

A comparative Analysis of Lexical Cohesion in Native and Non-Native
Speaker Writing: Text Linguistics and Corpus Perspectives

Eman Yonos

A thesis submitted to The University of Birmingham for the degree of
DOCTOR OF PHILOSOPHY

Department of English Language & Applied
Linguistics

School of English, Drama and American &
Canadian Studies

Collage of Arts and Law

The University of Birmingham

2019

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Abstract

Lexical cohesion is a major challenge for L2 writers. This study compares the frequency and function of forms of lexical cohesion in English argumentative writing by English native speakers (NSs) and Arab non-native speakers of English (NNSs) to gain insights into the specific challenges that Arab learners face. It overcomes some of the limitations of the classic models of lexical cohesion for the analysis of texts, and suggests a systematic approach to analysing lexical cohesion that is replicable across corpora. The study further explores the potential of a corpus approach to studying the functions of lexical cohesion.

The analysis identifies the frequency of ‘simple repetition’ and ‘derived repetition’ using wordlists, which pinpoint lexical cohesive networks in two corpora of argumentative essays, whereas ‘signalling nouns’ are quantified using a manual text analysis. The frequency results are interpreted by means of a text analysis to examine the paradigmatic choice of lexical cohesive relations and how they are associated at the non-linear level in both corpora. The text analysis is complemented by a corpus analysis to identify semantic preferences and prosodies of selected lexical items in both corpora. The frequency analysis showed that high frequency of lexical cohesion does not indicate good writing. The text analysis revealed that Arab speakers of English used lexically cohesive forms redundantly without using them to develop their argument across the text. The corpus analysis indicated that the Arab NNS essays diverged from the typical semantic field of a lexical item, which disrupted the cohesive structure of the writing. This study suggests that a combination of text analysis and corpus analysis provides a fuller picture of how NNSs use lexical cohesion at both the paradigmatic and syntagmatic levels. A better understanding of L2 writers’ use of cohesive devices will have implications for teaching lexical cohesion in the L2 writing classroom.

Acknowledgements

I would like to express my profound gratitude to my supervisor Professor Michaela Mahlberg for her patience, guidance and for sharing her expertise throughout my PhD journey. She has continually given me constructive suggestions and positive encouragement. Without her invaluable input, I could not have completed this research. I would also like to thank the examiners of my thesis, Dr. Nicholas Groom and Prof. Dan McIntyre for their insightful comments and constructive feedback, which helped me to improve this work. I would also like to extend my thanks to Dr. Jeannette Littlemore for her encouragement on the day of my viva.

I am also grateful to all wonderful staff in Hallward Library at the University of Nottingham where I wrote this PhD. I would also like to thank my parents, Ali and Aisha, for spiritually supporting me. Last but not least, my special thanks go to my husband for his patience, understanding and for his unfailing enthusiasm and support. I must not forget to mention my beloved daughter Leen, who has always been there cheering me on.

Table of Contents

Table of contents.....	ii
List of Figures.....	vii
List of Tables.....	viii
List of Abbreviations.....	ix

Chapter 1: Introduction

1.1 Introduction.....	1
1.2 The interplay between cohesion and coherence.....	4
1.3 The function of lexical cohesion.....	7
1.4 The importance of comparing NSs and NNSs in learner corpus research.....	8
1.5 Lexical cohesion in text linguistics and corpus linguistics.....	11
1.6 Research objectives and research questions.....	16
1.7 Organisation of the thesis.....	18

Chapter 2: Lexical cohesion in text-linguistics

2.1 Introduction.....	20
2.2 Halliday & Hasan's (1976) model of lexical cohesion.....	21
2.2.1 Reiteration.....	22
2.2.2 Collocation.....	23
2.3 Hasan's (1984) model of lexical cohesion.....	25
2.4 Hoey's (1991b) model of lexical cohesion.....	26
2.5 Other categories of lexical cohesion.....	30
2.6 Categories of lexical cohesion in the present study: an overview.....	31
2.6.1 A working definition of simple repetition in this study.....	34
2.6.2 A working definition of derived repetition in this study.....	36
2.6.3 Signalling nouns and their working definition in this study.....	39
2.7 Criteria for determining SNs in the present study.....	41
2.7.1 Signalling nouns and repetition.....	44
2.7.1.1 Signalling nouns and simple repetition.....	46
2.7.1.2 Signalling nouns and nominalisation.....	53

2.7.1.3 Signalling nouns and synonyms.....	56
2.7.1.4 Signalling nouns and text nouns.....	58
2.7.1.5 Partitives.....	61
2.8 Lexical cohesion in native and non-native English writing: A special reference to Arab speakers of English.....	63
2.8.1 Simple and derived repetition in native and non-native writing.....	64
2.8.1.1 Studies on Arabic speakers' English writing.....	64
2.8.1.2 Studies that compare NNS and NS writing.....	68
2.8.2 Signalling nouns in native and non-native writing.....	82
2.9 Applying classic models of lexical cohesion to text analysis.....	89
2.9.1 The notion of a tie and directionality of lexical cohesion.....	90
2.9.2 The boundary of the lexical item.....	97
2.9.3 The cross-categorisation of lexical cohesion.....	99
2.9.4 The selection of open-set words and the frequency factor.....	103
2.10 Conclusion.....	108

Chapter 3: Lexical Cohesion and Corpus Linguistics

3.1 Introduction.....	110
3.2 The definition of a lexical item according to corpus linguistics.....	111
3.3 Lexical cohesion in non-learner corpora from a corpus-linguistic perspective.....	116
3.4 Lexical cohesion in learner corpora from a corpus-linguistic perspective.....	122
3.5 Conclusion.....	128

Chapter 4: Methodology

4.1 Introduction.....	130
4.2 Building the Arab Learner English Corpus (ALEC).....	131
4.3 Ethical considerations.....	135
4.4 Finding a reference corpus of native data in this study.....	135
4.5 LOCNESS as reference corpus.....	138
4.6 Comparability between ALEC and LOCNESS.....	141
4.7 Combining quantitative and qualitative approaches in the analysis of lexical cohesion.....	143
4.7.1 The framework for the quantitative analysis.....	144
4.7.2 The framework for the qualitative analysis.....	145
4.8 The T-unit and the problem of punctuation in L2 writing.....	146

4.9 Exclusion and inclusion criteria for determining simple and derived repetition.....	150
4.10 Co-reference as a criterion in lexical cohesion analysis.....	152
4.11 Conclusion.....	159

Chapter 5: Quantitative analysis and results: Corpus-based analysis and text analysis

5.1 Introduction.....	160
5.2 The Lexical Repetition Network Model (LRNetM).....	161
5.3 The definition of ‘network’	164
5.3.1 Characteristics of a lexical repetition network within LRNetM.....	165
5.3.2 Lexical repetition networks compared to similar lexical networks in existing literature.....	168
5.4 Grouping a wordlist into lexical repetition networks.....	170
5.5 Types of lexical repetition networks within the LRNetM framework.....	173
5.6 Counting frequencies of simple and derived repetitions in lexical repetition networks.....	175
5.6.1 The length of the lexical repetition network.....	181
5.6.2 Frequency of simple repetition in a simple repetition network.....	182
5.6.3 Frequency of derived repetition in a derived repetition network.....	184
5.6.4 Calculating the total frequencies of simple and derived repetition in the lexical repetition networks in an individual essay.....	186
5.7 Comparing frequency counts of lexical cohesion between the classic models of lexical cohesion and the lexical repetition network model.....	187
5.8 Counting signalling nouns in ALEC and LOCNESS.....	205
5.9 Results.....	207
5.9.1 Frequency (tokens) of simple and derived repetition in ALEC and LOCNESS.....	209
5.9.2 Frequency (tokens) of signalling nouns in ALEC and LOCNESS.....	213
5.10 Conclusion.....	215

Chapter 6: A qualitative look at the results: A text-linguistic analysis and discussion

6.1 Introduction.....	217
6.2 Simple repetition.....	219
6.2.1 Text analysis of simple repetition in ALEC.....	220
6.2.2 Text analysis of simple repetition in LOCNESS.....	227
6.2.3 Simple repetition and lexical redundancy.....	233

6.2.4 Simple repetition as a cultural and rhetorical device in Arabic language.....	234
6.3 Derived repetition.....	238
6.3.1 Text analysis of derived repetition in ALEC.....	239
6.3.2 Text analysis of derived repetition in LOCNESS.....	247
6.4 Signalling nouns.....	254
6.5 Reconsidering the counting method of SNs in this study.....	257
6.5.1 Structure one (In-t-unit SN + across-t-unit SN).....	258
6.5.2 Structure two (Lexical couplets or strings).....	262
6.6 Other patterns of repetition in the use of signalling nouns by Arab speakers of English: Text analysis.....	266
6.7 The most common types of signalling nouns in ALEC and LOCNESS.....	273
6.8 Signalling nouns as devices for signalling the Problem-Solution pattern in ALEC and LOCNESS.....	280
6.9 General discussion on signalling nouns.....	282
6.10 Conclusion.....	284

Chapter 7: A corpus linguistic perspective on lexical cohesion: A qualitative analysis and discussion

7.1 Introduction.....	287
7.2 Lexical items selected for semantic preference and semantic prosody analyses in this study.....	290
7.3 Matching individual texts against a reference corpus: A corpus approach to textual cohesion.....	291
7.4 The textual behaviour of REDUCE in a reference corpus.....	294
7.4.1 The textual behaviour of REDUCE in LOCNESS.....	300
7.4.2 The textual behaviour of REDUCE in ALEC.....	303
7.5 The textual behaviour of AFFECT in a reference corpus.....	309
7.5.1 The textual behaviour of AFFECT in LOCNESS.....	315
7.5.2 The textual behaviour of AFFECT in ALEC.....	319
7.6 The textual behaviour of FACE in a reference corpus.....	322
7.6.1 The textual behaviour of FACE in LOCNESS.....	325
7.6.2 The textual behaviour of FACE in ALEC.....	328
7.7 Lexical cohesion through the interlocking of textual functions of lexical items.....	330
7.8 Conclusion.....	338

Chapter 8: Conclusions and implications for pedagogy

8.1 Introduction.....	341
8.2 Summary of the results.....	341
8.2.1 The frequency of lexical cohesion.....	342
8.2.2 The function of lexical cohesion.....	344
8.2.2.1 Text analysis.....	344
8.2.2.2 Corpus linguistics.....	346
8.3 Contribution of my thesis.....	347
8.4 Pedagogical implications for teaching lexical cohesion in an L2 writing classroom.....	351
8.4.1 L2 learners need to recognise that lexical cohesion should provide a framework for new information and not to be used redundantly.....	352
8.4.2 L2 learners need to be aware of the function of semantic prosody in creating textual cohesion.....	357
8.4.3 The application of a corpus linguistic approach to teaching cohesion in the L2 writing classroom.....	358
8.5 Directions for future research.....	363
8.6 Conclusion.....	368
Appendices.....	370
Appendix 1: Essay prompts and writing instructions template.....	370
Appendix 2: Learner profile template.....	372
Appendix 3: Participant Consent Form template.....	374
Appendix 4: ALEC and LOCNESS British A-level students' essays sub-corpus CD-ROM.....	376
Appendix 5: A complete essay from ALEC showing the intense use of commas.....	377
Appendix 6: An analysis of lexical cohesion in a complete essay from ALEC applying Halliday & Hasan's (1976) model.....	379
Appendix 7: An analysis of lexical cohesion in a complete essay from ALEC applying Hoey's (1991b) repetition matrix CD-ROM.....	382
References.....	383

