THE BUILDING BLOCKS OF BUSINESS CONTRACTS' RECITALS: A CORPUS-INDUCED APPROACH TO ASSESSING HOMOGENEITY

by

Reem Kamal Rabea

A Thesis Presented to the Faculty of the
American University of Sharjah
College of Arts and Sciences
In Partial Fulfillment of the
Requirements for the Degree of

Master of Arts in
Teaching English to Speakers of Other Languages (TESOL)

Sharjah, United Arab Emirates
December 2022

Declaration of Authorship

I declare that this thesis is my own work and, to the best of my knowledge and belief, it does not contain material published or written by a third party, except where permission has been obtained and/or appropriately cited through full and accurate referencing.

Signature: Reem Kamal Rabea
Date: 7th December 2022

Approvals

We, the undersigned, approve the Masters Thesis written by: Reem Kamal Rabea

Thesis Title: The Building Blocks of Business Contracts' Recitals: A Corpus-Induced Approach to Assessing Homogeneity

Date of Defense: 07/12/2022

Name, Title and Affiliation

Dr. Philip McCarthy
Assistant Professor
Department of English
Thesis Advisor

Dr. Tammy Gregersen
Professor
Department of English
Thesis Examiner (Internal)

Dr. Tor Brodtkorb
Assistant Professor
Department of Management
American University of Sharjah
Thesis Examiner (External)

Accepted by:

Dr. Mahmoud Anabtawi
Dean
College of Arts and Sciences

Dr. Mohamed El-Tarhuni
Vice Provost for Research and Graduate Studies
Office of Research and Graduate Studies

Acknowledgments

I would like to express my deepest gratitude to Dr. Philip McCarthy for his precious support and invaluable patience. I would also like to extend my sincere thanks to the entire defense committee. This project would not have been possible without the financial support of FRG22-Comprehensive: The Discipline Specificity Analysis Tool (DSAT). This educational journey was undertaken because of the assistance and the many opportunities I was given by the American University of Sharjah. Special thanks to Dr. Philip McCarthy and prof. Tammy Gregersen for the wonderful experience and the amazing learning opportunities facilitated throughout the program. I am also thankful to Dr. Tor Brodtkorb for sharing his valuable knowledge and legal expertise in his examination of the current study.

I acknowledge that the findings of the present study are dedicated to adding a legal aspect to The Discipline Specificity Analysis Tool (DSAT). Thus, I declare the entitlement of the findings produced in the present study to the DSAT project.

Abstract

By following a bottom-up approach based on principles of induction, the current study aims to provide the most typical lexical units of the provisional element of recitals in business contracts. In doing so, a novel methodological approach to corpus homogeneity is proposed to satisfy quality standards of the corpora serving the purpose of this study. Cosine values were utilized for homogeneity operationalization since they operate on the principle of similarity. Further, in producing the frequency wordlists, the study focuses on two methodological parameters, namely N-gram sequence length and frequency, which are proposed by Biber and Conard (2005). By means of a textual analysis tool named *Gramulator* (McCarthy et al., 2012), the N-grams common to two specialized corpora of several contract types were extracted. A preliminary frequency baseline (< 50%) was employed in the selection process. More notably, the statistical examination of the data suggests the presence of a *parent-child relationship* between both corpora. The findings of the current study are of interest to English for legal purposes (ELP) practitioners who are precisely teaching new law students as well as researchers that wish to investigate the discourse functions forming the conventions of business contracts drafting.

Search Terms: **Wordlists, Gramulator, N-grams, corpus homogeneity, Cosine values, specialized corpora, parent-child corpora**

Table of Content

List of Tables

Chapter 1. Introduction

The field of legal phraseology has received insufficient attention from scholars in general (Basaneze, 2017). Specifically, little research investigated the phraseological units shaping the genre of contracts. For instance, while Alasmary (2019) examined the lexical bundles (i.e., multi-lexical units) employed in various subjects in *contract law*, Cao and Zhao (2019) measured the frequency of several multi-word sequences in English business contracts. Meanwhile, by adopting a comparative research design, Arroyo and Pérez (2019) analyzed word clusters in corpora of real estate sales agreements in both English and Spanish, producing findings of high value to English for specific purposes (ESP) practitioners and contract drafters. These studies appear to collectively lack specificity with respect to contract analysis. That is, rather than analyzing specific clauses, elements, or provisions of legal contracts, mostly a holistic analysis approach was adopted. As such, offering the target audience, especially in an educational context, an in-depth analysis of one focused area is needed.

The current study adopts a direction that has been taken by Governatori and Milosevic (2006). As the authors explain, there is a rising need to study the cumulative evolvement of the language of business contracts (LBC) because such contract types involve complicated contracts' events. Therefore, two corpora of business contract families (*compensation* and *business operations*) are investigated. Further, the object of investigation opted for is *recitals,* which is a provisional element found in the introduction section of a contract. As far as I am aware, except for Arroyo and Pérez (2019), who touched upon the rhetorical structures of recitals of consideration, no other study has analyzed this provision. Thus, the current study acknowledges a gap in the literature of contract drafting.

In addition to the field of discourse analysis, the current study accounts for two important considerations presented in the literature of corpus linguistics. The first is Kilgarriff's advocacy (2001) for objective analysis of how two corpora relate to each other. The second is Miller and Biber's (2015) call for vocabulary studies to legitimize the reliability of a given corpus prior to claiming the validity of its findings. In responding to these calls, an in-depth statistical analysis was performed to uncover the underlying relationship between the target corpora. A novel method of internal corpus homogeneity adapting the notion of corpus saturation (McEnery & Wilson, 2001) is also carried out.

The present study can be classified as a corpus assisted discourse study (CADS), serving two distinct goals in the fields of corpus linguistics and discourse analysis. Accordingly, pedagogical implications pertaining to teaching the lexical conventions of recitals are suggested. The study also proposes methodological implications to corpus compilers engaging in corpus based phraseological studies.

Chapter 2. Literature Review

Since a CADS framework is followed in the current study, it is useful to understand how corpus linguistics contributed to the methodological advancement of discourse analysis. The historical roots of this consolidation will be briefly discussed first, followed by discussions on the corpus tools and types often employed in CADS as well as the methodological considerations that may affect the validity of such studies.

2.1 Corpus and Discourse Methodological Affinity

In the early years of their evolvement, there was a dissociation between corpus linguistics (CL) and discourse analysis (DA). The two fields were traditionally considered separate because of the dissimilar methodological roots each springs from. Mautner (2012) points out that, as far back as 1980 and 1990, corpus studies chiefly relied on quantitative approaches to investigate the lexicographic as well as linguistic features of large-sized corpora. Meanwhile, discourse related studies largely analyzed context-dependent data using qualitative methods. However, with the development of both paradigms, a new research line integrating CL and DA came into existence. Known as corpus assisted discourse studies (CADS), this type of research has been embraced for its mixed-methodological soundness (e.g., Marchi & Taylor, 2018), allowing triangulation of methods in investigating the phenomenon of interest (McEnery & Hardie, 2012). That is, the data is commonly viewed from quantitative and qualitative dimensions. In addition, sampling bias is effectively minimized as discourse analysis is no longer confined to some selected instances from texts (Mahlberg, 2014). In line with the authors' premises, it appears that the methodological assistance of CL indeed helps discourse analysts enhance the qualitative essence of discourse analysis framework.

