyzing text to find phrases that describe a text's theme or content or that are typical and/or unique to that text type. The term is usually a noun phrase. An English term, for instance, can be made up of nouns, adjectives, and prepositions.

N-grams – key multi-word expressions (any sequence of tokens). Only those items are included that occur more frequently in the chosen corpus than in the reference corpus. The findings show what distinguishes the chosen corpus from the reference corpus.

The reference corpus is selected for keywords, which are used to compare the focus corpus with. The largest corpus in the language is selected and recommended by default to represent the general language. Terminology extraction extracts words that are typical of the topic of the document or corpus, i.e., they appear in the corpus more frequently than they would in general language. A large non-specialized corpus in the language is used to represent general language to acquire a clearer picture of the typical uses of selected lexical items. (Computing, 2022)

Terminology extraction

Without a doubt, the terminology is crucial to many various industries, including localization, standardization, technical documentation, and translation. Numerous subject areas, including various legal and industrial sectors, as well as business, use a lot of language that is specific to those fields. The nomenclature used by many document authors may also be their own. It takes a lot of time to conduct the necessary research to produce any particular translation.

Although the extraction technologies make extraction easier, a human terminologist or translator must still validate the list of candidate terms that results. Therefore, rather than being totally automatic, the word extraction process is computer-aided.

Identifying term candidates in a text might be referred to as term extraction. It can either be a single language or several languages (usually bilingual). While multilingual term extraction examines existing source texts and their translations

in an effort to uncover possible terms and their counterparts, monolingual term extraction aims to study a text or corpus in order to identify candidate terms.

The process of extracting terms typically consists of four steps: compiling a corpus, extracting term candidates, validating the term candidates, and creating terminological records automatically or semi-automatically.

The setup of the employed software, the word lists that will be imported, and the creation of the extraction rules are all steps that must be completed by humans in the preparation of term extraction projects.

Linguistic term extraction

Language-based term extraction techniques often look for word combinations that fit specific morphological or syntactical patterns, such as "adjective + noun" or "noun + noun." The corpus's material is annotated for these purposes using parsers, part-of-speech taggers, and morphological analyzers. Different methods of pattern matching are used to filter term candidates. Because term creation processes vary from language to language, it is clear that the linguistic method is very language-dependent. As a result, linguistic word extraction technologies are typically made to function with just one language (or a small group of related languages). They cannot be simply modified to work with additional languages. As a result, they are not a good fit for integration with translation memory systems, which are often language-independent.

Statistical term extraction

The main goal of statistical term extraction techniques is to find recurring lexical item sequences. The user can frequently choose the frequency threshold, which denotes the minimum number of times a word or group of words must appear in order to be taken into account as candidate terms. The statistical approach's linguistic independence is one of its main advantages.

Terminology extraction from the focus corpus of business reports

Extracting terminology, as stated above, can identify single words and multi-word units which are typical of a corpus, and it defines its content or topic. Firstly, we will look at the keywords in the focus corpus of business reports.

The chart in Figure 1 demonstrates a clustered column chart to compare values of frequency in the focus corpus.

Upon examination of the keyword frequency, the most prominent keywords can give us an idea about the topic of the typical language of the corpus. An even