List of Figures

Figure 1 Linear representation of a chain.....	28
Figure 2 Web representation of a chain.....	28
Figure 3 The connection between T-unit 1 and all successive T-units.....	75
Figure 4 A sample of Hoey’s (1991b) coding matrix of repetition links applied by Reynolds (1995).....	76
Figure 5 Hoey’s (1991b) identification of a network of repetition links between lexical items across sentences.....	94
Figure 6 A screenshot of the CLAWS Web POS Tagger.....	172
Figure 7 Sample of lexical repetition networks in an Arab NNS essay.....	173
Figure 8 An illustration of lexical repetition networks in an NS essay from LOCNESS, ranked in alphabetical order.....	176
Figure 9 An analysis of lexical cohesion in an NNS essay from ALEC applying Hoey’s (1991b) repetition matrix.....	191
Figure 10 An analysis of lexical cohesion in a complete essay from ALEC applying the LRNetM model.....	192
Figure 11 Individual mean frequencies of simple repetition, derived repetition and SNs in ALEC & LOCNESS.....	214
Figure 12 Arabic thought.....	236
Figure 13 English thought.....	236
Figure 14 Sample of concordance lines of REDUCE in the ‘traffic’ reference corpus.....	295
Figure 15 Concordance lines of REDUCE in a native text	300
Figure 16 Concordance lines of REDUCE in an Arabic non-native text.....	304
Figure 17 Sample of concordance lines of AFFECT in the ‘traffic’ reference corpus.....	310
Figure 18 Sample of concordance lines of FACE in the ‘immigration’ reference corpus.....	324
Figure 19 REDUCE, AFFECT and FACE cohesive networks in a native text from LOCNESS.....	331
Figure 20 REDUCE, AFFECT and FACE cohesive networks in an Arab non-native essay from ALEC.....	335
Figure 21 A Lexical Repetition network of an Arab L2 essay	362

List of Tables

Table 1 Hoey’s (1991b) categories of lexical repetition.....	30
Table 2 Different strategies of creating lexical cohesion in a text.....	31
Table 3 Categories of lexical cohesion in the present study.....	33
Table 4 Summary of lexical cohesion studies on Arab L2 English writing.....	65
Table 5 Summary of cohesion studies on NS and NNS English writing.....	69
Table 6 Overview of research into SNs in NS and NNS English writing.....	83
Table 7 Summary of the basic variables in the learner profiles by Arab L2 learners.....	133
Table 8 The composition of LOCNESS	139
Table 9 Topics of argumentative essays in LOCNESS sample and ALEC.....	140
Table 10 A comparison between the LRNetM model and the classic models of lexical cohesion.....	163
Table 11 An example of a simple repetition network in an NS essay from LOCNESS.....	182
Table 12 An example of a derived repetition network in an NS essay from LOCNESS	184
Table 13 An example of a lexical network that combines simple & derived repetition in an NS essay in LOCNESS.....	185
Table 14 An analysis of lexical cohesion in an NNS essay from ALEC applying Halliday & Hasan’s (1976) model.....	190
Table 15 Comparing frequency counts of lexical cohesion applying the classic models of lexical cohesion & the LRNetM model in a complete essay from ALEC.....	193
Table 16 Frequency of simple and derived repetition per thousand words in ALEC and LOCNESS.....	210
Table 17 Frequency of SNs per thousand words in ALEC & LOCNESS.....	213
Table 18 Overall frequencies of the ‘in-t-unit SN + across-t-unit SN’ structure per thousand words in ALEC & LOCNESS.....	259
Table 19 Individual text-based mean frequencies of SNs per thousand words in ALEC & LOCNESS excluding the ‘in-t-unit SN + across-t-unit SN’ structure.....	260
Table 20 Overall frequencies of lexical couplets per thousand words in ALEC & LOCNESS.....	262
Table 21 Individual text-based mean frequencies of SNs per thousand words in ALEC & LOCNESS excluding the ‘lexical couplets’ structure.....	263

List of Abbreviations

ALEC: Arab Learner English Corpus

CLAWS: Constituent Likelihood Automatic Word-tagging System

ESL: English as a second language

EFL: English as a foreign language

L1: First language

L2: Second language

SLA: Second language acquisition

SNs: Signalling nouns

SR: Simple repetition

DR: Derived repetition

TOEFL: Test of English as a foreign language

IL: Interlanguage

ICLE: The International Corpus of Learner English

LOCNESS: The Louvain Corpus of Native English Essays

NSs: Native speakers of English

NNSs: Non-native speakers of English

Chapter 1

Introduction

1.1 Introduction

Foreign and second language learners need to be aware of the links that hold chunks of text together and that contribute to the creation of a text as a unit of meaning. L2 learners could learn this skill by using ‘cohesion’. The term ‘cohesion’ is used to refer to the property of connectedness that characterises a text in contrast to a mere sequence of words (Halliday and Hasan 1976). Mahlberg (2006) suggests that language learners need to use cohesive forms appropriately to achieve a native-like level of writing. She further stresses that cohesion makes the text easy to read and helps writers to produce a clear argument. ‘Cohesion’ is different from ‘coherence’, where the former refers to the formal features on the surface of a text and the latter characterises the underlying meaning of a text. Section 1.2 provides definitions of both ‘cohesion’ and ‘coherence’ and how they interact to produce a unified text. Halliday and Hasan (1976) classify cohesion into grammatical and lexical. Lexical cohesion is the main focus of the present study. Hoey (1991b) suggests that the study of lexical cohesion is chiefly the study of repetition and semantically related sets of lexical items. Lexical repetition does not cover only the repetition of the same lexical item but there are other degrees of lexical relations that enable repetition to take place. A case in point is the use of signalling nouns (Flowerdew 2003).

The use of lexical cohesion poses a challenge for writers, whether they are writing in their first or additional language (Flowerdew 2009). Scarcella (1984) finds that native speaker writers use a variety of grammatical and pragmatic techniques of cohesion and coherence

more successfully than non-native speaker writers who prefer same item repetition to connect their sentences. McCarthy (1991) claims that reiteration using same item repetition is not always found in English discourse, whereas substantial variation from sentence to sentence in writing is more common. Despite the difficulties that L2 learners encounter with lexical cohesion, it has been overlooked in English language teaching (Flowerdew 2006). This situation is certainly true with reference to English language teaching in Arabic countries, where the focus is placed on teaching rules of grammar rather than discourse competence. Although some L2 teachers teach students how to link sentences together through the use of cohesive devices, they introduce this in the form of a mechanical test. In this test, students need to look, for example, at a number of highlighted words in the text and match them with their synonyms. Alternatively, students are asked to group words according to categories such as, vehicles: *car*, *motorcycle*; or they have to match adjectives with their opposites, e.g., *happy* – *unhappy*. These exercises do not help L2 learners to create a unified text. Therefore, any pedagogical treatment of lexical cohesion needs to make Arab speakers of English aware that writing meaningful and coherent texts in English requires learning more than grammar knowledge. Such an aim could be achieved through instructing L2 learners that cohesive forms contribute to only one part of the total unity of a text, and it is necessary for L2 learners to know how to employ cohesive forms functionally. In this regard, McCarthy and Carter (2014: 175) suggest incorporating lexical cohesion into the syllabus at the discourse level. They claim that lexical cohesion is one of the linguistic elements of the language system, and it is as a result possible to be taught as language knowledge. This indicates that teaching lexical cohesion should not only focus on teaching learners what the synonyms and hyponyms of a certain word or group of words are, but also explaining how these cohesive forms

function in text to create a successful writing. Section 1.3 will highlight the function of lexical cohesion.

While there are a number of studies that have been conducted to analyse lexical cohesion in the writing of Arab speakers of English, these studies focus on forms and frequency of lexical cohesion rather than on how cohesion functions in text. These studies are principally concerned with quantifying the number of cohesive forms in the English writing of non-native speakers without an adequate attempt to analyse this frequency data qualitatively. Additionally, studies conducted by researchers such as Khalil (1989) and El-Gazzar (2006) do not use a reference corpus of native English writing against which they can compare the English writing by Arab L2 learners. Section 1.4 will highlight how important it is to compare native and non-native speakers in learner corpus research. My research will further identify some of the limitations of the classic models of lexical cohesion for the analysis of texts. Based on this examination, I will suggest a model of analysis that is based mainly on a corpus-linguistic perspective, and which is capable of capturing cohesive forms in text more systematically than the classic models (see Chapter 5 for an explanation of this model). There are no studies that have used corpus analysis in examining lexical cohesion in the English writing of Arab speakers. Therefore, an evaluation of lexical cohesion from a corpus-linguistic perspective is required. Section (1.5) underlines the value of combining text linguistics with corpus linguistics to describe the frequency and function of lexical cohesion.

Overall, to the best of my knowledge there are no comparative studies that have analysed the patterns, the frequency and the functions of lexical cohesion in English written texts by Arab speakers of English and compare them with English native speaker writing. The present study

will help address this gap in the literature by gaining insights into the specific challenges that Arab learners face when they use lexical cohesion compared to native speakers. It might also help understand Arabic learners' interlanguage system, namely interlanguage cohesion. The present research therefore aims to improve the teaching of writing to Arab speakers of English by suggesting pedagogic intervention in the use of lexical cohesion. It will particularly suggest an application of corpus theoretical concepts to teaching lexical cohesion in an L2 classroom. This research is, as a result, constructive for both English language teachers and learners and for research into areas related to second language writing, interlanguage, learner corpora and corpus linguistics.

1.2 The interplay between cohesion and coherence

Even though the present study focuses on the role of cohesion in text in the first place, the analysis will in a few cases highlight how cohesion can signal coherence (see Chapter 6). Researchers agree that cohesion and coherence are different, but there is disagreement on what differentiates the two. Widdowson (1978: 29), for example, distinguishes between cohesion and coherence, arguing that 'cohesion' is "the overt, linguistically-signalled relationship between propositions", while 'coherence' refers to "the illocutionary function of these propositions, with how they are used to create different kinds of discourse" (Widdowson 1978: 52). He acknowledges that all discourse can be characterised in terms of the relationship between cohesion and coherence. However, he claims that some texts are still coherent in spite of the complete absence of cohesion. As Seidlhofer and Widdowson (1999: 207) put it, one can "derive a coherent discourse from a text with no cohesion in it at all". Brown and Yule (1983: 196) argue that we still can identify a text as a text if it contains few, if any, explicit markers of cohesive relationships. They further allege that formal cohesion

alone is not sufficient to guarantee the production of a text. Enkvist (1978: 10) also contests that cohesive ties are inadequate for guaranteeing textness. De Beaugrande & Dressler (1981: 4) further maintain that although overt markers of cohesion are reliable clues, “[t]he surface is [...] not decisive by itself; there must be interaction between cohesion and other standards of textuality to make communication efficient”. For de Beaugrande and Dressler (1981), ‘cohesion’ and ‘coherence’ constitute two of their seven parameters of textuality. In their model, ‘cohesion’ includes the techniques by which the surface features of the text function as progressive occurrences to ensure continuity. ‘Coherence’, on the contrary, is concerned with the techniques that help to maintain the connectedness of concepts in a text (de Beaugrande & Dressler 1981: 3–10). These researchers, however, do not completely deny the role of cohesion in establishing unity in discourse as long as it is differentiated from coherence. Nonetheless, Morgan and Sellner (1980) give little value to the importance of cohesion in interpreting the text. Carrell (1982) agrees with Morgan and Sellner (1980) and affirms that ‘cohesion’ is just an illusion; “[t]he illusion of lexical cohesion is created by the text’s coherence” Carrell (1982: 484).

There are views which regard ‘cohesion’ and ‘coherence’ as separate but related concepts. Halliday and Hasan (1976) argue that cohesive forms alone are not enough to provide an interpretation for the text; the text requires also coherence, because the two together successfully characterise a text. These researchers explain that a text is “coherent with respect to the context of situation, and it is coherent with respect to itself, and therefore cohesive” Halliday and Hasan (1976: 23). What seems to be clear is that to Halliday and Hasan cohesion, being limited to the text, is more restricted than that of coherence, which also involves the context. They consider cohesion as a necessary property for coherent discourse.