In discussing the merits of the mixed-method approach further, CL software packages offer several benefits to discourse analysis. Mautner (2015) highlights that CL software provides statistical computations, coupled with textual data dedicated to qualitative analysis. Further, the author adds that the data outputs produced by the software enable discourse functions identification, detection of salient patterns, and investigations of word frequencies as well as the lexical strings these words are accompanied with. More importantly, a prominent issue related to examining respectively large volumes of data is also elevated (see Sinclair & Carter, 2004).

Accordingly, CL software permits the discourse analyst to widen the research scope by examining large data sets that manual approaches would not allow.

2.2 Corpus Tools

As previously delineated, it is commonplace for CADS to exploit the tools offered by CL software. With reference to Partington and Marchi (2015), the most frequently deployed corpus tools are frequency lists, key-cluster lists, and concordance searches. However, given the current study's focus, the discussion below will solely concentrate on frequency lists.

2.2.1 Frequency wordlists

Frequency words inform the common phraseological clusters that are distinctly specific to a discourse. Greaves and Warren (2012) highlight that these words come in several terms, namely, N-grams, clusters, lexical bundles, fabricated patterns, or chunks. The authors explain that these clusters are concerned with generating patterns of meaning that rest on word associations frequently collocating beyond mere arbitrary co-occurrences. One of the vital considerations multi-word sequence studies is concerned with is the length of the N-gram. In respect to this point, Mautner (2012) emphasizes the importance of determining the specificity degree of a word bundle through its length. The author maintains that, "the shorter the N-gram, the less distinctive of a particular corpus it is likely to be" (p. 254). For instance, a study using a two-word cluster would produce more general results in comparison with another employing a four-lexical unit frame. However, the author holds that generic and peculiar bundles are both beneficial when corresponding to rightly proposed research questions. As such, the goal of the study is what determines the bundle focus.

Prior studies have paid particular attention to determining an appropriate N-gram length. To provide illustrative examples, Cortes (2004) ran a preliminary frequency analysis to decide on the bundle length in students' disciplinary writing, considering four-word clusters instead of five. In doing so, the author selected the bundle with the higher recurrence rate (< 10%). Furthermore, in light of Biber and Conard (2005) findings, Breeze (2013) employed bundles ranging from four to eight lexical units as four legal genres were under examination. It should be noted that Breeze restricted bundle selection from the texts based on a frequency threshold determined beforehand. On this evidence, it appears that methodological considerations of bundle length correspond with frequency accounts and other factors pertaining to the object of investigation.

In addition to length, other N-gram characteristics require attention. Biber and Conard (2005) managed to single out the crucial methods of inspecting phraseological units, despite researchers' varying views. These methods entail considering the following characteristics of lexical bundles: "fixedness; idiomaticity; frequency; completeness in syntax, semantics, or pragmatics; and intuitive recognition by native speakers of a language community" (p. 57). The authors also note that the study's purpose motivated the method of choice.

To operationalize some of the previously mentioned features, certain statistical means are often designated. For instance, Brezina (2018) highlights that the use of frequency threshold often avails the researcher to filter the multi-word sequences with highest relatedness in the target discourse (i.e., Biber and Conard (2015) call them *building blocks*). However, there are debates with regards to determining cut-off frequency baselines (Breeze, 2013). The author also recommends the use of parametric (i.e., t-test and ANOVA) and non-parametric tests (i.e., Kruskal Wallis) in case of discourse or multiple genre comparisons. In brief, lexical bundles echo discourse conventions that need to be inspected cautiously in association with well-defined criteria.

2.3 Corpus Type

In addition to the previous considerations, choosing a suitable corpus type is important. In this regard, there are mainly two corpora types with differential focus: *general* and *specialized.* These corpora types seem to be mostly employed in corpus assisted discourse studies (CADS). Both corpora types are discussed as follows.

2.3.1 General corpus

As defined by Koester (2012), g*eneral corpora*, also known as *mega corpora*, incorporate millions to billions of tokens, such as The Bank of English, The Brown Corpus of Standard American English, and The British National Corpus. Such commercialized corpora can achieve representativeness (i.e., the variability degree of sample texts) through heavy reliance on a broadly diversified sample (McEnery & Wilson, 2001), involving a comprehensive range of text types. Accordingly, general corpora compilers need to reflect the variability in the text population in the sample.

General corpora may offer more benefits to studies with a general focus. Connor and Upton (2004) acknowledge the importance of general corpora in providing valuable evidence about linguistic patterning, chiefly instrumental to construction of dictionaries (e.g., Neufeldt, 2019) and main grammar rules (e.g., Biber

et al., 1999). However, the authors argue that such corpora are less instrumental in scrutinizing specific professional and academic contexts, giving rise to the use of specialized corpora. Likewise, Mautner (2015) conceives general corpora as an optimal tool for CADS concerned with broader aspects of the discourse of interest (e.g., Baker et al., 2008). In line with these arguments, it appears that general corpora may be an apt fit for studies with a general linguistic focus.

2.3.2 Specialized corpus

As opposed to general corpora, specialized corpora are suitable for studies with a specific focus. De Geaugrande (2001) explains that specialized corpora are "delimited by a specific register, discourse domain, or subject matter" (p. 11), such as the Michigan Corpus of Upper-Level Student Papers (MICUSP) and the European Corpus of Academic Talk (EuroCoAT). Koester (2012) reports that unlike the decontextualized nature of general corpora, specialized corpora permit a deep analysis of corpus data in relation to the context under analysis. The author also highlights that the specificity of such types of corpora are mostly convenient for informing pedagogical implications for English for specific and academic purposes. Nonetheless, there have been several debates over the issue of representativeness in the context of specialized corpora. For instance, Williams (2002) expresses the inapplicability of accomplishing representativeness for narrowly focused corpora. The justification provided by the author is that targeting a specific theme in a given discipline may constrain diversity.

In contrast to Williams (2002), Vaughan and Clancy (2013) claim that representativeness of small or specialized corpora can be measured. The authors emphasize that representativeness should be treated with equal caution as general corpora, given the importance of such a parameter, regardless of the size. In a similar vein, while McEnery and Wilson (2011) hold a similar point of view, they argue that ways to gauge representativeness in general and specialized corpora are fundamentally different. As indicated above, sampling lies in the heart of general corpora representativeness. However, with respect to specialized corpora, an approach founded on the grounds of measuring lexical finiteness, called *corpus closure* or *corpus saturation*, is employed, which will be discussed in detail in the following section. In all respects, the sampling frame of a corpus may be liable to vary, depending on the goal and type of the compiled corpus.

2.4 Quality Assurance

In corpus-based vocabulary studies, it appears that the process of corpus quality assurance is a methodological precondition for producing valid results. More specifically, prior to claiming that a wordlist presents the lexical units typical of a particular discourse, it is important to ensure the quality of a corpus from the outset. For instance, Schmitt (2010) primarily bases the quality of a wordlist on the good state of its corpus. This methodological issue has been further raised by Miller and Biber (2015), advocating for employing a quality measure to assure that a corpus would yield the same wordlist every time another corpus representing the same target domain is investigated. The authors framed this idea into the notion of *internal representativeness*, meaning that the linguistic variability of the discourse under investigation is entirely reflected in the corpus. This notion can be operationalized by having a corpus partitioned into sub-corpora, and if the same quantitative results are found across the segments, then it is safe to contend that the whole corpus contains the diverse linguistic features.