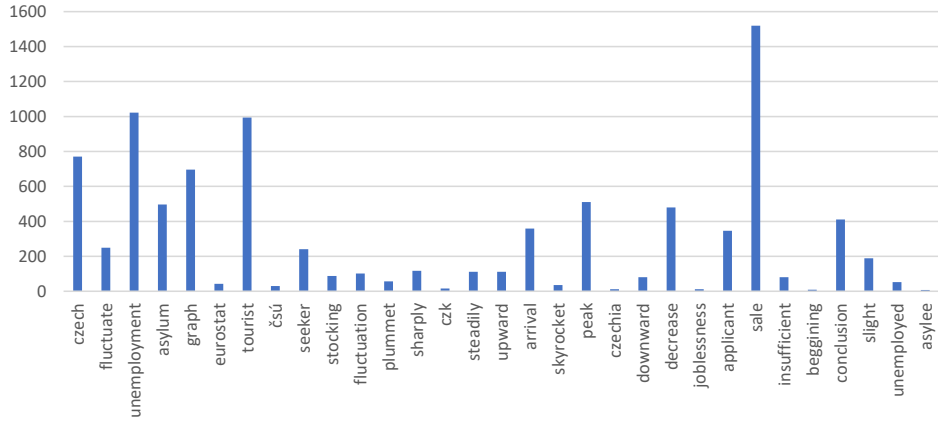

Fig. 1: *list of keywords from the focus corpus of business reports*

better tool is to use the multi-word terms to demonstrate the most frequently occurring word combinations used in the business reports of interest.

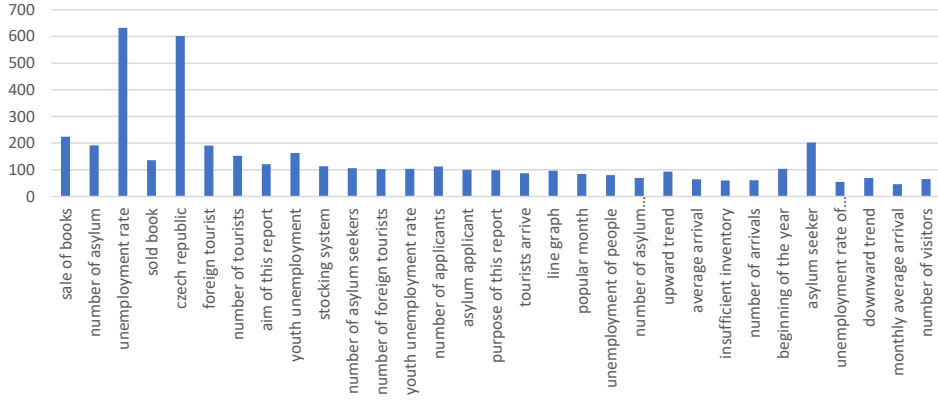

Fig. 2: *List of multi-word terms in the focus corpus*

The chart in Figure 2 gives us an overview of the most frequently used multi-word terms. Upon examination of the frequently used terms, one can deduce the examined texts deal mostly with sales of books, asylum seekers, tourists, and unemployment rates in the Czech Republic.

Upon examination of the list of n-grams, the investigator can get an even more detailed idea of the typical language and topic used in the focus corpus as follows:

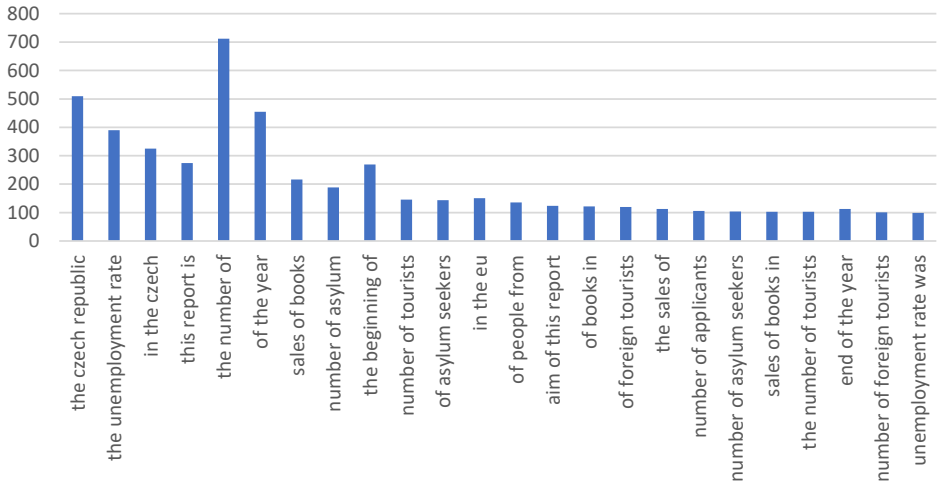

Fig. 3: *List of n-grams in the focus corpus*

As already mentioned, the n-grams present those items of the corpus that occur more frequently in the chosen corpus than in the reference corpus. The findings show what distinguishes the selected focus corpus from the reference corpus. Taking into consideration that the focus corpus was compiled out of business reports written by students from the Czech Republic, one can assume that "the Czech Republic" will be one of the most frequently occurring n-grams in the corpus, as is visible in the above-mentioned chart.