In a later work, Hasan (1984) highlights that ‘coherence’ is a feature that can be measured by the reader of a text. The recognition of coherence is based on the interaction of cohesive forms, which she calls ‘cohesive harmony’; the text is more coherent when the cohesive harmony of a text increases. She argues that cohesion is perhaps not a requirement for coherence. Nonetheless, a text which contains cohesive devices is expected to be more coherent than one without them. Hasan (Halliday and Hasan 1989: 94) claims that “cohesion is the foundation upon which the edifice of coherence is built. Like all foundations, it is necessary but not sufficient by itself”. Daneš (1974) also stresses that lexical cohesion is the main contributor to the coherence of text. He, however, looks at ‘coherence’ purely from a linguistic perspective. His view does not focus on the addressee’s knowledge of the context and situation, but concentrates on the formal linguistic matters. Daneš suggests that the extent of coherence in the text depends on the degree of the continuity of the thematic progression (see Chapter 6 for an example of this case). Successive units need to have a link between them, in the form of similar linguistic elements. If this link is missing, a discontinuity will occur in the progression, resulting in a break in the text’s cohesion.

There is no simple consensus on the relationship between cohesion and coherence. However, the crucial point to stress in the present study is not to find the correlation between cohesion and coherence, but to examine whether high frequency of lexical cohesive markers in NNS/NS writing indicates a unified coherent text. This point has been investigated by a number of researchers such as Hartnett (1986) who observes that the only existence of cohesive forms does not necessarily lead to successful writing. She maintains that researchers should not link overall writing quality to the quantity of cohesive devices. They should instead give value to the function or aptness of these devices. Witte and Faigley (1981)

observe that quantitative measures of cohesive ties cannot serve as simple indices of writing quality. This study will therefore adopt the view that successful communication involves both cohesion and coherence, which are separate but interlinked, as Tanskanen (2006) describes. ‘Cohesion’ in my study is manifested through surface features. However, these cohesive forms have to serve a function. This function is to develop the argument of the text, as Widdowson (1978: 27) claims “a discourse is cohesive to the extent that it allows for effective propositional development”.

1.3 The function of lexical cohesion

Hoey (1991b: 16) emphasises that “cohesion can only be satisfactorily understood if it is described functionally and taken as a piece”. The function of lexical cohesion has not been discussed by many researchers. Winter (1979), for example, points out that identifying the common function of the different cohesive forms is much more important than distinguishing them in terms of their taxonomy. He explains that the common function is to repeat. He maintains that the function of repetition is to draw attention to what has been replaced. That is, this repetition provides a clause (or other stretches of text) constant by which the new information is noticed and its importance to the context assessed. Winter (1979: 101) further highlights that “[i]n such repetition, there are obligatory changes or additions to the repeated clause structure which give it new meaning as clause. These have been called replacement”. Winter (1979) emphasises that it is this semantic process which is significant to the function of repetition. Hoey (1991b) comments that Winter’s (1979) use of the term ‘replacement’ does not indicate the concrete replacement of one item by another, but it refers to the supply of new information in a context that has been previously mentioned. Knight (1958: 176, cited in Winter 1974: 10) stresses that “[a]lthough repetition can be used very effectively [...] for

emphasis [...], it is nevertheless an error when used unnecessarily, without attaining any effect except monotony". Winter's idea is useful for the present study not so much in terms of how clauses in a text relate to each other, but rather, as we will see in Chapter 6, in that his idea shows that repetition should have a function in text by working as a trigger for re-entering new information.

However, the function of lexical cohesion is more complex than that. Stubbs (2001) observes that the overlapping of co-occurrence patterns is cohesive. This function of lexical cohesion emphasises the phraseological character of natural language (Chapters 3 and 7 will explain this function of lexical cohesion in more detail). Nevertheless, this function is rarely formally taught to learners of English, as Cheng (2009) points out. From my previous experience in teaching English to EFL learners in an Arabic speaking country, English course books and other teaching materials used in L2 classrooms have produced very little evidence of the relationship between lexical cohesion and overlapping patterns of co-selection. In the final chapter, I will therefore suggest a corpus linguistic approach to teaching cohesion in an L2 classroom. The aim of this approach will highlight such a relationship.

1.4 The importance of comparing NSs and NNSs in learner corpus research

The current study compares a learner corpus of non-native speakers of English to a reference corpus of native speakers of English (details of the type of each corpus are provided in Chapter 4). Hunston (2002) suggests that the essence of learner corpora is primarily that they are for comparison. She claims that these corpora provide information about the difference between heterogeneous groups of learners, and between learners and native or expert speakers. Researchers such as Granger (2002) and Leech (1998) argue that comparing native/

and non-native speakers can highlight different characteristics of non-nativeness in learner writing, which include errors and other cases of under- and over-use of words, phrases and structures. Leech (1998: 20) maintains that a comparison of learner corpora with native speaker corpora can give information about the features of 'interlanguage'. Waelateh (2016) defines interlanguage as a transitional phase between a learner's first language and their second language, in which the learner adopts rules from both languages in order to produce texts in the second language. Some linguists have criticised this type of comparison (e.g., Widdowson 1997) because they consider that interlanguage has to be studied separately and not as an incorrect form of language when compared to the native 'norm'. But Granger (2002) maintains that this comparison can determine the degree of divergence between native and non-native speakers. The comparison with native data is essential since all foreign language teaching tends to improve the learner's proficiency, which means bringing it closer to some native speaker (NS) norms. The most influential work in learner corpora has been conducted by Sylviane Granger. Granger (2002) adopts the following definition of learner corpora, which is based on Sinclair's (1996) definition:

Computer learner corpora are electronic collections of authentic foreign or second language textual data assembled according to explicit design criteria for a particular SLA/FLT purpose (Granger 2002: 7)

Sinclair (1996) distinguishes between data that are collected from natural communication and data collected in experimental or in artificial conditions of different kinds. He considers the first type of data as authentic, while the latter is not. However, Granger (2002) argues that the idea of 'authenticity', that is mentioned in Sinclair's (1996) definition, is questionable in the case of learner language. Granger (2002) explains that many learner corpora entail some 'artificiality' and control. To clarify this point, Granger (2002) gives the example of 'free

compositions' and considers them as natural in that they feature as free writing in which learners are writing freely rather than write something that the researcher is keen to examine. At the same time, however, Granger (2002) adds that these compositions are somewhat elicited (not authentic) because a number of task requirements, such as the topic or the time limit, are usually set for the learner. She then concludes that learner corpora of essay writing can be dealt with as authentic written data because learner data is rarely as fully natural as native speaker data. This observation is of value to the present study because I use a learner corpus of written essays which according to Granger's (2002) argument can be considered authentic to some degree.

According to the definition above, Granger (2002: 8) points out that learner corpora can be categorised into English as a Second Language (ESL), which means English acquired in an English-speaking setting (such as Britain or the US), and English as a Foreign Language (EFL), which is concerned with English learned mainly in a classroom context in a non-English speaking country (e.g., Spain, Libya, etc.). This distinction is important as the interest of the current study is to analyse an ESL corpus of essays written by Arab L2 learners who are studying in the UK. Furthermore, learner corpora present a new type of data which can develop our understanding of both Second Language Acquisition (SLA) research, which attempts to recognise the mechanism of foreign/second language acquisition, and foreign language teaching (FLT) research, which aims to promote the learning and teaching of foreign/second languages (Granger 2002). The current study aims to shed light on both of these areas of research. On the one hand, it endeavours to analyse lexical cohesion in learner writing to understand how learners of English use lexical cohesive devices in their writing. On the other hand, the results of this study also intend to serve a pedagogic purpose by

suggesting multiple strategies that help L2 learners use lexical cohesion in their writing at the functional level.

1.5 Lexical cohesion in text linguistics and corpus linguistics

Lexical cohesion, particularly the study of lexical repetition patterns, is studied both in text linguistics and corpus linguistics. Connor (1996: 19) points out that ‘text linguistics’ is sometimes referred to as ‘text analysis’ or ‘written analysis’. Connor (1996: 80) defines ‘text linguistics’ as “an analysis of texts that extends beyond the sentence level and considers the communicative constraints of the situation”. De Beaugrande and Dressler (1981: 3) maintain that texts need to carry a ‘communicative function’, and this communicative function is defined according to seven principles of textuality, which include ‘cohesion’, ‘coherence’, ‘intentionality’, ‘acceptability’, ‘informativity’, ‘situationality’ and ‘intertextuality’. These principles are important for defining a text as a communicative unit. A number of linguists distinguish between the concepts of ‘text’, viewed as “a physical product, and ‘discourse’, viewed as a dynamic process of expression and interpretation, whose function and mode of operation can be investigated using psycholinguistic and sociolinguistic, as well as linguistic, techniques” (Crystal 2008: 482). Crystal’s (2008) definition shows that ‘discourse’ analyses language from diverse levels. However, the present study will not focus, for instance, on dimensions such as psycholinguistic context. Rather, it analyses lexical cohesion from a linguistic perspective. It suffices to stress at this point that text-linguistics is the most widely used approach to the study of lexical cohesion. This approach looks at language “above the sentence or above the clause” (Stubbs 1983: 1), with a focus on the ways in which texts are structured through language. Nevertheless, the term ‘discourse’ might be used throughout this thesis but the use of the term will not have any theoretical implications.

With a specific focus on the principle of cohesion, particularly lexical cohesion, Adorjan (2013) states that within the framework of the traditional approach, different cohesive forms are categorised in a single text based on properties of semantic relatedness. Then, a theoretical model such as Halliday and Hasan's (1976) model is applied, which is subsequently tried on a small number of texts. The fundamental concern of this traditional approach is to emphasise the inter-clausal nature of lexical cohesion. It also describes cohesion with relation to two general categories: 'grammatical cohesion' and 'lexical cohesion'. The description of cohesion based on these two categories indicates that grammar and lexis are dealt with separately in the language system. Sinclair (2004: 169) claims that in the classic model, "meaning is largely held to reside either in the grammatical choice – on the paradigmatic axis – or in the lexical choice of a word to deliver a meaning". The paradigmatic axis describes the vertical connection between lexical items that belong to the same item class and which can be replaced by other items in the same position within a given sentence. Sinclair (2004) labels this model of language as the 'slot-and-filler' model in which the syntactic structures constitute a set of slots, which are filled with lexical items from the dictionary. Altenberg and Granger (2002) explain that paradigmatic relations are typically described in terms of such relations as synonymy, antonymy, hyponymy, meronymy, etc. These paradigmatic relations are the main linguistic cohesive markers that the classic approach to lexical cohesion is concerned with. If we look at this traditional approach from a methodological perspective, Teich and Fankhauser (2005) mention that the traditional framework of cohesion analysis can only allow for observing a limited selection of texts which means that results gained cannot be generalised. Besides, they maintain that such a method of analysis is manual and hence extremely laborious.

It is important to stress that the aforementioned argument does not mean to devalue or undermine analyses that are either led or supported by the traditional approach to lexical cohesion. In fact, this approach which is based on human analysis is necessary particularly when analysing small size corpus or identifying cohesive relations that have textual features such as signalling nouns. However, analysing lexical cohesion through this traditional approach needs to be supported with another approach in order to eliminate any weakness associated with it. Corpus linguistic work by researchers such as Sinclair (2004) and Hoey (2005) suggests that lexis needs more attention than it has received in traditional approaches to language. As a result, extra consideration is also required for lexical cohesion – or even an approach that fundamentally differs from the traditional text-linguistic one (Flowerdew and Mahlberg 2009). With advancement in corpus linguistic techniques, research into lexical cohesion may take new routes. A corpus theoretical approach to the description of English emphasises the importance of lexis and takes the view that lexis and grammar cannot be divided into distinct groups. Sinclair (2004: 141) suggests a model in which the ‘paradigmatic’ and the ‘syntagmatic’ dimensions of lexical items can be identified by studying the contextual patterning – or co-selection – of words in text corpora. The syntagmatic dimension relates words to the linguistic context, lexically, grammatically and semantically (see Chapter 3, Section 3.2 for more detail on how syntagmatic phenomena are described in corpus linguistics). Thus, the meaning of lexical items must be determined with respect to these two linguistic dimensions in order to describe the forms and the functions of the language satisfactorily. Consequently, cohesion can be conceptualised in a different way from how has been described in the traditional literature: cohesion is established through the interconnection of lexico-grammatical patterns and the overlap of lexical items (see Chapter 3 for further literature on lexical cohesion in corpus linguistics).