Similar to Miller and Biber's (2015) approach, another quantitative procedure aiming to assure the lexical quality of a corpus is proposed by McEnery and Wilson (2001*). Corpus saturation/closure* assumes the saturation of a corpus if the incorporation of new segments of texts into already existing ones, leads to inconsequential changes in the new lexical units in comparison with the original segment (Xiao, 2010). The usefulness of this approach was documented by Łukasik (2014), who emphasized the usefulness of the closure parameter in studies dealing with corpora dedicated to terminologies extraction, or analysis of collocations as corpus saturation assists in the omission of lexically inappropriate items. That is, as a quality assurance instrument, this procedure permits corpus verification. Overall, the validity of vocabulary related results seems to heavily rely on the lexical saturation of a corpus.

2.4.1 Corpus homogeneity

Corpus homogeneity has been approached quite differently by scholars. For instance, Gries (2006) conceives internal corpus homogeneity to be associated with internal variability, prompting the target linguistic phenomenon under analysis to vary across the corpora. Meanwhile, Kilgarriff (2001) links homogeneity with internal similarity existing between the distributions of two sub-divided segments of compiled documents. In this regard, the author proposes using word-frequency measures for

internal homogeneity evaluation. The first step is portioning the corpus into slices, followed by a random allotment of each slice into sub-corpora created earlier. Subsequently, the sub-corpora similarity should be evaluated, then the entire procedure can be iterated with successive sets of texts. The last step requires the calculation of mean and standard deviation to assess the similarity of the iterations as a whole.

Having closely looked at internal homogeneity, it is vital to understand this property from a holistic perspective. In the literature, homogeneity has been generally associated with corpora comparison, either for similarity (e.g., Denoual, 2006; Kilgarriff, 2001; Rose et al., 1997) or variability (e.g., Rayson et al., 1997). Nevertheless, various statistical applications have been used in comparing corpora, such as log-likelihood statistic, cross entropy, chi-squared test, and chi-squared statistic. Most studies that investigated corpora similarity seem to rely on different variations of Kilgarriff's (2001) methodology. For instance, although Rose et al. (1997) employed the same algorithms, they computed correlation rank using a wordlist for measuring similarity between corpora. Overall, upon review of corpus homogeneity literature, this property has received little attention, causing variation in the way it is conceptualized.

Chapter 3. Current Study

As indicated earlier, one of the main purposes of the current paper is to acknowledge the gap of phraseological studies investigating the genre of contract drafting. The existence of such a gap was further confirmed by Alasmary (2019) claiming that, "despite decades of research into the distribution of multiword patterns in various disciplines, *contract law* seems to escape the attention of researchers" (p. 252). Given the nature of the genre under investigation, two specialized corpora were employed, where each was analyzed using a corpus-induced approach (Lee, 2008).

To claim the validity of the target corpora, a novel procedure based on internal lexical corpus homogeneity (LCH) is proposed. This procedure originates from the notions of internal representativeness (Miller & Biber, 2015) and corpus closure (McEnery & Wilson, 2001). However, some changes were made to the algorithms used to measure these parameters. In this regard, I adopt Kilgarriff's (2001) definition of homogeneity in formulating the theoretical framework of the newly proposed approach. Overall, the present paper seeks to address the following research questions:

1. How do the two corpora under investigation compare to each other?
2. Does the lexical corpus homogeneity (LCH) procedure measure what it purports to measure?
3. What are the building blocks of business contracts recitals?

Chapter 4. The Method

This chapter draws upon the phases of corpus collection, refinement, and cleaning. The tool employed for the textual analysis carried out in the current study is also described.

4.1 Phase I: Preliminary Database

The current study employs two specialized corpora that were collected and constructed according to specific criteria provided in McNamara et al. (2014). With reference to Coh-Metrix studies, which are primarily concerned with computational text analysis, the authors define a *corpus* as, "a set of written, representative and balanced, computationally readable texts that form a reasonable point of departure as a thematically related language variety, register, genre, or text-type" (p. 146). In defining the term *corpus*, the authors explicitly identify a set of elements that should be carefully considered when compiling a corpus. The following sections demonstrate how the current study addresses the previously mentioned elements of corpus design.

I initially compiled a database of legal business contracts from a legal website called *FindLaw.com.* The website belongs to Thomson West, who owns one of the most well-known legal materials publishing firms in the U.S. The website provides a freely accessible archive of sample business agreements obtained from publicly traded companies in a variety of industries, namely, *drugs*, *biotech*, *technology*, and *real estate*. The website has a wide collection of contracts grouped into several categories according to contract types. Each category typically comprises relevant contracts types that are, from a legal standpoint, distinguished according to a single theme. To attend to the thematic representativeness aspect to corpus compilation, c*ompensation* (COMP) and *business operations* (BO) categories were chosen for analysis for having the greatest diversity of contract types. That is, the compensation category consists of *consulting*, *employment*, and *severance* as well as *separation* agreements. Meanwhile, the business operations category includes *sales*, *marketing,* and *intellectual property* agreements. With reference to each category (i.e., *compensation* and *business operations*), the relevant contracts were systematically drawn from the population text for each contract type. Specifically, the selection criterion for contracts was solely based on the order in which contracts were displayed on the website, regardless of size. In fact, the contracts were chosen from the top down, with each sample member having an equal probability of being chosen, as each contract list was exhausted to the fullest.

Having compiled the preliminary dataset, a number of contracts were eliminated. To illustrate, contracts containing many repetitive language items, particularly the ones written by the same business entity were eliminated. Additionally, extension and amendments contracts were excluded since copies of the original contracts were unobtainable. If the linguistic features of modified contracts were to be accurately analyzed, it was necessary to compare the existing and amended contract terms to have an adequate understanding of the context. Promissory notes were ineligible for selection as they do not necessarily require having an important element of legal contracts called *consideration*, meaning inducement given by the promissor to the promisee (McCamus, 2020). In sum, it can be established that there were no instances of biased or favored contracts upon sample selection.

4.2 Phase II: The Corpus

After collecting the business contracts, it was important to refine both corpora in terms of topic focus. In essence, business contracts formation generally entails various provisions that serve different functions, such as *preamble*, *recitals, covenants*, *rights*, *remedies*, and many others. Performing an in-depth textual analysis would indeed necessitate limiting the scope of the study to investigating one of the provisions in question. Thus, among the numerous provisions that constitute a business contract, I particularly opted for *recitals* to be the object of investigation, given the insufficient attention they receive. *Recitals* are a group of paragraphs that state the background information outlining the nature of the agreement (Adams, 2018). Draftspeople normally use recitals to describe the context of the contract.

Having decided upon the provision of interest, both corpora were refined to strictly rest on recitals (see Table 1). More importantly, it was crucial to maintain a sufficiently reasonable number of examples with respect to each contract type to ensure the balance of each corpus. The outline of the structure of compensation and business operations corpora is provided below.