The next step in the analysis of the students' business reports corpus is to compare it to the reference corpus to ascertain the contexts, similarities, and differences in writing and style of the small group of investigated students and the large natural English corpus compiled by linguists. We will be using the relative frequency, aka frequency per million, since that is the best criterion to compare frequencies between corpora of different sizes. Since the focus corpus of business reports is relatively small in comparison to any natural English language corpus used as a reference corpus, the relative frequency can give us a better picture of the typically used words and word combinations and ascertain trends.

Noticeably, number eight in the Table 8 shows the "čsú" word, in fact, an abbreviation meaning "the Czech Statistical Office" has a very low relative frequency in the reference corpus as it is an abbreviation originating in the Czech language. It has been omitted from further investigation. Nevertheless, it was chosen to illustrate the distinctive use in the business reports included in the focus corpus, as many

Tab. 8: *Relative frequencies of the first ten words based on their keyness*

Item	Relative frequency (focus)	Relative frequency (reference)	Score
1. Czech	4450.79004	9.59981	419.99
2. fluctuate	1439.28137	3.20773	342.29
3. unemployment	5907.41211	16.43521	338.88
4. asylum	2867.0022	9.33691	277.45
5. graph	4023.05151	19.20575	199.15
6. Eurostat	242.77036	0.30513	186.78
7. tourist	5745.56494	31.53668	176.62
8. čsú	173.40739	0.00051	174.32
9. seeker	1387.25916	7.69974	159.58
10. stocking	508.66171	3.41601	115.41

reports were written based on statistical data provided by the Czech Statistical Office, thus demonstrating the heavy L1 interference in the investigated text.

Table 8 shows the first ten words and their relative frequencies across the corpora ordered based on their keyness score. This score is used based on simple maths to identify keywords of one corpus vs. another. A higher value (100 and more) focuses on high-frequency words and a lower value (1 and less) focuses on low-frequency words.

According to the statistic used for keywords in the Sketch Engine (Computing, 2022) it is a variation on "word W is so-and-so times more frequent in corpus X than corpus Y". The keyness score of a word is calculated according to the following formula:

$$\frac{f_{pm_{rm\text{focus}}} + N}{f_{pm_{rm\text{ref}}} + N},$$

where
$f_{pm_{rm\text{focus}}}$ is the normalized (per million) frequency of the word in the *focus corpus*,
$f_{pm_{rm\text{ref}}}$ is the normalized (per million) frequency of the word in the *reference corpus*,
N is the so-called smoothing parameter ($N = 1$ is the default value).

Fig. 4: *Formula used to calculate the keyness score (Kilgarriff, 2009)*

Next, we will look at the first ten most frequently occurring multi-word terms based on their keyness and check their relative frequencies.

Tab. 9: *The first ten multi-word terms and their relative frequencies and keyness score*

Item	Relative frequency (focus)	Relative frequency (reference)	Score
1. sale of books	1294.77527	0.0365	1250.15
2. number of asylum	1109.80737	0.05331	1054.59
3. unemployment rate	3653.11572	3.07704	896.267
4. sold book	786.11353	0.01185	777.896
5. Czech Republic	3479.7085	3.50646	772.381
6. foreign tourist	1104.0271	0.48197	745.648
7. number of tourists	878.59747	0.27636	689.146
8. aim of this report	699.40985	0.02261	684.925
9. youth unemployment	942.18018	0.42302	662.8
10. stocking system	653.16785	0.00218	652.745

When we examine the first ten n-grams, we can sometimes observe overlaps with the multi-word phrases, with articles and prepositions acting as typical dictionary expressions rather than standalone lexical bundles. It is desirable to mix multi-word terms and phrases with n-grams as decided by the statistical approach since we are examining the typical language and word combinations of the students/users of the language by focusing on the sample of their language. The following steps in the investigation are decided by the overlapping phrases.