However, few studies have used corpus analysis in examining lexical cohesion in text. This may be due to the fact that the focus of corpus-linguistics is on large collections of texts instead of single texts. This method, as Adorjan (2013) points out, does not allow observing individual differences within texts in a corpus, which is a key feature in the analysis of lexical cohesion that entails analysing single texts. However, I argue that a corpus linguistic approach can still be beneficial for studying cohesion. One approach to achieve this aim is to identify the cohesive behaviour of certain lexical items in a general corpus, and then reassess this behaviour within an individual text (see Chapter 7 for this approach). This approach to the analysis of lexical cohesion has been used by researchers such as Stubbs (2001a) and Hoey (2005). This view is supported by Flowerdew and Mahlberg (2009: 2) who posit that “corpora provide huge amounts of real evidence and at the same time make it possible to focus on specific types of texts and on specific patterns of words”.

Thornbury (2010) highlights another limitation of the study of cohesion using corpus linguistics. He claims that the use of corpus tools to analyse individual texts in terms of lexical cohesion, or the type of internal organisation these texts create by means of discourse markers, is challenging. He points out that corpus tools cannot easily detect cohesive ties unless they have been tagged as such. And even so, it is another matter to identify what a device is cohesive with. Thornbury (2010) refers to another difficulty of analysing lexical cohesion which is the problem of corpus annotation. Teich and Fankhauser (2005) argue that there are many tools that support the corpus annotation of grammatical units, which include ‘part-of-speech taggers’. However, the corpus annotation of textual features such as ‘cohesion’ is still problematic. Teich and Fankhauser (2005) explain that one reason for this difficulty is because a fully electronic annotation of cohesion cannot usually be achieved.

McEnery and Wilson (2001: 63) support this view and state that “aspects of language at the levels of text and discourse are probably the least frequently encountered annotations in corpora”. They observe, for instance, that the ‘pronoun reference/anaphor’ – one device of cohesion – has not been marked up sufficiently in corpora with exceptions of few corpora that annotated pronouns such as the Lancaster/IBM anaphoric Treebank. The inadequacy of such anaphorically annotated corpus seems likely due to the complexity of determining to what or whom those pronouns are referring. In addition, McEnery and Wilson (2001) comment that this limited number of annotated corpora is because that pronoun reference annotation (anaphoric annotation) is a type of annotation which can only be performed by human analysts. Thus, it could be inferred that this situation in corpus linguistics can also be applied to lexical cohesion and this also may explain why there are few corpora, if any, which are annotated for lexical cohesive devices.

However, researchers such as Thornbury (2010) and Cheng (2009) suggest that corpus tools, such as frequency/word lists, can provide a starting point to track, count and plot discourse features at the surface level. These features include discourse connectors, and occurrences of lexical repetition. Furthermore, some researchers (e.g., Scott and Tribble 2006) use the keyword method which plays a role in revealing the ‘aboutness’ of a text. Thornbury (2010) emphasises the importance of key words, as a form of repetition, in creating cohesion. These quantitative methods remain, however, surface features, which constitute only one part of what ‘cohesion’ is in the present study. These methods do not explain how cohesion functions in text, which is a key element of cohesion. Therefore, this quantitative analysis of lexical cohesion needs an interpretation in the light of existing models. The current study combines both approaches of text-linguistics (see Chapter 6) and corpus linguistics (see Chapter 7) to

analyse forms and functions of lexical cohesion (further information on how each approach is used is explained in Chapter 4). As Thornbury (2010: 276) points out, integrating different related disciplines offers the most promising way forward.

1.6 Research objectives and research questions

In this study I aim to examine three categories of lexical cohesion in two corpora: ALEC (The Arabic Learner English Corpus) and the British sub-part of LOCNESS (The Louvain Corpus of Native English Essays). These two corpora contain argumentative essays written by ESL Arab non-native speakers of English (NNSs), who are studying in the UK, and English native argumentative essays that represent the baseline for the comparison in this study (NSs). Chapter 4 provides more detail on these two corpora. From these essays in both ALEC and LOCNESS, I intend to identify and quantify the three devices chosen of lexical cohesion to examine which is the most dominant category of cohesion used by both groups. On the basis of this information, I further want to find out how lexical cohesion is used in text at the paradigmatic level. Another more methodological objective of this study is to investigate to what extent a corpus approach can shed new light on the type and functions of lexical cohesion that can be explained through the overlapping of co-occurrence lexical patterns in texts as well as links between the units across stretches of text. These objectives are to answer the following research questions:

RQ1: What are the relative frequencies of the tokens of each lexical cohesive category in each variety (NNS vs. NS)?

- How many instances/tokens of simple repetition can be counted in each corpus (NNS vs. NS)?

- How many instances/tokens of derived repetition can be counted in each corpus (NNS vs. NS)?
- How many instances/tokens of signalling nouns can be counted in each corpus (NNS vs. NS)?

RQ2: What are the advantages and disadvantages of a text linguistic approach in describing the function of the lexical cohesive forms in each corpus (NNS vs. NS)?

RQ3: What are the advantages and disadvantages of a corpus-linguistic approach in adding further detail to the description of the function of the lexical cohesive forms in each corpus (NNS vs. NS)?

The first main research question with its three sub-parts will be answered quantitatively. Simple and derived repetitions will be quantified with the lexical repetition network model (LRNetM) that I will suggest in Chapter 5. This model is mainly based on corpus tools (i.e. word lists) (see Chapter 5 for a comprehensive view of this model). Signalling nouns, the third category of lexical cohesion, are counted through a complete text-oriented approach. The initial results will be built on and will lead to the analysis of qualitative data in the subsequent phase. Firstly, a text-linguistic analysis of a number of extracts from both corpora will be applied to address the second research question that examines the functions of cohesion from the angle of text-linguistics (see Chapter 6). Then, I will use corpus-linguistic descriptive concepts (namely, semantic preference and prosody) to answer the third research question (see Chapter 7). The procedure of analysis (quantitative and qualitative) will be explained in Chapter 4.

1.7 Organisation of the thesis

This thesis is organised into eight chapters. Following this introduction, Chapter 2 will firstly review the key traditional models of lexical cohesion identifying how each model categorises and analyses lexical cohesion. Secondly, an overview of simple repetition, derived repetition and signalling nouns will be given along with the working definitions of each category according to the present study. This will be followed by a survey of the previous studies in NNS writing that analysed lexical cohesion in native and non-native writing. This survey also highlights how lexical cohesion is quantified in learner writing. The chapter then examines a number of issues in the application of the principles of the traditional model of lexical cohesion to text analysis. This will lead to Chapter 3, which looks at lexical cohesion in the light of corpus linguistics. The chapter reviews two types of corpus-based studies of lexical cohesion: Studies that analyse non-learner corpora such as newspapers, and studies that analyse learner corpora. The chapter ends with introducing the theoretical concept of the ‘lexical item’ explaining how it could be used as an analytical tool to analyse lexical cohesion. Subsequently, Chapter 4 considers the type of participants, and provides an account of the design and construction of the ALEC corpus. It also suggests LOCNESS as a reference corpus of native English writing, and discusses the problem of selecting a proper reference corpus in studies that involve learner corpora. The chapter further highlights a number of comparability issues that are related to the structure of the two corpora. It also discusses briefly how combining quantitative and qualitative analyses is a constructive approach to the analysis of lexical cohesion. It then explains the general framework for the data analysis describing the procedures of using text analysis and corpus analysis. After that, the chapter underlines a number of methodological points that have to be considered before starting the

analysis. These points include the use of the T-unit in the analysis of lexical cohesion and the issue of co-reference.

Next, Chapter 5 introduces the Lexical Repetition Network Model which is based on creating lexical networks that are grouped through wordlists. This model will be used to quantify both simple and derived repetition. Based on this quantification, this chapter will also provide a comparison of how the counting method of lexical cohesion in the Lexical Repetition Network Model differs from that applied by the classic models. The chapter then suggests a purely text-analysis approach to analyse signalling nouns. This chapter then introduces the main quantitative findings. Following this, Chapter 6, a discussion chapter, interprets the quantitative findings by examining the three lexical cohesion categories in the light of text analysis. Chapter 7 offers another method of analysing lexical cohesion by applying descriptive concepts of corpus linguistics, namely, semantic preference and prosody. The eighth and final chapter of this thesis concludes the study by summarising its main findings and considering the implications that these findings might have for second language writing research and corpus linguistics. Chapter 8 also highlights the key contribution of this thesis, and discusses areas for future research.

Chapter 2

Lexical cohesion in text-linguistics

2.1 Introduction

As mentioned in Chapter 1, the traditional text-linguistic models of lexical cohesion focus on the paradigmatic relations, which refer to different kinds of semantic associations between lexical items that can be substituted with another item in the same category. These paradigmatic relations are the main cohesive categories that are addressed by the classic models of lexical cohesion. A number of models of cohesion analysis have been introduced. However, the work on cohesion by Halliday and Hasan (1976) as well as Hoey (1991b) are the most influential studies in the field. Restricting a discussion of cohesion to the work of these researchers does not indicate that other publications (e.g., Gutwinski 1976; Martin 1992) that have contributed to the research in cohesion are less important. But this chapter will focus on the models that are central for the purpose of the present study (cf. Sections 2.2, 2.3 and 2.4). In Section 2.5, the chapter will also refer to other categories or strategies that can be employed to create lexical cohesion in text. This survey will then lead to an introduction of the categories of lexical cohesion which will be included in the present study (cf. Section 2.6). Sections 2.6.1, 2.6.2 and 2.6.3 will provide the operationalised definitions of ‘simple repetition’, ‘derived repetition’ and ‘signalling nouns’ respectively. Section 2.7 will discuss in more detail the criteria for determining what constitute signalling nouns in the present study. Afterwards, a review of the previous studies that compare native and non-native writing in terms of the three lexical cohesive categories will be presented in Section 2.8. Most of these studies based their analysis of lexical cohesion on traditional models of lexical cohesion but none of these studies have examined the applicability of these models to analyse large-scale

corpora. Therefore, Section 2.9 will highlight some of the limitations of the classic models of lexical cohesion for the analysis of texts. This will be followed by a conclusion (cf. Section 2.10).

2.2 Halliday & Hasan's (1976) model of lexical cohesion

Halliday and Hasan (1976) have had the widest audience because they provide the most detailed description of a relatively neglected part of the linguistic system: cohesion. Their book *Cohesion in English* describes a clear method for analysing and coding sentences in terms of cohesive devices. They divide cohesion into two broad categories: grammatical and lexical. Grammatical cohesion includes five types: 'reference', 'substitution', 'ellipsis' and 'conjunction', whereas 'lexical cohesion' is created through semantic relationships among lexical items. Halliday and Hasan (1976) maintain that lexical cohesive relations are used in English to create 'texture', which is a principle that differentiates a text from a non-text. They claim that texture is created by the presence of cohesive relations between lexical items in the text that provide its unity, and they call the cohesive relationship between two related lexical items in a text as a 'tie'. They describe that a 'tie' is formed when the relation of cohesion within a text is established where one item in one sentence presupposes another for its interpretation, and the two elements are thereby tied into a text. Halliday and Hasan (1976) add that a tie is a useful concept that can be used to analyse a text based on its cohesive features, and it provides a systematic description of its patterns of texture. Furthermore, they classify cohesive ties according to the distance separating the presupposing from the presupposed items (text-span classes: immediate, mediated, remote, mediated and remote). In their analysis, they allocated numbers to each example of a non-immediate tie. This is to show the number of intervening sentences in order to know the way cohesive relations build up the

text. Herein they also introduce the notion of ‘a cohesive chain’, which is described as what is established when a cohesive item refers back to an item that is itself cohesive with a still earlier element, and so forth. Halliday and Hasan (1976) identify two major sub-classes of lexical cohesion which are ‘reiteration’ and ‘collocation’.