Table 1: The mapping of compensation and business operations corpora

Contract Provision	Category	Contract Type
Recitals (111)	Compensation (60)	Consulting (15)
		Employment (30)
		Severance & Separation (15)
	Business Operations (51)	Sales & Marketing (21)
		Intellectual Property (30)

Overall, I ascertained the representativeness and balance of the current corpora on the grounds of having an appropriate collection of text varieties, with the pervasiveness of appropriate frequency of examples for each contract type. This being said, no corpus can be regarded as fully representative (McEnery & Hardie, 2012), nor do we have a concrete scientific method to gauge the balance of a given corpus (Xiao, 2010). However, by relying on best estimates and informed judgements (Sinclair, 2005), I was able to determine the most suitable range of text membership. More importantly, McNamara et al. (2014) point out that Coh-Metrix based studies do not need to employ an exhaustive, unequivocal corpus, such as The British National Corpus. Instead, the authors argue that having "a reasonable point of departure" by constructing a "practical and suggestive" corpus that is directed at answering the research questions and able to guide subsequent studies is adequately sufficient (p. 150). Therefore, the aim of the current corpora is not to establish a reference repository, but instead, to conveniently serve the purpose of the current study and other disciplinary studies interested in studying recitals.

4.3 Corpus Cleaning

The process of corpus cleaning concerns making changes to texts, either by removing or adding elements into the text to qualify raw data for inspection. The reason for cleaning the data is to mitigate having unreliable output based on flawed input, which in computer science is called *garbage in, garbage out* (Powell et al., 2022). For instance, when uploading data into a computational tool, researchers commonly remove annotations and picture captions to avoid misleading evaluations (McNamara et al., 2014). However, in corpus linguistics, corpus texts that could adversely lead to inaccurate data analysis are called *boilerplate* texts (McEnery & Brookes, 2012). Such texts are generally a source of noise, leading to complications in the data analysis process, such as repetitive headers that can influence the results of frequency analysis tools.

Although neither corpora employed in the current study featured any boilerplate texts, some discourse features required *dummy coding* to avoid any possible analytical errors, in addition to conserving the authenticity of the data. In this process, Sinclair (1991) recommends replacing the removed data with codes or tags rather than complete elimination. To encode the spaces that existed between the paragraphs of the recitals, the word *PARG* was consistently added throughout. In sum,

it was imperative to reserve the line breaks, as I noticed the frequent use of these breaks in association with certain words.

4.4 The Tool

The present study employs two tools to carry out the textual analysis of the data. Both tools are explained as follows.

4.4.1 Gramulator

The current study employs an exhaustive textual analysis tool called *Gramulator* (McCarthy et al., 2012). The authors define the *Gramulator* as a computational tool that facilitates the identification of "indicative lexical features of texts and text types" (p. 1). The tool principally analyzes *two sister corpora* in relation to each other, meaning two corpora under the assumption of being closely related, yet respectively distinct in some aspects.

Gramulator relies on analytical entities, termed *N-grams,* which are concerned with discerning patterns in the data under analysis. These lexical frames are articulated by Kopaczyk and Tyrkkö (2018) as, "any sequence of tokens that is of a specific length regardless of its frequency or distribution in a corpus" (p. 294). The notation *N* relatively specifies the sequence length of the frame. For instance, a two-word gram is called a *bigram,* whereas a *trigram* refers to a three-word cluster. By default, the Gramulator provides three sets of output containing lexical characteristics termed *differentials*, *typicals*, and *reciprocals*.

The current study solely focuses on investigating the features of *reciprocal* indices. *Reciprocal indices* constitute the shared N-grams (above 50% frequency) in two given corpora. However, their frequencies are likely to be different in each corpus. For instance, in analyzing the reciprocals of two sub-corpora (A and B), the bigram *enter into* exhibited different recurrence values in A(B) in comparison with B(A) (5.4% and 10.9% respectively). This example indicates that although the reciprocal bigrams occur at different frequencies, they still appear in both corpora.

Each reciprocal bigram is typically assigned a numerical value by the software. This value represents the *term frequency: document frequency* (TF:DF), which is based on the frequency percentage of an N-gram occurring in the corpus documents (McCarthy et al., 2012). For instance, if the bigram *of the* occurs 15 times in the entire corpus, but only emerges in 7% of the files, then a value of 1.05 is assigned. Thus, this value represents the recurrence rate of a bigram in the corpus.

The Gramulator consists of several modules, four of which are specifically used for the textual analysis of the current study. The *Sorter* module was substantially used to partition the corpus data into three sets, where random text assignments were made by the software. The module normally splits "the data into *training*, *validation*, and *testing* sets" (McCarthy et al., 2012, p. 10). However, for convenience, I opted for alphabetical letters, namely A, B, and C to label the three sets. Depending on the size of the data, the weight of each set is determined, with training set normally possessing the highest weight and validation having the lowest.

As for corpus assessment, *the Evaluator* module was selected. A t-test was conducted based on the value input involving the percentage of which a text is made of the bigram of interest. The *Evaluator* also features a measure of lexical diversity termed Measure of Textual Lexical Diversity (MTLD), which is explained below. *The Viewer* was used to access the reciprocals with their TF:DF. This module also contains an important measure of similarity, known as *cosine similarity*, which can be used to measure the lexical similarity of target corpora. Finally, *the Concordancer* was employed to obtain examples of a determined bigram within the entire corpus. The module also presents the number of instances and the emergence percentage of a bigram in the corpus.

4.4.2 Measure of textual lexical diversity (MTLD)

McCarthy and Jarvis (2010) define MTLD as, "an index of a text's lexical diversity (LD), evaluated sequentially." (p. 5). MTLD computes the average length of lexical clusters in their respective sequential order based on calculations of type-token ratio (TTR). For instance, the word cluster of *in consideration of* is sequentially evaluated according to the TTR number of each lexical unit. The authors highlight that once "the default TTR factor size value (here, .720) is reached, the factor count increases by a value of one and the TTR evaluations are reset" (p. 5). In doing so, word clusters that constitute insufficient factors are still measured. To complete these factors, approximate mean values for the lexical units in question are assigned a value of one until the target (.720) is attained. After the factor value is computed, the MTLD final value is generated through a division of the aggregate number of the lexical units by the aggregate factor value.

One of the shortcomings of MTLD is its sensitivity to short texts. McCarthy and Jarvis (2010) report that the insertion of approximate values for insufficient factors caused texts below 100 tokens to probably constitute more estimated mean

values than they should. Consequently, the reliability of the analysis may be affected. Despite this limitation, MTLD was employed in the current study because it is the most widely used lexical diversity tool, and the only tool that is not affected by differences in text length. As opposed to other measures, MTLD does not require the researcher to limit all text lengths to obtain valid lexical diversity findings.

Chapter 5. Results

The results section presents the stages of performing the internal lexical corpus homogeneity test (LCH), as a novel method to measure the degree of the lexical homogeneity of a specialized corpus. Four major stages are discussed below, including the statistical procedures required to ensure the validity of the results. The relationship between COMP and BO is also examined below.

5.1 The Stages of Performing the Lexical Corpus Homogeneity (LCH) Test

5.1.1 Stage zero: Corpus collection

Prior to performing the LCH test, I collected two corpora. One corpus (i.e., compensation) consisted of 60 texts, while the other (i.e., business operations) comprised 51 texts. However, for statistical reasons, at least 60 texts are recommended, although a greater number is preferable.

5.1.2 Stage 1: Data division

By using the *Sorter* module in Gramulator, each corpus was randomly divided into three equal-sized groups (A, B, and C). While 20 texts were randomly allocated to each group belonging to COMP, 17 texts were distributed to each of the groups entitled to BO in the same fashion. Nevertheless, since BO comprised only 51 texts, the entire LCH procedure was replicated to establish confidence upon comparison of the results. In addition to BO, COMP was also qualified for replication, although the number of COMP texts reached the recommended threshold (60 ≥). The reason for repeatability of measurements is to ensure consistent, reliable results (Laak et al., 2013). The overall goal is not to obtain identical results. Instead, similar results pointing in the same direction is the aim. For these reasons, the LCH test was conducted twice for each corpus.