Tab. 10: *The first ten n-grams and their relative frequencies and keyness score*

Item	Relative frequency (focus)	Relative frequency (reference)	Score
1. the Czech Republic	2942.14551	0.05185	2798.07
2. the unemployment rate	2254.29614	0.02901	2191.72
3. in the Czech	1878.5802	0.01825	1845.89
4. this report is	1583.7876	0.04111	1522.21
5. the number of	4115.53564	1.86881	1434.93
6. of the year	2630.01221	1.01671	1304.6
7. sales of books	1248.5332	0.00028	1249.19
8. number of asylum	1086.6864	0.00123	1086.35
9. the beginning of	1554.88635	0.77985	874.169
10. number of tourists	843.91602	0.00512	840.608

After narrowing down the most prominent multi-word terms and combining them with the distinctive n-grams, following is a graphic representation of the relative frequencies and scores of the multi-word terms and n-grams of both corpora to demonstrate the most pronounced word combinations of interest:

The narrowed-down list gives us the most frequently used lexical bundles across the corpora, with distinctive scores, even though showing low relative frequencies

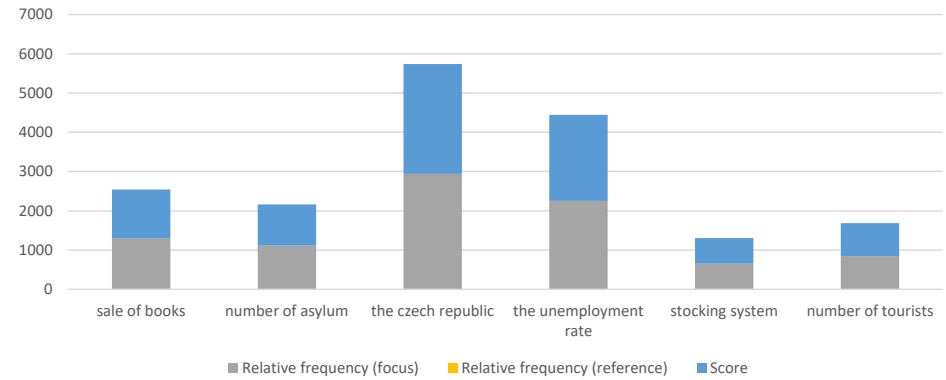

Fig. 5: *Combinations of multi-word terms and n-grams across the corpora in graphic representation*

in the reference corpora. The low relative frequencies can be observed in the numerical values presented in the corpus (tables 9 and 10), which is a result of the considerable difference in frequencies since the focus corpus is a much smaller body of text than our reference corpus. However, the keyness score indicates good comparability with the reference corpus leading to a further investigation into the expanded context via example concordance sentences as follows:

Focus corpus:

In the first three months of the year 2020 the *sales of books* were stable.

Reference corpus:

All proceeds from the *sale of books* are invested back into the running of the Book Festival, a not-for-profit charity organization.

Focus corpus:

The highest *number of asylum seekers* occurred in July 2015, when the number of applications exceeded the 160 000 mark.

Reference corpus:

By world standards, we have a tiny *number of asylum seekers* and accept only a small number of refugees.

Focus corpus:

We believe this alone could raise the number of tourists coming to *the Czech Republic* in the winter months.

Reference corpus:

CzechTourism, an agency run by the state, informed about the most visited places in *the Czech Republic.*

Focus corpus:

The Czech Republic's youth *unemployment rate* is displayed by line graph.

Reference corpus:

A 24% *unemployment rate* is just one of the many job challenges faced by military husbands and wives.