2.2.1 Reiteration

‘Reiteration’, according to Halliday and Hasan (1976: 275-278), includes the repetition of the same word (e.g., *mushroom – mushroom*), the use of a synonym (e.g., *sword – brand*), the use of a superordinate (e.g., *Jaguar – car*), and the use of a general word (e.g., we all kept quiet. That seemed the best *move*). All these lexical forms share the function of repeating the previously mentioned item either identically or in a modified form. Halliday and Hasan (1976: 278) explain that the clearest type of reiteration is where two lexical items share the same referent (i.e. co-referential). They claim that such co-referentiality will often add strength to the cohesive relation between the items. An example of this case is the following (Halliday & Hasan 1976: 281):

- (1) Just then *a fawn* came wandering by: it looked at Alice with its large gentle eyes, but didn’t seem at all frightened.... ‘What do you call yourself?’ *the fawn* said at last.

The category of reiteration and its subclasses have been broadly used in lexical cohesion analyses, with sometimes further modifications. Halliday and Hasan’s (1976) model includes another category which is the category of ‘collocation’.

2.2.2 Collocation

Halliday and Hasan (1976) extended the range of lexical relationships that have a cohesive effect to include collocation. ‘Collocation’, in their study, is created by means of the association of lexical items that regularly co-occur in the same lexical environment or when they are connected through lexico-semantic relationships. For example, *boy* and *girl* establish cohesion through collocation because they have opposite meanings, but examples such as *laugh* and *joke*, and *boat* and *row* are also cohesive despite the fact that they are not systematically related, only “typically associated with one another” (Halliday & Hasan 1976: 284-286). They admit that collocation is the most ambiguous category in their analytical model. Such relations do not depend greatly on systematic semantic relationship. Instead, cohesion by means of collocation is always possible between any pair of lexical items that appear in similar contexts if they occur in adjacent sentences.

Collocation in Halliday & Hasan’s (1976) study is not the same as the concept that is introduced by Firth (1957) who defines collocation as words that co-occur regularly in each other’s company. This definition by Firth implies that collocation spreads over short distances. Tanskanen (2006) adds that this is the sense of collocation according to lexicography which has also been stressed by Sinclair (1996) who continued the study of the lexicographic aspects of collocation, using large computerised corpora. His analyses show the importance of collocational patterns in language use and highlight the fact that there is significant co-selection among words (see, for example, Sinclair 1966, 1991, and 2004). In contrast, collocation according to Halliday and Hasan (1976: 286) spans over larger distances “weaving in and out” of successive sentences building up long cohesive chains with word patterns like “*candle...flame...flicker; hair...comb...curl...wave*” Halliday and Hasan (1976:

286). This type of collocation is called ‘cohesive collocation’ as Tanskanen (2006) describes. Although the present study does not analyse a wide range of semantic relations, the concept of cohesive collocation or word association is still relevant. The present study analyses what Hoey (2005) describes as the collocation of a word with itself which produces cohesion by repetition. That is, lexical items that stretch over a text associate with each other through lexical repetition forming a network of lexical cohesion (e.g., *computer...computers...computerise...computation*) (see Chapter 5 for the type of lexical repetition networks included in the present study).

Generally, Halliday and Hasan’s (1976) book discussed lexical cohesion in less than twenty pages compared with over fifty, for example, for the grammatical category of ‘substitution’. However, the list of cohesive ties that are lexical in Halliday and Hasan’s (1976) analysis of sample texts in their final chapter appear rather detailed. Hoey (1991b) counted these examples of lexical cohesion that were analysed by Halliday and Hasan (1976), and observed that these examples account for over 40% of the total lexical ties. Tanskanen (2006) points out that Halliday and Hasan’s (1976) book devoted insufficient space for lexical cohesion, which is important to highlight the importance of this linguistic feature in discourse. However, the value of Halliday and Hasan’s work lies not only in the model of analysis they propose, but also in the fact that their discussion of lexical cohesion pays attention to the interplay between cohesive relations and coherence of a text. This interplay is essential to the functioning of cohesion. Halliday and Hasan (1976: 23) explain that one can produce texts which are cohesive but which are not treated as texts because they do not have continuity of meaning with regard to coherence. This fact is relevant to the present study, as I have been

arguing in Section 1.2 that cohesive forms are only useful if they predispose the receiver to successfully interpret the message.

2.3 Hasan's (1984) model of lexical cohesion

Hasan (1984) revised lexical cohesive relations in her study on coherence and cohesive harmony. In her paper, she redefines the lexical cohesive categories, and the obvious change that she has made concerns the category of collocation. She admits that “unless we can unpack the details of the relations involved in collocation [...], it is best to avoid the category in research” due to its inter-subjective nature (Hasan 1984: 195). She breaks down the relations that have earlier been studied under collocation into separate categories in her revised version of lexical cohesion. Hasan's new model consists of two main categories: ‘General’ and ‘Instantial’. The general category, for example, includes ‘repetition’ and other relations that can be explained by the general semantic system of English: ‘synonymy’, ‘hyponymy’, ‘meronymy’ (part-whole relation) and ‘antonymy’. Some collocation relations are thus now considered under the general category. For example, Hasan notes that it is no longer necessary to think of *go* and *come* as related to each other through collocation; they fall within the same chain on the ground of being ‘antonyms’.

What is particularly important about Hasan's (1984) study is that it recognises the chain-forming property of lexical cohesion instead of concentrating on individual ties as in earlier studies in lexical cohesion. Hasan (Halliday & Hasan 1985) divides cohesive chains into ‘identity chains’, which are made of various co-referential items across a text, and ‘similarity chains’, which are created by means of co-classification (substitution and ellipsis) or coextension (repetition, synonymy, antonymy, hyponymy, and meronymy). Hasan (Halliday

& Hasan 1985: 91) introduces the concept of ‘chain interaction’, which means that cohesive ties that enter into chains should be considered in combination with other ties. She considers that a chain interaction requires at least two candidates of one chain to be connected cohesively with at least two candidates of another chain. This interaction will therefore result in a network of chain relationships. She calls this interaction of cohesive ties as ‘cohesive harmony’, which is a key aspect for a text’s coherence. Hasan’s (1984, 1985) idea is insightful for the current study, as demonstrated in Chapter 1, Section 1.2.

2.4 Hoey’s (1991b) model of lexical cohesion

Hoey (1991b) gives more attention to lexical cohesion than Halliday and Hasan (1976). In his book *Patterns of Lexis in Text*, he provides a comprehensive analytical model of lexical cohesion. He considers that “[l]exical cohesion is the only type of cohesion that regularly forms multiple relationships” (Hoey 1991b: 10) between items in the text. He also remarks that cohesive studies are, somewhat, studies of lexical relations, particularly those that allow repetition. His model focuses not only on counting cohesive categories, but also on observing how such cohesive devices connect to organise text. As demonstrated in Chapter 1, Hoey (1991b) considers Winter’s (1979) study on ‘repetition-replacement relations’ to be helpful in explaining the function of lexical cohesion, namely repetition. Winter emphasises that the function of repetition is to give a framework for interpreting what is changed (see Example 2).

- (2) Pressures built up on all sides: *his father*, a ‘moderately successful plumbing contractor’ (said Time) *demanded performance*. *His mother*, who left her husband in Florida and moved to Austin to be near her son, *demanded love*.

(Winter 1979: 220, cited in Hoey 1991b: 18)

In Example (2), the italicised items represent ‘simple repetition’. This repetition, as Hoey (1991b) highlights, creates an environment where the focus can be given to what is ‘replaced’ (which is indicated by the underlining). In Example (2), the repetition of the possessive pronoun *his* attracts the eye to a contrast relation between *father* and *mother* (*his father – his mother*). What is more, the verb *demanded* is repeated and its repetition creates the condition for replacement: *love* replaces *performance*, which they are also a type of repetition by contrast. Chapter 6 refers to this function of repetition with examples from ALEC and LOCNESS.

In line with Winter’s (1979) study, Hoey’s (1991b) model is concerned with the use of different types of repetition as clues that can help signal those sentences in a text that are most central to the meaning of the text and that also contribute to the organisation of the text. In this context, Hoey differentiates two types of sentences: ‘central’ and ‘marginal’. Central sentences need to share multiple repetition links (at least three links) with other sentences, and develop the topic of the text. In contrast, marginal sentences have fewer than three connections with other sentences and do not offer valuable information in the text. In order to create coherent summaries from a long text, Hoey (1991b) removed marginal sentences and gathered central sentences, or selected topic opening and topic closing sentences. Hoey’s (1991b) model of lexical cohesion is also inspired by Hasan’s (1984) chaining approach but he claims that Hasan’s (1984) approach did not consider the relationship of cohesion and how sentences connect as whole units in the text. Thus, Hoey (1991a: 389) suggests that interactions of chains are assumed to interrelate and make a nest. In other words, a chain of items should not be thought of linear as in Figure 1 but as a web as in Figure 2.

a
↑
a
↑
a
↑
a

Figure 1 Linear representation of a chain
(cf. Hoey 1991a: 389)

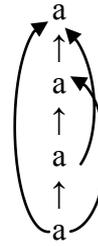


Figure 2 Web representation of a chain
(cf. Hoey 1991a: 389)

Based on Hoey's (1991a) view of a chain, he creates a manual matrix to explore the ways by which the different categories of repetitions link sentences and form bonds. His repetition matrix reflects a kind of non-linear complexity of connected sentences through different types of repetitions. According to Hoey (1991b), links refer to both lexical and non-lexical cohesive relations, whereas 'bonds' are created when sentences share three or more links in a text. He describes the interconnection of these bonded sentences as 'nets' (see Example 3). Links and bonds are key analytical concepts in Hoey's (1991b) model. Bonds are useful because they can identify adjacent and non-adjacent related sentences in texts, and the nets that connect these sentences can reflect the organisation of the text. The following sentence pair, which is taken from Hoey's (1991b: 37) data, constitutes a bond because this pair shares three repetition links (*drug – drugging; humans – humans; bears – them, animals*).

- (3) A *drug* known to produce violent reactions in *humans* has been used for sedating grizzly bears.
- To avoid potentially dangerous clashes between *them* and *humans*, scientists are trying to rehabilitate the *animals* by *drugging* them and releasing them in uninhabited areas.

Example (3) shows examples of ‘simple repetition’ (e.g., *humans – humans*) and ‘complex repetition’ (e.g., *drug – drugging*). Hoey (1991b) explains that ‘simple repetition’ is the repetition of the same lexical item with an inflectional change only (e.g., the singular and plural distinction). In contrast, ‘complex repetition’ involves repeating the same root of the lexical item with a derivational change. These two types of repetition are key categories in the present study and will be discussed in detail in Sections 2.6.1 and 2.6.2. Hoey’s (1991b) taxonomy of lexical cohesion was revised in Károly’s (2002) study in which she applies Hoey’s (1991b) model to analyse English learner writing by Hungarian students. Károly (2002) agrees with Hoey (1991b) that the greater part of cohesion is produced by lexis, which has an outstanding function in creating texture in English written discourse. She also follows Hoey (1991b) in differentiating the two types of repetition ‘simple’ and ‘complex’. She, however, classifies both ‘simple’ and ‘complex repetition’ under a broad term which she calls ‘same unit repetition’ and also re-named ‘complex repetition’ as ‘derived repetition’. However, this is only a difference in terminology. The definitions of simple and complex/derived repetition are the same by both researchers. Károly (2002) prefers to use simple terminologies and criticises Hoey for including obscure category labels. For example, she uses the term ‘synonym’ instead of ‘paraphrase’ to include all those lexical units that are interchangeable in the particular environment where they occur. On the other hand, ‘paraphrase’ in Hoey’s (1991b) model is more complicated. Hoey (1991b) divides it into simple and complex. Hoey (1991b: 62) suggests that ‘simple paraphrase’ occurs “whenever a lexical item may substitute another in context without loss or gain in specificity and with no discernible change in meaning”. On the other hand, ‘complex paraphrase’ occurs when “two lexical items are definable such that one of the items includes the other, although they share no lexical morpheme” (Hoey 1991b: 64). He classifies complex paraphrase into three

subcategories: ‘antonymy’, ‘a link triangle’, and the ‘mediator’ missing. Other types of complex paraphrase include: ‘superordinates’ and ‘coreference’. Károly (2002) recommends that it would be safer and more practical to retain the original category names adopted by previous researchers such as Halliday & Hasan (1976), and Hasan (1984) than using new terms. I outlined Hoey’s (1991b) categories of lexical repetition in Table 1.