5.1.3 Stage 2: Identification of super indices

The following sections in this stage of the LCH test delineate the procedure to formulating a *super index* composed of the highest frequency bigrams (HFBs) occurring in a corpus. With the utility of a cosine similarity measure, it is illustrated below how to compute the lexical homogeneity of the groups from which the super indices are extracted.

5.1.3.1 Preparing the data

Having divided the data of each corpus, I paired the three groups into six subsets, namely A(B), A(C), B(A), B(C), C(A), and C(B) for COMP and BO separately. In the pairing A(B), A is the first group being analyzed in relation to B (the second group).

The next step was to identify the *reciprocal indices* of each pairing. However, it was essential to determine the length of the N-gram first. A two-word unit (i.e., *bigram*) was selected as all previous Gramulator work verified the use of bigrams. The use of bigrams also enabled the analysis of certain single words that were constantly used on separate line breaks (i.e., PARG_whereas). According to Arroyo and Pérez (2019), the opted N-gram length appears to be useful in readily identifying the keywords that are typical of the target domain. For these reasons, the bigram length was chosen.

Having specified the N-gram length, the reciprocal bigrams of each pair were processed by Gramulator, and the outputs were automatically obtained through *the viewer* module. Overall, corresponding to the six matched pairings, six reciprocal indices were obtained.

5.1.3.2 Cosine similarity computation

This step is dedicated to test the six produced indices for similarity prior to extracting the HFBs. The measure employed to compute the groups similarity at the lexical level is cosine similarity. Han et al. (2012) argue that cosine similarity is often implemented in textual analysis to compute the similarity of a given document. In technical terms, it is concerned with calculating the degree to which two vectors (i.e., sequences of numbers) are similar by assigning a value between zero and one, with zero indicating dissimilarity, and one evidencing complete resemblance (Park et al., 2020). The closer the value obtained is to one, the more similar the two vectors are expected to be.

Another point explained by Han et al. (2012) is that cosine similarity accounts for frequencies of certain words, known as *term frequency vector*. These frequencies are equally investigated in the reference documents so that multiple comparisons could be subsequently made. In case of the absence of a lexical unit in one of the texts, a value of zero is recorded. As such, to perform the cosine calculations, Gramulator was used, specifically by clicking on *corpus similarity* button that can be found in *the viewer* window. To derive the cosine similarity between the groups of each corpus, fifteen pairwise comparisons were executed (see Table 2). The comparison procedure was conducted twice for both respective corpora.

Table 2: Cosine values of BO and COMP

Group Pairs	COMP I	COMP II	BO I	BO II
AB-AC	0.995	0.988	0.988	0.993
AB-BC	0.929	0.976	0.837	0.898

BC-AC	0.931	0.970	0.876	0.900
AC-BA	0.930	0.967	0.874	0.902
AB-BA	0.936	0.980	0.882	0.906
BA-BC	0.991	0.995	0.989	0.993
BC-CA	0.920	0.941	0.834	0.931
BA-CA	0.919	0.938	0.831	0.933
AC-CA	0.947	0.959	0.922	0.866
AB-CA	0.942	0.934	0.909	0.858
CB-CA	0.994	0.971	0.992	0.990
CB-BC	0.924	0.969	0.843	0.931
CB-BA	0.917	0.964	0.834	0.933
CB-AC	0.941	0.953	0.916	0.859
CB-AB	0.939	0.960	0.912	0.857

All cosine values obtained suggest that all groups compared are closely similar, if not, almost identical. These findings indicate that a high degree of lexical similarity, precisely a bigram-to-bigram similarity consistently occurred in the groups of each corpus. The findings of this stage are important for two reasons. The first is that the cosine values establish a powerful indication of internal-corpus homogeneity, knowing that the frequencies of the reciprocal bigrams of each group are considered. The second reason is that the reciprocals of each group qualify to produce the super indices, given the great similarities found. In brief, this stage was concerned with deploying a corpus similarity measure to ascertain whether major variations in the lexical features of BO and COMP groups existed.

5.1.3.3 The formulation process

To create this exclusive index, called *super index*, all six subsets were assembled and combined. Only the bigrams occurring consistently in all subsets were considered, that is, the ones occurring six times across A, B, and C combined. Alternatively, all non-HFBs occurring with frequency (≤ 5) could be removed from the three groups as they are not lexically representative of the corpus. Overall, two super indices were developed with respect to each corpus. The super indices of BO constituted 25 and 27 bigrams respectively, whereas those of COMP consisted of 23 and 25 bigrams.

5.1.4 Stage 3: Super indices validation

This step is concerned with testing whether the super indices are "measuring what they purport to measure" (Oluwatayo, 2012, p. 391). The core function of these indices is to measure similarity through the identified set of the word clusters that are

typical of each corpus in relation to the other, that is, the HFBs as I called them. As a consequence, the situated validity in the LCH context would postulate that for HFBs to measure similarity, they should be occurring at similar frequency ratings within the entire corpus. In this respect, a t-test between groups (A, B, and C) and their respective corpora was performed using the HFBs previously identified. The p-values obtained are shown in Table 3.

Table 3: The t-test between the divided groups and each corpus

GROUP	BO I	BO II	COMP I	COMP II
A	.419	.530	.934	.937
B	.456	.905	.406	.776
C	.894	.500	.330	.713

As depicted in the Table, no p-values computed are significant, suggesting that HFBs of BO and COMP are not different in relation to their corpora. Therefore, the t-test results signify the validity of the super indices.

5.1.5 Stage 4: Uber index creation

With the compelling evidence that the super indices measure what is intended to be measured, an *uber index* can be derived from the HFBs. The uber index, as the name suggests, constitutes the reciprocal bigrams that are ubiquitous in the corpus, representing the most prevalent collocates requisite to writing recitals. On a frequency spectrum, the uber index would surpass the frequency of the super index. To create the uber index, a comparison between the super indices of both corpora were made, where only the common bigrams were considered. Accordingly, for BO, 23 common bigrams were found, while four were excluded. The same number was found for COMP; however, only two bigrams were eliminated (see Table 4).

Table 4: The uber indices of BO and COMP

BO	COMP
consideration of	_parg_ now
in consideration	_parg_ whereas
_parg whereas	agree as
now therefore	and _parg_
the parties	and the
parg now	as follows
and _parg_	company and
as follows	consideration of
desires to	desires to
the mutual	in consideration
set forth	in the

terms and	now therefore
the terms	of the
agree as	set forth
with the	terms and
for the	the company
in the	the executive
of the	the mutual
on the	the parties
to the	the terms
and conditions	therefore in
this agreement	to the
therefore in	whereas the

There were overall 20 bigrams commonly shared by both corpora in the uber index. This finding suggests that these bigrams are not merely indicative of the contract families under analysis. Instead, they represent the building blocks of business recitals composition, given the high systematicity of these bigrams across the five different contract types. However, while two bigrams (*and conditions* and *this agreement*) were peculiar to BO, three bigrams (*company and*, *the company*, and *the executive*) were unique to COMP. These peculiar bigrams can be considered as keywords of each contract family.