Focus corpus:

Our company should focus on applying *the stocking system* we applied in those months to assure a stable increase in book sales in upcoming months.

Reference corpus:

A rotational *stocking system* controls the timing and intensity of grazing by rotating animals among paddocks, and gives the pastures time.

Focus corpus:

This report is about the *number of tourists* arriving in the Czech Republic in the course of a year.

Reference corpus:

The *number of tourists* visiting Egypt rose in the first four months of 2012, the cabinet said on Tuesday.

We can determine the language context of the most often occurring word combinations from our sample of language by examining a section of examples from both the focus corpus and reference corpus. Users can examine the reference corpus' extended contexts in greater detail to learn more ways to utilize the language and improve their language proficiency.

According to the data, the most commonly used lexical bundles from the focus corpus of business reports were applied in a way that was appropriate and equivalent to that of the reference corpus, creating a standard for natural language use for the students who made up the sample under study.

Conclusion

The usage of specialized language, in our case, the typical language used in business English, can be clarified by limiting the keywords and n-grams, finding the typical lexical bundles in the focus corpus, and using statistics to compare the terms across corpora. A reference corpus can be used to support or refute the appropriateness of language use and to learn more about future usage in various settings.

A corpus analysis can be used to advance research, broaden the focus to less frequently used phrases and keywords, investigate broader settings, examine a range of sources, and identify errors, all of which can improve general and specialized language communication.

This brief study has demonstrated that examining a very small corpus of written reports by L2 English language speakers can yield a wide range of information and uses, as well as suggestions for language improvement and further research. The reported results, which are supported by trustworthy statistics, are simply the start of a more thorough investigation into an expanding database of linguistic resources utilized to increase the instructive value of corpus-based investigations.

Being a relatively new scientific field, corpus linguistics is expanding quickly, and the materials created for research are consistently being updated and enriched. Working with corpora offers a wide range of nearly limitless research opportunities as well as a very quick and efficient response when employing computer tools for corpus processing. Therefore, gathering large amounts of data to process and extract pertinent and useful information for definitive outcomes is a relatively simple operation.

The need for increasing multicultural collaboration and recent advancements in communication place expectations on today's populace's ability to communicate effectively. Since languages make up the majority of a society's culture, it is crucial to become fluent in them and make an effort to prevent ambiguities and misunderstandings. This paper just introduces one potential strategy for completing the goal, exhibiting approaches and viewpoints on a small subset of linguistic traits while making use of the data at hand.

References

Computing, L. (2022, October 22). *Sketch Engine*. Retrieved from Sketch Engine EU: www.sketchengine.eu

Kilgariff, A. (2001). Comparing corpora. *International journal of corpus linguistics*, pp. 97-133. Retrieved from Sketch Engine.

Kilgariff, A. (2005). Language is never ever random. *Corpus Linguistics and Linguistic Theory 1 (2)*, 263–276.

Kilgarriff, A. (2009). Simple maths for keywords. *Proceedings of Corpus Linguistics Conference CL2009.* Mahberg: University of Liverpool.

Ltd, L. C. (2015, July 8). *sketchengine.eu.* Retrieved from Statistics used in Sketch Engine (chapter 5): https://www.sketchengine.eu/documentation/simple-maths/

Author

Michelle Duruttya, M.A., Ph.D., Prague University of Economics and Business, e-mail: michaela.duruttya@vse.cz and the University of West Bohemia, e-mail: durm10@kaj.zcu.cz
She is a researcher, linguist, English language teacher, translator based in Prague, Czech Republic. She cooperates with several Czech as well as international educational institutions as a teacher trainer and teacher of the English language. She has taught English in Hungary, Slovakia, the U.K., and China for more than 15 years while being an Oral Examiner and Team Leader of Oral Examiners of all the main suite Cambridge Examinations for the British Council. She works for the Prague University of Economics and Business and the University of West Bohemia in Pilsen. She cooperates with the Czech Academy of Sciences and several language schools.