Category of lexical cohesion	Examples
a. Simple lexical repetition	<i>a bear – bears</i>
b. Complex lexical repetition	<i>a drug (n) – drugging (v)</i>
c. Simple paraphrase	<i>to sedate – to drug</i>
d. Complex paraphrase	<i>heat – cold</i>
e. Substitution	<i>a drug – it</i>
f. Co-reference	<i>Mrs Thatcher – the Prime Minister</i>
g. Ellipsis	<i>a work of art – the work</i>
h. Deixis	<i>The works of Plato and Aristotle – these writers</i>

Table 1 Hoey’s (1991b) categories of lexical repetition

2.5 Other categories of lexical cohesion

Lexical cohesion is achieved through a number of cohesive devices other than the categories that are introduced by the aforementioned models of lexical cohesion. For example, Flowerdew (2002) suggests the category of ‘signalling nouns’ (henceforth SNs). Flowerdew

and Forest (2015) define an SN as an abstract noun that must require lexical realisation to provide its specifics for the current discourse (e.g., *problem*). They maintain that this is the feature that makes this type of nouns cohesive. Hinkel (2001: 129) explains that these nouns are effective strategies for creating cohesion in texts because they “have specific identifiable referents in text, to which these nouns are connected”. SNs are one of the lexical cohesive categories that the present study analyses (Section 2.6.3 provides the working definition of SNs and their criteria of analysis). Apart from SNs, previous studies on cohesion have analysed other strategies that can be used to connect the text cohesively. In Table 2, I list a number of these strategies suggested by different researchers.

Strategies of lexical cohesion	The author
a. Parallelisms and adjacency pairs	Morley (1999)
b. The flow of information	Biber et al. (1999)
c. The choice of tense and aspect	Quirk et al. (1985)
d. The use of punctuation	Gutwinski (1976); Baker (2011); Evtushenko and Butuzova (2014)

Table 2 Different strategies of creating lexical cohesion in a text

2.6 Categories of lexical cohesion in the present study: an overview

After reviewing the classic models of lexical cohesion and their different taxonomies, I need to determine categories of lexical cohesion that will be analysed in the present study. Strauss and Fiez (2014) notice that although all cohesive devices serve to create textual and discursive cohesion, some of these categories cross-cut each other in multiple ways. These researchers maintain that it is difficult to label and (at times impossible) to discretely label one particular

type of cohesive resource as separate and distinct from another type. This view is also shared by Morris and Hirst (2006), who observe that it is not often an easy task to assign a specific label to lexical cohesive devices. This is because the semantic meaning of these forms is interlinked. Sense relations overlap and there is an obvious fuzziness of the boundaries of the categories. Károly (2002), for example, comments that it is sometimes impossible to decide if a word is a synonym or a hyponym. Another example of such a difficulty in identifying the right type of lexical cohesion is the category of 'paraphrase' that is suggested by Hoey (1991b). This category is broad and opens the possibility for including different lexical items into the lexical cohesive relationship. In this regard, de Beaugrande and Dressler (1981) point out that paraphrasing could be of a single concept, or of a more complex configuration. The latter type of paraphrasing is further discussed by Teubert (2001: 133), who argues that paraphrasing is part of the negotiation of meaning in discourse. Mahlberg (2005: 163) adds that all language use is paraphrase in that it depends on previous occurrences of language and patterns of words. Thus, this uncertainty of what might be included under a specific category could confuse the analysis of lexical cohesion.

Consequently, for purely practical and analytical reasons, this study will choose categories that are somehow quantifiable and not subject to intuition. From Hoey (1991b), I will take the category of 'simple repetition'. I will further adopt his category of 'complex repetition'. However, I will prefer to use the label 'derived repetition' that is suggested by Károly (2002) instead of the term 'complex repetition'. This is because the term complex repetition is a broad label and can encompass a wide range of lexical cohesive relations other than derived repetition. For example, 'paraphrase' may also be considered a type of complex repetition as I explained earlier in this section. Therefore, I suggest that the term 'derived repetition'

characterises the features of this category better than the term ‘complex repetition’. The third category will be ‘signalling nouns’ that is taken from Flowerdew (2002). Table 3 illustrates this classification in the present study compared to Hoey (1991b) and Károly (2002).

Hoey (1991b)	Károly (2002)	The present study	Examples
1.Simple repetition	1.Same unit repetition a. Simple repetition	1. Lexical repetition a. Simple repetition	<i>brain – brain</i> <i>go – went</i> <i>small – smaller</i>
2.Complex repetition	b. Derived repetition	b. Derived repetition	<i>calculations – calculate</i> <i>human (adj) – human (n)</i>
		2. Textual repetition Signalling nouns	<i>argument,</i> <i>problem,...etc</i>

Table 3 Categories of lexical cohesion in the present study

Table 3 shows that the present study classifies both simple and derived repetition as lexical repetitions whereas it identifies signalling nouns as textual repetition. That is simple and derived repetition, in the lexical cohesive relationship; usually link an individual lexical item or a phrase to another individual item, whereas SNs have the ability to label stretches of discourse, rather than one element. In his discussion of the category of general superordinates, McCarthy (1991) describes that these nouns (including SNs) enclose many elements of the text in a single, more general label. Table 3 also illustrates that the term ‘simple repetition’ is present in each classification by each researcher, while ‘complex repetition’ is named differently. The three researchers agree that simple repetition needs be distinguished from

derived repetition because ‘simple repetition’ represents ‘lexical recurrence’, which means same lexical items are plainly repeated and because this category includes inflectional morphemes which do not change the class of words they are added to. On the other hand, ‘derived repetition’ can add lexical variation to the text because it involves new, though morphologically related items, which makes the text lexically rich.

I acknowledge that any text can have various lexical cohesive devices not only the three categories that I selected. I also recognise that the exclusion of some categories from the analysis prevents to provide a complete picture of cohesion in the text. However, what is significant in the current study is that its primary aim is not to classify lexical cohesion into multiple categories but to focus on how lexical cohesion functions in Arab NNS writing. By putting an emphasis on specific categories, I can, therefore, study their frequency and function in detail rather than analysing texts based on a whole set of lexical cohesive categories and making the analysis too general. Many researchers followed the same approach and selected one category of lexical cohesion to make a detailed study on it. For example, Flowerdew (2009) studied signalling nouns in students’ writing, while Mahlberg (2005) devoted a complete book to general nouns. Additionally, the fact that the present study compares NSs and NNSs justifies my selection of three categories of lexical cohesion due to the time and effort required during the analysis of two data sets.

2.6.1 A working definition of simple repetition in this study

Hoey (1991b) points out that simple repetition is the plainest form of repetition, which most people expect when they deal with repetition. Simple repetition is further named as ‘direct’, ‘exact’, ‘recurrence’ or ‘formal’ by different researchers (e.g., Gutwinski 1976; de

Beaugrande and Dressler 1981). The difference is only in the terminology because the definition of this category is similar in all these approaches. In the present study, Hoey's (1991b) definition of 'simple repetition' is used.

Simple lexical repetition occurs when a lexical item that has already occurred in a text is repeated with no greater alteration than is entirely explicable in terms of a closed grammatical paradigm (e.g., *bear – bears*).

(Hoey 1991b: 53)

The closed grammatical paradigm refers to the possible inflectional changes in the lexical item. These changes produce syntactically motivated variants of the same lexeme, as Hoey (1991b) clarifies. This means that the lexical item keeps the same grammatical function when it is inflected. Inflectional changes in English are: singular and plural; verb form differences '3rd -singular present agreement, past tense, past participle, and *-ing* form'; the possessive; and comparative or superlative morphemes. For the possessive, only possessive nouns (e.g., *computer's keyboard*) are included in the present study. With respect to the *-ing* form, Hoey (1991b) did not explain clearly in presenting his method of analysis whether we count this form as a noun gerund or a verb when it is attached to a word. However, in his repetition matrix, he considered lexical items that function as a noun gerund to be a verb. Károly (2002) also treated noun gerunds as verbs but she made this clear in her method of analysis and illustrated the different uses of the inflectional morpheme *-ing* with this example.

- (4) a. Will you *interview* the president? Yes, I'm *interviewing* great people today.
b. She must *interview* the president. *Interviewing* is her source of inspiration.

In Example (4-a), *interviewing* functions as a continuous form whereas in Example (4-b) *interviewing* functions as a noun gerund and establishes simple repetition with *interview*, not a

derived repetition. Károly (2002) highlights that the ability of forming a gerund is characteristic of every English verb despite the word class change that occurs. She points out that this is a result of a mechanical grammatical rule which complies with Spencer's (1991: 193-194) observation that gerunds are considered to be part of the inflectional paradigm of the verb, not as a class of derivational morphology. Therefore, I will adopt this view and consider the noun gerund as a verb in the present study.

2.6.2 A working definition of derived repetition in this study

Hoey (1991b) supports Stotsky's (1983) view that complex (derived) lexical repetition is not only an extreme version of simple repetition. Instead, it is quite distinct from it in that the collocation that a particular word forms with other words may be quite different from those that a closely related word may form (Renouf 1986). For example, *economist* and *economy* collocate with quite distinct sets of words, whereas items in true simple repetition have similar collocational profiles. Hoey (1991b: 56) points out "we should be wary, however, of assuming that complex repetition is just simple repetition with knobs on". Furthermore, Gutwinski (1976: 81) also acknowledges the importance of derived repetition by claiming that a lexical item, which is created from the same root, can have cohesive characteristics that resemble those of a synonym. He illustrates this similarity by the lexical item *marriage* which coheres with *marry* just as *matrimony* coheres with *marriage*. De Saussure (1915) justifies the grouping of such words as *marriage* and *marry* on the basis of what he calls 'associative relations' obtaining between such items. He cites examples *enseignement* (*teaching*), *enseigner* (*teach*), and *enseignons* (*we teach*). Likewise, de Beaugrande and Dressler (1981) include complex repetition in their taxonomy of lexical cohesion, but they call it 'partial recurrence', which means the use of the same lexical item but with a change in its part of

speech. Similarly, Károly (2002) also includes ‘complex repetition’ in her study but labels it ‘derived repetition’, as alluded to in Section 2.6.

Moreover, complex repetition is also present in Halliday and Hasan’s (1976) model but they do not consider it as a separate category from simple repetition. They classify lexical items with a derivational change (e.g., *noun*, *nominal*, *nominalise*, and *nominalisation*) as a repetition of the same item. A lexical item, according to Halliday and Hasan (1976), is not constrained by “a particular morphological form” (Halliday and Hasan 1976: 291). Stotsky (1983) asserts that Halliday and Hasan’s (1976) model neglects the cohesive relationships that can be created by derivatives, because it regards them as simply a repetition of the same item. This, as Stotsky (1983) points out, will not describe the semantic relationships between lexical items accurately, and important information about the textural patterns in a particular text might not be obtained. Stotsky (1981: 20) underlines the role of derivatives in allowing writers to create a nominalised style, which increases the density of ideas required for concise writing. The repetition of derivational items provides the writer with an additional stylistic advantage because it gives already connected lexical items a further source of cohesion. In this study, I support Stotsky’s (1981, 1983) view that the use of derived repetition should be considered as distinct from simple repetition in creating cohesive relationships. In this study, Hoey’s (1991b) definition of complex repetition (named as derived repetition in the present study) is adopted.

Closely related to simple lexical repetition is complex lexical repetition. This occurs either when two lexical items share a lexical morpheme, but are not formally identical (e.g., *drug* (n) – *drugging* (v)), or when they are formally identical, but have different grammatical functions (e.g., *human* – *humans*).