Inspection of COMP recitals revealed that the drafter constantly made reference to *the company* as a contractual party. For instance, in one of the consultation contracts, the drafter stated that *the company wishes to assure itself of the services*. This finding is logical since compensation contracts are mainly based on defining what the company owes the employee or the third party and vice versa, either in the event of termination or recruitment of a party. However, the recurrence of the word *executive* cannot be considered as a keyword as most of the separation contracts entailed termination of company executives. With respect to BO, as suggested earlier, referring to the terms and conditions of conducting a business operation is highly important for defining the legal obligations of each party, in case of any breach in the future. Thus, the keywords detected were mainly related to emphasizing the contract conditions. For example, *enter into this agreement on the terms and conditions set forth.*

To measure the recurrence rate of each bigram in the corpora, The *Concordancer* module in Gramulator was used. Upon comparison of all frequency rates, the phrase *in consideration of,* which is composed of the two bigrams *in*

consideration and *consideration of*, recorded the highest value in both COMP and BO (see Table 5). Although this phrase appears to occur more frequently in COMP (71.7%) than in BO (64.7 %), the difference between both corpora is not significant (p = .684). In recitals, the phrase *in consideration of* is usually written in the last sentence, signaling the end of the recitals. Adams (2014) articulates that this last statement is formally called *the recital of consideration*, which is a traditional component of the recital provision. This statement simply means *in exchange for* and it is used as a reference to the element of *consideration* in a contract, meaning the exchanged value between the promisor and promisee. As shown in Table 5, the phrase in *consideration of* is often preceded by *now therefore* to signal a transitioning to the body section of the contract.

Table 5: Concordances of the phrase *in consideration of* in BO and COMP recitals

BO
set forth in section a below _PARG_ **in consideration of** the foregoing , the mutual covenants and agreements
set forth therein _PARG_ therefore **in consideration of** the mutual promises contained herein the parties
this agreement_PARG_, now, therefore, **in consideration of** the mutual covenants and conditions contained
forth in this agreement, therefore, **in consideration of** the mutual covenants hereinafter set forth
under this agreement. now, therefore, in consideration of the above recitals, the terms and conditions
COMP
such services. _parg_now, therefore, **in consideration of** the mutual covenants hereinafter recited,
performance of such services; now, therefore, **in consideration of** the mutual covenants contained in this consulting agreement,
described in paragraph 5 below. therefore, **in consideration of** the mutual covenants and agreements set forth in
conditions set forth below. now therefore, **in consideration of** the promises and of the mutual covenants, conditions

In brief, all statistical procedures carried out confirm the validity of the LCH procedure. The results obtained are in the anticipated direction that the LCH test is valid to measure the homogeneity of a given corpus. The super/uber indices, which this test fundamentally rests on, have important uses. As mentioned earlier, the indices could be employed as a homogeneity measure. However, if the cosine values obtained are closer to zero as well as statistically significant results in stages two and

three are obtained, these findings would indicate a level of lexical heterogeneity in the corpus.

5.2 The Relationship Between COMP and BO

To examine the relationship between BO and COMP, I conducted a between corpora t-test using HFBs (i.e., highest frequency bigrams). As shown in Table 6, the results suggest that COMP is significantly different from BO, with the exception of one test producing a p-value slightly over the significance level ($p = .063$).

Table 6: The between corpora t-test

	BO			**COMP**	
t-test	I	II	**t-test**	I	II
COMP	$p < .001$		**BO**	.025	.063

To assess the representativeness of BO's HFB in COMP data set and BO's HFB in BO itself, the average values were calculated. These averages were computed by using the *values* (i.e., percentage of which a text is made of a bigram) produced by *the evaluator* function in Gramulator for every bigram contained in the super indices. The same computations were simultaneously performed for COMP corpus (i.e., COMP's HFB representativeness in BO and COMP itself). The results reveal a new direction in seeking to understand the relationship between both corpora (see Table 7). The average values of COMP relative to BO (COMP $\propto$ BO) reveal that COMP encompasses a higher amount of BO's indicative language than BO itself. Thus, the results suggest that COMP is a subset of BO. When the entire statistical procedures were reiterated using the uber indices, all results obtained were in agreement.

Table 7: Average values of HFBs in each corpus and its relative

BO corpus			**COMP corpus**		
t-test	I	II	**t-test**	I	II
COMP	7.15	7.66	**BO**	5.67	5.80
BO	5.87	6.03	**COMP**	9.30	9.63

To verify the previous result, I measured the lexical diversity of BO and COMP. Specifically, Measure of textual, lexical diversity (MTLD) by (McCarthy, 2005) was employed to have a better understanding of the power dynamics underlying this relationship. MLTD shows that BO is lexically more varied than COMP, with BO possessing a significantly higher average than COMP (COMP (M) = 46.5; BO (M) = 52.5, $p = .001$). The magnitude of BO's lexical diversity caused COMP to be

statistically different from BO despite the existence of commonalities. The reason for this dispersion in the results can be attributed to the contractual nature of BO. In fact, BO-based contracts comparatively have more lexical variation since more terms on contractual parties' rights reservation and exhaustive payment-related obligations are discussed in great detail. Furthermore, on closer inspection for texts comprising more than 100 tokens, the results obtained are in agreement with the overall MTLD measurements (number of texts: N >100, COMP (M) = 45.6; BO (M) = 52.8, p = .023). Nonetheless, with respect to texts under 100 tokens, the average difference between COMP and BO was not statistically significant (number of texts: N <100, COMP (M) =47.1; BO (M) = 51.9, p = .198), qualifying COMP to be statically more diverse than BO. Under these circumstances, only the results of texts constituting 100 ± tokens are considered, given MTLD sensitivity to text sizes below 100.

5.2.1 Relationship formalization

The results above offer compelling evidence that there is a part-whole relationship between COMP and BO. In the field of lexical semantics, this hierarchical relationship is called *meronymy-holonymy*. For example, *leaf* can be viewed as a part (i.e., holonym) of a *tree* (i.e., meronymy), which constitutes the whole. In connection with this semantic relationship, BO is designated as the *meronymic corpus*, whereas COMP is denoted as the *holonymic corpus.* The previous terms were assigned to each corpus based on the average values of HFBs occurring in each corpus and its counter (i.e., HFB (BO): COMP and HFB (COMP): BO) as well as the MTLD results. That is, as the results suggest, COMP corpus is composed of a higher percentage of BO's language than BO itself.

Although BO and COMP embody a part-whole relationship, I decided to designate different labels to both corpora. Rather than *meronymic and holonymic,* BO was denoted as *the parent corpus,* whereas COMP was designated as *the child corpus.* The reason for choosing these terms is that previous Coh-Metrix studies used the notion of sister corpora to refer to correlative corpora (McCarthy et al., 2012). Thus, it is more meaningful to establish relevance by adopting closely linked terms that are conducive to such studies. For convenience also, parent and child corpora are more facilitative, promoting an easier understanding of the underlying relationship.

Chapter 6. Discussion and Conclusion

The chapter below provides a discussion of the findings of Gramulator and the results of some of the decisions pertaining to setting certain parameters in the analysis. The implications, limitations, and the conclusion of the current study will also be presented below.

6.1 Discussion of the Findings

In connection with the parent-child relationship finding, analyzing the child corpus first may be more convenient. To identify the indicative language of the parent corpus, the reciprocal N-grams of the child corpora need to be determined first. To provide an example, in the domain of biology, to identify the type of a tree, scientists normally study the features of the leaf. Likewise, the linguistic features of the child corpus are likely to guide our understanding of the language variation present in the parent corpus, as the MTLD results suggested. In sum, I established in this paper a new type of a corpus relationship that may affect our understanding of how two corpora could relate to one another.

The uber index findings agree with Arroyo and Pérez (2019) in terms of the appropriacy of the chosen N-gram length. On the one hand, the use of bigrams readily facilitated the observation of the building blocks, including functional and content word clusters corresponding to the legal discourse in general, and the genre of business contracts in particular. On the other hand, the keywords that were peculiar to each contract category were simultaneously obtained. The adequate use of the N-gram length also enabled an appropriate positioning of the current study on the spectrum premise proposed by Mautner (2012), as I originally aimed to establish a point on the general end of the indices spectrum. My goal was successfully met through the use of the uber index. That is, irrespective of the contract type, I was able to determine the building blocks of business recitals in general.

In response to Miller and Biber's (2015) contention, I maintain the validity of the produced wordlist on the reliability of the measures. Prior to making validity claims of the produced indices, I assured that the repeatability of the LCH procedure would yield approximate results. Thus, the entire procedure was replicated for each corpus twice using both the super and uber indices. Furthermore, as indicated in the introduction section, the LCH approach was based on the premises of the corpus saturation parameter proposed by McEnery and Wilson (2001). However, while the authors adopted a top-down approach in measuring the finiteness of the target

corpora, I adopted a bottom-up approach for the LCH procedure. In all respects, the goal of measuring the stability of the lexical features constituting BO and COMB was met.

In conclusion, the super/uber index is useful for studies based on discourse analysis. Specifically, the indices offer benefits to ESP/EAP studies that are dealing with moves analysis as identifying HFBs could facilitate our understanding of the discourse functions behind certain word choices, and the frequencies at which they occur. More importantly, the LCH presents itself as an indispensable tool for extracting the most consistently occurring words in a corpus, leading to the identification of fixed phrases that are typical of a certain discourse. Specifically, this tool seems to be appropriate for discourses that "to a large degree created and strengthened by phraseology—fixed recurrent patterns at the textual, grammatical, and collocational level." (Biel, 2017, p. 177). For this reason, the legal discourse was, specifically, opted for analysis.

6.2 Implications

The current study offers important pedagogical as well as methodological implications. With reference to the parent-child corpora relation findings, in an ELP course, the recitals of COMP are suggested to be taught first since they allow the use of less diversified words, as MTLD results also suggest. Furthermore, the phraseological units list could be employed for assessment purposes. The ELP instructor could assess whether the students met certain writing standards corresponding to the conventions of writing recitals. This evaluation could be conducted through measuring the degree to which students' production contains some of the lexical frames provided in the list.

Instead of manual evaluations, instructors are recommended to use The Discipline Specificity Analysis Tool (DSAT). This tool screens the student's production by measuring the percentage of recitals indicative language inclusion in the uploaded paper. The tool provides the feedback in the form of suggestions and interpretations based on detecting the problematic segments not addressing the written conventions of the discipline. The tool also shows the segments where appropriate indicative language use was observed. Accordingly, instructors could create peer-review practices or feedback sessions using (DSAT).

With respect to the methodological implications, the study stated in great detail the steps to using Gramulator in case prospective researchers in other

disciplines wished to perform a similar investigation. Accordingly, the steps of compiling and measuring the lexical consistency of a specialized corpus as well as the process of extracting the word list in a convenient fashion using the software are all clearly outlined. Finally, the study provided a methodological contribution to the field of corpus linguistics. The proposed LCH addresses many concerns demonstrated in previous relevant work.

6.3 Limitations and Future Work

One of the limitations of the current study pertains to the categorization of the contract types. The contracts were classified according to an element that was common to each contract category. However, the categorization in question was not based on the legal implications of each contract type under *contract law*. Another methodological concern worth reporting was the sample size of BO corpus. More BO-based contracts were unobtainable and since many of them were eliminated because of the sampling frame restrictions, only 51 contracts remained eligible for inclusion. In fact, previous studies have reported that t-test results reliability may stay unaffected for small sample size. For example, Winter (2013) confirms t-test robustness for a small sample size ($N \leq 5$). Therefore, BO sample size may not represent a limitation that would affect the findings validity. Nevertheless, this concern was still addressed through replications of the statistical procedures to ensure reliable results. Given also the MTLD sensitivity to texts below 100 tokens, it would have been insightful to check the diversity of both corpora beyond this restriction.

Finally, because of practical constraints, this thesis does not engage in an in-depth qualitative analysis of the proposed phraseological units. Therefore, further work is required to establish the structure patterns as well as the functions served by the bigrams.

6.4 Conclusion

The current work adopts a corpus induced form (Lee, 2008) as a corpus assisted discourse study, where quantitative methods were mainly employed in the investigation. The main aim was to illuminate an uncharted research area in the genre of contract drafting to assist novice legal drafters in their quest for writing recitals with precision. Language is a powerful tool that could be employed to maximize clarity in our writing or leave perplexed readers attempting to account for ambiguities. Such a scenario is inescapable when

drafting in the legal context. Therefore, providing new law students with a list of fixed phrases they could adapt in accordance with the specificity of the contractual arrangement is a good start in an ELP course. However, to make such a proposed list, it was important to comply with certain standards in compiling and validating the target corpora based on insights from previous seminal vocabulary studies.

References

Adams, K. A. (2018). *A manual of style for contract drafting*. American Bar Association.

Alasmary, A. (2019). Lexical bundles in contract law texts: A corpus-based exploration and implications for legal education. *International Journal of English Linguistics*, *9*(2), 244–257. https://doi.org/10.5539/ijel.v9n2p244

Arroyo, B., & Pérez, L. M. (2019). Lexical chunks in English and Spanish sales contracts. *Terminology*, *25*(1), 32–59. https://doi.org/10.1075/term.00027.lop

Basaneze, K. (2017). Investigating "concgrams" in the language of contracts and legal agreements. *Fachsprache*, *37*(3-4), 176–192. https://doi.org/10.24989/fs.v37i3-4.1289

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & society*, *19*(3), 273-306.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English*. London: Longman

Biel, L. (2017). Phraseology in legal translation: A corpus-based analysis of textual mapping in EU Law. In Cheng, L., Sin, K. K., & Wagner, A. (Eds.), *The Ashgate handbook of legal translation* (pp. 177-192). Routledge.

Breeze, R. (2013). Lexical bundles across four legal genres. *International journal of corpus linguistics*, *18*(2), 229-253.

Brezina, V. (2018). Statistical choices in corpus-based discourse analysis. In Marchi, A. and Taylor, C. (Eds.), *Corpus approaches to discourse: A critical review* (pp. 259-280). Routledge

Cao, M., & Zhao, W. (2019). A corpus-based analysis of business contract English registers. *International Journal of English Linguistics*, *9*(3), 158–163. https://doi.org/10.5539/ijel.v9n3p158

Connor, U., & Upton, T. A. (2004). *Discourse in the professions: Perspectives from corpus linguistics*. John Benjamins.

Conrad, S., & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*, *20*, 56–71. https://doi.org/10.1515/9783484604674.56

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*(4), 397–423. https://doi.org/10.1016/j.esp.2003.12.001

De Geaugrande, R. (2001). Large corpora, small corpora, and the learning of language. In Ghadessy, M., Henry, A., & Roseberry, R. L. (Eds.). *Small corpus studies and ELT theory and practice (pp. 3-28)*. Benjamins.