(Hoey 1991b: 55)

Hoey's (1991b) definition of complex repetition indicates that this category involves a derivational change in the lexical item by adding affixes, and this derivational change usually entails a change in the word class. This rule is illustrated by *drug*, which is a noun, and *drugging* where the suffix *-ing* is attached to *drug* and change it into a noun gerund that functions as a verb. However, Hoey (1991b) mentions that some lexical items could still be classified as a type of complex repetition while they do not undergo any derivational change (e.g., *human* (adj) – *humans* (n)) – see Example 5 below. He contends that such words constitute a marginal case of complex repetition because it could be classified as 'simple repetition' or 'complex repetition'. He adds that this classification is based on the syntactic position the word has in the sentence. In Example (5), the word *human* in sentence (b), is in complex repetition with *humans* in sentence (a) if *human* assumed to be an adjective. If, however, it is regarded as a noun modifier, then it is in simple repetition with *humans* because this is not a different grammatical function. The following example is by Hoey (1991b: 37):

- (5) a. A drug known to produce violent reactions in *humans* has been used for sedating grizzly bears [...]
- b. Many wild bears have become 'garbage junkies', feeding from dumps around *human* developments.

Consequently, identifying such a case of 'zero derivation' – or 'conversion' (cf. Twardzisz 1997), becomes completely biased from an empirical analysis point of view, and could confuse the counting process for analysts particularly in comparative studies where frequency is used as a distinguishing factor between NSs and NNSs. Hoey (1991b), however, gives this case of repetition a secondary importance because his aim is not to count and compare frequency of lexical repetition whereas in the current study such an aim is important. Károly

(2002) stresses the importance of including words that have the same word-forms with different parts of speech in the analysis of derived repetition (e.g., *work* v – *work* n). Biber (2006: 243) observes that different part of speech realisations of the same word form can have quite different meanings. Therefore, it is sensible to consider the same word-forms of a different part of speech (i.e. zero derivation) as different word-forms. In addition, in learner corpora, the use of zero derivation might indicate an improvement in students' writing. Schmitt and Zimmerman (2002) emphasise that the ability to use the appropriate form of a word in a given grammatical context is necessary for developing grammaticality suitable language. Therefore, this case of derived repetition will be considered in the present study during the counting process.

As discussed just earlier, the addition of a derivational affix induces a change in the word class of the lexical item. However, Spencer (1991) notes that such an addition does not necessarily change the word class. Károly (2002) suggests that when word formation produces a change in meaning while the two lexical items remain members of the same word class, they will fall also under derived repetition because they undergo a morphological change (e.g., *green* (adj) – *greenish* (adj); *brother* (n) – *brotherhood* (n)). Accordingly, these cases will be included under derived repetition in this study. Furthermore, following Hoey (1991b), morphologically related antonyms will be included (e.g., *happy* – *unhappy*). I will further include transparent derivatives (e.g., *surely*, *ensure*) and less transparent derivatives (e.g., *surety*, *assure*) if they occur in any of the corpora that will be used in this study.

2.6.3 Signalling nouns and their working definition in this study

As mentioned in Section 2.6, the term 'signalling nouns' suggested by Flowerdew (2002) is used in the present study. SNs are similar to general nouns that are proposed by Halliday and

Hasan (1976). However, Flowerdew (2009) mentions that general nouns can be considered as a superordinate term that incorporates SNs, which means that general nouns are more general than SNs. SNs have a number of features that have been described in the previous research as ‘type 3 vocabulary’ (Winter 1977), ‘anaphoric nouns’ (Francis 1986), ‘advance labels’ (Tadros 1985), ‘carrier nouns’ (Ivanič 1991), ‘metalanguage nouns’ (Winter 1992), ‘enumerative nouns’ (Tadros 1994) and ‘shell nouns’ (Hunston & Francis 2000; Schmid 2000). All these labels describe the same category, which Flowerdew (2002) names as SNs. In spite of the different opinions about the terminology, these nouns share a number of features in that they all represent a sub-category of abstract nouns, and have a signalling function (Flowerdew 2003). These nouns tend to refer to or encapsulate or package (Sinclair 1993: 7; Francis 1994) stretches of discourse. Flowerdew’s (2009) definition of SNs will be adopted in this study. He defines SNs as,

Signalling nouns are nouns which have cohesive properties across [...] clauses. A signalling noun is potentially any abstract noun the full meaning of which can only be made specific by reference to its context. Examples of nouns which can function as signalling nouns are *attitude, difficulty, process, reason, result* etc.

(Flowerdew 2009: 85)

Flowerdew and Forest (2015) note that SNs can operate across clauses – either cataphorically (referring forward) or anaphorically (referring backward – or within the clause). The current study focuses on SNs that operate only across T-units cataphorically and anaphorically. The T-unit will be used in the present study. It is defined by Hunt (1965) as the shortest unit of language that could be considered a grammatical complete sentence (Chapter 4 provides further detail on the T-unit). Examples 6 and 7 illustrate the across clause function with the noun *problems* signalling cataphorically, and the noun *fact* signalling anaphorically. In all

following examples on SNs, SNs are italicised, while their realisations are underlined. Also, two slashes // are used to mark the end of each T-unit in the examples from ALEC and LOCNESS.

- (6) T Cartels encounter two characteristic *problems*. The first is ensuring that members follow the behaviour that will maximize the industry's joint profits. The second is preventing these profits from being eroded by the entry of new firms.

(Flowerdew & Forest 2015: 1)

- (7) Electricity is used to drive the motor of an electric train, but inevitably some of the energy is lost as heat. This unavoidable *fact* is of great importance in biology.

(Flowerdew 2003: 330)

Flowerdew's (2009) definition of SNs is too general and it will not completely help to identify SNs in the essays by Arab speakers of English and native speakers. Therefore, I need to set up more specific criteria by which I can capture instances of SNs in ALEC and LOCNESS, whenever possible. I will allocate more space to SNs in the following sections than the one that is given to simple and derived repetition. This is because SNs cross-cut with simple and derived repetition and hence it requires extra attention. Nevertheless, this argument does not indicate that my analysis is devoted to analyse SNs, but SNs represents one part of a larger argument about other lexical cohesive categories (i.e. simple and derived repetitions).

2.7 Criteria for determining SNs in the present study

Flowerdew and Forest (2015: 46) admit that "there is no single test or method which neatly identifies all and only SNs and which can be applied without reliance on expert judgment of borderline cases". Furthermore, SNs are not strictly a lexical or grammatical category but also

have semantic and discursal characteristics (Flowerdew and Forest 2015: 7), which makes their analysis not precise. Schmid (2000) adds that it is not easy to determine an all-inclusive list of shell nouns (SNs in the present study) because identifying these nouns is only possible inside their text. Francis (1994) in a discussion of ‘retrospective labels’ articulates a general criterion of how SNs (anaphoric nouns in Francis’s terms) might be determined:

A retrospective label serves to encapsulate or package a stretch of discourse. My major criterion for identifying an anaphorically cohesive nominal group as a retrospective label is that there is no single nominal group to which it refers: it is not a repetition or a ‘synonym’ of any preceding element. Instead, it is presented as equivalent to the clause or clauses it replaces, while naming them for the first time.

(Francis 1994: 85)

Francis’s (1994) criterion raises a key issue in linguistics where semantic categories overlap in discourse and could carry out more than one linguistic function at a time. A lexical item could, for instance, act as a synonym or has a function of signalling or signposting simultaneously. Therefore, it is vital to determine clear criteria to distinguish SNs from other lexical cohesive categories, particularly because this study analyses other categories such as lexical repetition. Francis’s (1994) criterion will be taken into account in the present study but to a considerable degree as we will see in the following sections. Besides, her criterion only deals with the retrospective function of SNs while the present study analyses both functions of retrospection (i.e. anaphoric) and prospection (i.e. cataphoric). Accordingly, the criteria of coding SNs in the present study have to consider both categories.

A further principle of identifying any noun as an SN in this study is that an SN has to provide a new characterisation of the discourse segment that it encapsulates or prospects. That is, an SN needs to categorise its textual referent differently. Characterisation is one of the three

functions that Schmid (2000: 14) identifies to determine shell nouns (SNs in the present study). These functions are: ‘temporary concept-formation’, ‘linking’, and ‘characterisation’. The first function involves that a shell noun encapsulates the discourse segment into single concepts that are related to the temporary situation of speech. This concept is of a temporary nature because its meaning varies according to the surrounding discourse. The linking function stresses that SNs create textual links in the text in that they help to show cohesive links with preceding or subsequent stretches of discourse. The third function which I consider as the most revealing one in setting up my criteria of analysis is characterisation. This function requires an SN to propose any different conceptual shell producing some fresh insight into the writer’s understanding of the segment. Schmid (2000) illustrates these functions as shown in Example (8); the two noun phrases that contain shell noun phrases are in italics whereas their lexical realisation is underlined.

- (8) *The Government’s aim* is to make GP’s more financially accountable, in charge of their own budgets, as well as to extend the choice of the patient. Under *this new scheme*, family doctors are required to produce annual reports for their patients.

(Schmid 2000: 7 – BBC material, COBUILD corpus)

In Example (8), the noun phrase *the Government’s aim* encloses the underlined clause in a single temporary nominal concept *aim*. It also links *aim* with the clause which contains the actual details of information. In addition, this shell noun phrase characterises the chunk of information as an *aim*, i.e. as something the British government wants to achieve. Schmid (2000) points out that the characterisation function depends on the speaker’s choice of a particular shell noun and modifier that he or she wants to describe in order to characterise their own ideas. In Example (8), Schmid (2000) explains that the speaker could have

characterised the information by introducing other shell nouns such as *endeavour*, *effort* or *need* instead of *aim*. In the second sentence in Example (8), Schmid (2000) notes that the speaker reactivates the same information given previously but characterises it with a different conceptual shell through the use of the noun phrase *this new scheme*. By using *scheme* with the pre-modifier *new* instead of just repeating the noun *aim*, the speaker adds fresh information that the *scheme* is *new* and also uses *this* to link the second shell noun to the information expressed in the previous sentence. We could conclude from this discussion that the characterisation function is an important criterion to determine SNs. Schmid (2000) underlines that while most shell nouns (SNs in the present study) succeed in fulfilling temporary concept-formation and linking functions, they fail to satisfy characterisation. In the present study, one of the reasons that might prevent SNs from meeting the characterisation function successfully is ‘repetition’, as we will see in the following section.

2.7.1 Signalling nouns and repetition

Flowerdew and Forest (2015: 53) observe that the relation between an SN and its lexical realisation is complicated by the phenomenon of repetition. More specifically, they explain that an SN may be repeated in the environment of its realisation and once more throughout the text. The form of an SN could be simply repeated or reiterated through other means of repetition such as paraphrases and synonyms. Flowerdew and Forest (2015) establish a principle when such types of repetition take place in the text. They posit that when the same SN is repeated in adjacent pairs of sentences – one of the SNs functions as an across-clause SN and the other one as an in-clause SN – a priority will be given to count the in-clause SN (see Example 9). Flowerdew and Forest (2015) refer to the first function of the SN as a discourse relation whereas the latter one as a syntactic relation. They prioritise the syntactic

over the discursive features in cases of double realisation because they observe that the across-clause SNs outnumber (60%) the in-clause relations (40%) even without counting these cases of double realisation. They explain that both functions of SNs could have been coded but excluding one of them is just a methodological decision. Flowerdew and Forest's claim implies that their direction of the analysis is based on a personal preference and on frequency results. Example (9) illustrates the overlap between SNs and repetition where a single stretch of text acts as the specifics for multiple SNs (Flowerdew & Forest 2015: 53):

- (9) Let's go back over them a little more slowly now. Let's look at this second *objection*. Er, in a nutshell, this *objection* is that the original position is set up in a way that's biased against ever er yielding procedural principles of justice and historical entitlement-type principles. OK, we can call that the bias *objection*.

In Example (9), the lexical item *objection* is repeated three times throughout three successive sentences. The first mention of *objection*, which functions as a cataphoric across-clause SN, is repeated in the context of its realisation that is expressed in the subsequent sentence. The second mention of *objection* functions as in-clause SN. In line with Flowerdew and Forest's principle articulated above, only the in-clause SN is counted in sentence two: this *objection* is that the original position is set up in a way that's biased against ever er yielding procedural principles of justice and historical entitlement-type principles. There is also a third mention of *objection* in the last sentence: *OK, we can call that the bias objection*. This SN has an anaphoric across-clause function pointing backward to the same lexical realisation indicated by the underlining above. Again, Flowerdew and Forest (2015) do not count it here. They control the number of repeated SNs not because they deny the role of continuity that

repetition plays in discourse but only for them to create a confined method of counting and not to exaggerate the significance of SNs in their findings.