Denoual, E. (2006). A method to quantify corpus similarity and its application to quantifying the degree of literality in a document. *International Journal of Technology and Human Interaction*, *2*(1), 51–66. https://doi.org/10.4018/jthi.2006010104

Efe, I., & Ozer, O. (2015). A corpus-based discourse analysis of the Vision and mission statements of universities in Turkey. *Higher Education Research & Development*, *34*(6), 1110–1122. https://doi.org/10.1080/07294360.2015.1070127

Ghadessy, M., Henry, A., & Roseberry, R. L. (2001). *Small corpus studies and Elt theory and practice*. Benjamins.

Governatori, G., & Milosevic, Z. (2006). A formal analysis of a business contract language. *International Journal of Cooperative Information Systems*, *15*(04), 659–685. https://doi.org/10.1142/s0218843006001529

Greaves, C., & Warren, M. (2012). Building a written corpus: What can a corpus tell us about a multi-word units? In O'Keeffe, A., & McCarthy, M. (Eds.), *The Routledge handbook of corpus linguistics* (pp. 35-47). Routledge, Taylor & Francis Group.

Gries, S. T. (2006). Exploring variability within and between corpora: Some methodological considerations. *Corpora*, *1*(2), 109–151. https://doi.org/10.3366/cor.2006.1.2.109

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, *6*(1), 97–133. https://doi.org/10.1075/ijcl.6.1.05kil

Koester, A. (2012). Building small specialised corpus. In O'Keeffe, A., & McCarthy, M. (Eds.), *The Routledge handbook of corpus linguistics* (pp. 35-47). Routledge, Taylor & Francis Group.

Kopaczyk, J., & Tyrkkö, J. (2018). Blogging around the world: Universal and localised patterns in online Englishes. In Kopaczyk, J., & Tyrkkö, J. (Eds.), *Applications of pattern-driven methods in corpus linguistics* (pp. 277–308). John Benjamins Publishing Company.

Lee, D. (2008). New ways of doing old things. In Bhatia, V. K., Flowerdew, J., & Jones, R. H. (Eds.), *Advances in discourse studies* (pp. 86-99). Routledge.

Łukasik, M. (2014). Compiling a corpus for terminographic purposes. *Komunikacja Specjalistyczna*, *7*, 71-83.

Mahlberg, M. (2014). Corpus linguistics and corpus analysis . In Schneider, K. P., & Barron, A. (Eds.), *Pragmatics of discourse* (pp. 189–213). De Gruyter Mouton.

Marchi, A. and Taylor, C. (2018). Introduction: Partiality and reflexivity. In Marchi, A. and Taylor, C. (Eds.), *Corpus approaches to discourse: A critical review* (pp. 1-15). Routledge

Mauntner, G. (2015). Checks and balances: how corpus linguistics can contribute to CDA. In Meyer, M., & Wodak, R. (Eds.)., *Methods of critical discourse studies* (pp. 154-179). Sage.

Mautner, , G. (2012). What can a corpus tell us about a discourse? In O'Keeffe, A., & McCarthy, M. (Eds.), *The Routledge handbook of corpus linguistics* (pp. 35-47). Routledge, Taylor & Francis Group.

McCamus, J. (2020). *The law of contracts, 3/E*. Irwin Law.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, VOCD-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/brm.42.2.381

McCarthy, P. M., Watanabe, S., & Lamkin, T. A. (2012). The Gramulator: A Tool to Identify Differential Linguistic Features of Correlative Text Types. In P. McCarthy, & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation and resolution* (pp. 312-333). IGI Global. https://doi.org/10.4018/978-1-60960-741-8.ch018

McEnery, T., & Brookes, G. (2012). Building a written corpus: What are the basics? In O'Keeffe, A., & McCarthy, M. (Eds.), *The Routledge handbook of corpus linguistics* (pp. 35-47). Routledge, Taylor & Francis Group.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: method, theory, and practice*. Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction*. Edinburgh University Press.

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Miller, D., & Biber, D. (2015). Evaluating reliability in quantitative vocabulary studies. *International Journal of Corpus Linguistics*, *20*(1), 30–53. https://doi.org/10.1075/ijcl.20.1.02mil

Neufeldt, V. (2019). DCHP-2: The dictionary of Canadianisms on historical principles, second edition ed. by Stefan Dollinger. *Journal of the Dictionary Society of North America*, *40*(2), 242–246. https://doi.org/10.1353/dic.2019.0025

Oluwatayo, J. (2012). Validity and reliability issues in educational research. *Journal of Educational and Social Research, 2*(2), 391–400. https://doi.org/Doi:10.5901/jesr.2012.v2n2.391

Park, K., Hong, J. S., & Kim, W. (2020). A methodology combining cosine similarity with classifier for text classification. *Applied Artificial Intelligence*, *34*(5), 396–411. https://doi.org/10.1080/08839514.2020.1723868

Partington, A., & Marchi, A. (2015). Using corpora in discourse analysis. In Biber, D., & Reppen, R. (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 216 – 234). Cambridge University Press.

Powell, W., Foth, M., Cao, S., & Natanelov, V. (2022). Garbage in garbage out: The precarious link between IoT and blockchain in food supply chains. *Journal of Industrial Information Integration*, *25*, 100261.

Rayson, P., Leech, G. N., & Hodges, M. (1997). Social differentiation in the use of English vocabulary. *International Journal of Corpus Linguistics*, *2*(1), 133–152. https://doi.org/10.1075/ijcl.2.1.07ray

Rose, T., Haddock, N., & Tucker, R. (1997). The effects of corpus size and homogeneity on language model quality. *Proceedings of the ACL SIGDAT Workshop on Very Large Corpora*, pp. 178–91, Hong Kong

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.

Sinclair, J. (2005). Corpus and text – basic principles. In Wynne, M (Eds.), *Developing linguistic corpora a guide to good practice* (pp. 1-16). Oxbow Books.

Sinclair, J. (1991). *Corpus, concordance, Collocation*. Oxford University. Press.

Sinclair, J., & Carter, R. (2004). *Trust the text : language, corpus and discourse*. Routledge

Vaughan, E., & Clancy, B. (2013). Small corpora and pragmatics. In J. Romero-Trillo (Eds.), *Yearbook of corpus linguistics and pragmatics*, (pp. 53–73). Springer

Williams, G. (2002). In search of representativity in specialised corpora. *International Journal of Corpus Linguistics*, *7*(1), 43–64. https://doi.org/10.1075/ijcl.7.1.03wil

Winter, J. (2013). Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation*, *18*(10), 1–12.

Xiao, R. (2010). Corpus Creation. In Indurkhya, N., & Damerau, F. J. (Eds.). *Handbook of natural language processing* (pp. 149-165). Chapman and Hall.

Vita

Reem Kamal Rabea was born in 1995, in Alexandria, Egypt. She studied literature in high school and received her education in 2012. She received her B.A. degree in international logistics management. She graduated from Altinbas University as valedictorian in 2017.

Ms. Reem taught academic English at university level for three years in Turkey. In December 2020, she began her master's program in TESOL (Teaching English to Speakers of Other Languages) in the American University of Sharjah as a graduate research assistant. Her research interests are in English for legal and professional purposes.