Flowerdew and Forest's (2015) technique of dealing with the issue of SNs and repetition is less strict than Francis's (1994) who accepts only SNs that are mentioned for the first time as stated in Section 2.7. Flowerdew and Forest's (2015) principle does not completely apply to the present study because in-clause SNs are not included. Instead, the present study analyses SNs that function cohesively across T-units. Thus, in Example (9), I would have counted the across-clause SN which is the first mention of the lexical item *objection* rather than the one with the in-clause SN function (Section 2.7.1.1 illustrates this principle). Flowerdew and Forest's (2015) principle is still useful though because it puts restrictions to count the frequency of SNs when they are repeated in the same text. As demonstrated in this section, repetition could take a number of forms which overlap with SNs. These forms of repetition include 'simple repetition', 'nominalisation' and 'synonyms'. In the forthcoming sections, I will therefore suggest principles of analysis for counting SNs when they are repeated either in their contextual environment or throughout the text.

2.7.1.1 Signalling nouns and simple repetition

1. (a) When the same SN is repeated in successive pairs of T-units – the first of which functions as an across-t-unit SN and the other one as an in-t-unit SN – only the SN with the function of across-t-unit will be counted. Example (10) illustrates this principle of analysis. The example is taken from the ALEC corpus, which I will use in the present study. A detailed description of this corpus is presented in Chapter 4. The T-units in Example (10) and in all following examples are numbered and separated by two slashes.

- (10) **1** Immigration, undoubtedly, is an essential *requirement* to the continuity of many countries. // **2** This *requirement* is not only to replenish the decrease in the number of populations in those countries, // **3** but immigration offers those populations the treatment of the major deficiencies in certain sectors. (ALEC)

In Example (10), the first mention of *requirement* in T-unit (1) is a cataphoric SN which refers forward to what a requirement is to be. T-units (2) and (3) contain the lexical realisation of *requirement*. *Requirement* is then repeated in T-unit (2) and functions as an in-t-unit SN. However, based on the principle of analysis suggested in (1-a), I will only count the initial mention of *requirement*, giving that my aim is only to analyse across-t-unit cohesion.

(b) In contrast, when the initial occurrence of an SN acts as an in-t-unit label, it will be discarded and priority of counting will be given to the across-t-unit function which represents the other mention of the SN (cf. Example 11). This rule is applied if the two SNs share the same referent.

- (11) **1** I am of the *opinion* that governments should increase the price of gas to solve ‘some’ traffic problems. // **2** However, my agreement with such an *opinion* has limitations. (ALEC)

In Example (11), the first mention of the noun *opinion* (T-unit 1) acts as an in-t-unit SN and hence will not be considered. However, it will be picked up as simple repetition in the analysis throughout the present study. Conversely, the second mention of *opinion* (T-unit 2) acts as an anaphoric across-clause SN whose lexical realisation is the same as the first mention and it will thus be counted. This counting procedure is according to the principle I established in (b) which gives the priority of counting to the across-t-unit SN function, which

is the main function that the present study analyses. Although the second mention of the SN with the across-t-unit function is only a repetition of the earlier noun and does not fulfil the characterisation function by labelling anything fresh, it is still counted. This counting decision is not in line with Francis's (1994) criterion which discards anaphoric nouns from the count if they are preceded by their repeated form. My decision to count the second mention of *opinion* in Example (11) as an SN (not a simple repetition) is because *opinion* is still picking up the previous SN. Also, the structure in which this SN occurs represents a common feature in the Arab learner writing where participants employ the same SN which functions initially as in-t-unit and then is followed by its repeated form that acts as across-t-unit. Thus, it is worth capturing the second mention of *opinion* as an instance of an SN instead of repetition to examine how frequent this kind of structure is in learner writing (see Chapter 6 for more examples of this use of SNs).

2. (a) When the same word-form of an SN is repeated within adjacent T-units and the other mention of this word-form is still signalling forward helping to carry the first SN until it is resolved, I will consider the second mention as an example of repetition. This repetition functions primarily to provide a further description of the previous text. Therefore, I will only count the initial mention as an SN as long as it functions as across-t-unit SNs. Consider Example (12).

- (12) **1** The immigrants would bring their inherent *skills* and *behaviours* into the new country. // **2** Indeed, some of those *skills* and attitudes might be of considerable interest to the original population // **3** as they widen their horizon about the varieties of options available worldwide. // **4** For example, in the UK many restaurants started adding new recipes to their menu which are taken from chef immigrants and gained the interest of the original British costumers. (ALEC)

In Example (12), the first abstract noun *skills* (T-unit 1) functions as an across-t-unit SN whose lexical realisation is underlined (T-unit 4). Thus, *skills* will be counted. However, the second mention of *skills* in T-unit (2) could be considered as a continuation of the first SN in that it holds the idea of *skills* for a while in order to keep it in the reader's mind until it is resolved later. As a result, the second mention of *skills* is counted here as simple repetition.

(b) Looking again at Example (12), it contains another interesting feature which is common in the writing of Arab L2 learners. It is the use of lexical strings or couplets. Lexical strings or couplets are the use of “a phrase or sentence coordinating two or more words with shared semantic components and a single referent” (Rieschild 2006: 6; Johnstone 1983). In English, lexical couplets are mostly used in idioms like *bits and pieces*. However, the analysis of the present study does not analyse this idiomatic use but it looks at non-idiomatic uses of this structure when it contains SNs. The phrase *skills and behaviours* in T-unit 1, for instance, coordinates two abstract nouns, both of them are SNs of cataphoric function. These two SNs also share a common lexical realisation (underlined in T-unit 4) despite the fact that the Arab L2 writer does not produce a proper specific to reflect both SNs successfully. In this case, the methodological question is: Do I need to count SNs that are part of a lexical string/couplet as two elements – or do I need to count only one SN considering the second mention as an empty element? Or I perhaps count the couplet as one unit? In the present study, I will count both SNs in the couplet whose lexical content is the same. This decision is built up on a methodological basis. Arab speakers of English use a number of lexical strings/couplets that contain two different SNs which might be worth including in the counts of SNs (see Chapter 6 for further examples of lexical couplets/strings).

3. (a) When an SN whose function is across-t-unit is reused throughout the text referring to the same idea that is previously signposted by the same form of the SN, it will be counted as an instance of simple repetition not an SN. Consider example (13) (Note: T-units do not form a complete essay. The square brackets indicate T-units that are not presented).

(13) **1** With regard to the positive impacts and *advantages* that immigrants can bring to those host countries; there are many. // **2** At the economic level, immigrants can have a contributing impact on job creation. // **3** First, with money they may bring, immigrants can invest, make businesses and create new jobs. // **4** Second, creating jobs means employing workers and helping the jobless earn their living. // **5** Third, among immigrants, there can be a number of talented persons from whom the host countries can benefit. [...] **15** Based on the above mentioned points and *advantages* of immigration flows, it may not be very wise to stop such empowering gains as their benefits to the host country may outweigh their drawbacks. (ALEC)

Example (13) contains two instances of the same word-form of the SN *advantages*. The first word-form in T-unit 1 functions as an across-t-unit SN with a cataphoric function. The second mention then occurs in T-unit 15 and also acts as an across-t-unit SN but with an anaphoric function. Both nouns refer to the same information (i.e., lexical realisation) which is underlined (T-units 2, 3, 4 and 5). In this case, we could count both nouns giving that each one has a different function, and also this might be helpful later if we intend to distinguish the different functions of SNs (i.e. anaphoric and cataphoric). However, I would prefer to count only one instance which is the initial mention of the SN, and consider the other mention(s) to be simple repetition. This is because the repeated mentions of SNs do not refer to a different lexical content in their text to be counted. Besides, this counting decision is to leave room for

SNs with fresh insight in the text to be distinct from simple repetition, which is another category of lexical cohesion that will be counted in a separate stage throughout this study.

(b) On the other hand, if the same SN is repeated over the course of the text but has a different lexical realisation to its initial lexical realisation, then it will be counted. Example (14) demonstrates this principle.

- (14) **1** A wave of immigration has spread throughout the world especially among developed countries in Europe and United States of America. // **2** This has created a number of controversial arguments regarding whether this flow of immigrants should be banned, controlled or left as it is. // **3** People who argue that it can be left as it is think that banning it may lead to undesirable economic and cultural consequences. // **4** Still, this *issue* has both positive and negative effects. [...] **29** Waves of immigrants can cause a rise in the cost of living. // **30** Large number of people means more need for food especially savoury ones and other products, rise in costs of renting houses. // **31** But if we look at the *issue* from a different perspective, we can see that more demand on food for example means more food stores to employ workers and create more jobs.// (ALEC)

In Example (14), the SN *issue* is repeated over the text in T-units 4 and 31. However, as the example demonstrates through the underlining, *issue* has different realisations every time when it is repeated. That is, *issue* in T-unit 4 encapsulates T-units 1, 2 and 3 whereas *issue* in T-unit 31 packages the information in T-units 29 and 30. Accordingly, both instances of *issue* will be counted. Example (15) below is another case where the Arab L2 writer uses the SN *impact(s)* in two different ways as follows:

- (15) **1** With regard to the positive *impacts* and advantages that immigrants can bring to those host countries; there are many. // **2** At the economic level, immigrants can have a contributing *impact* on job creation. // **3** First, with money they may bring, immigrants can invest, make businesses and create new jobs. // **4** Second, creating jobs means employing workers and helping the jobless earn their living. // **5** Third, among immigrants, there can be a number of talented persons from whom the host countries can benefit // [...] **7** The sense of belongingness that immigrants may develop will motivate them to serve that country whether in the army or in other aspects of life. // **8** Once they earn their degree, they will probably choose to serve the country to which they feel they belong. // **9** This means that more expertise and talents are brought with these waves of immigrants. // [...] **11** Politically socialised, they can have a loyal and emotionally charged feelings towards a certain party and can lead a powerful movement towards change in benefit of the country and its people. (ALEC)

In Example (15), the first use of *impacts* in T-unit 1, which is a plural form, is a broad one in that *impacts* compresses all the underlined text as its lexical realisation whereas the singular noun *impact* in T-unit 2 labels only a specific point that is related to job creation (indicated by T-units 3, 4, 5, 7, 8 and 9). Thus, in this case the lexical realisation of both nouns overlaps but is still slightly different and hence both SNs will be counted.

4. If the same SN fulfils two functions at the same time by acting as an anaphoric across-t-unit SN, and as a cataphoric across-t-unit SN, I will count one function only – the one that comes first in terms of order of occurrence. Very few examples of this case exist in the ALEC corpus. Example (16) illustrates this case in the following way:

- (16) **1** Human brain processes data in a parallel way, which means that all the processes that are involved in processing a particular task occur simultaneously. // **2** Brain has millions of neurons or cells that are connected to each other and many of them can operate at the same time to process this particular task. // [...] **4** Accordingly, human brain has a greater processing *capacity* than any advanced computers. // **5** A practical example of this processing power, human brain can process and analyse very complex pictures far quicker than the most advanced computer. (ALEC)

In Example (16), the noun *capacity* in T-unit (4) can act as an anaphoric SN labelling the prior segment that is indicated by the underlining in T-units 1 and 2 as a *capacity*. Simultaneously, *capacity* can function as a cataphoric label preparing the reader to anticipate more information in the following discourse.

Overall, all the aforementioned examples in Section 2.7.1.1 suggest that the phenomenon of repeating SNs in a text is very common in Arab L2 writing. The following sections will cover other types of repetition, and the way they overlap with SNs.

2.7.1.2 Signalling nouns and nominalisation

Flowerdew (2009) observes that SNs are firmly linked to ‘nominalisation’. Flowerdew and Forest (2015: 12) maintain that ‘nominalisation’ is a productive process for the creation of SNs. McArthur (1996: 403) defines ‘nominalisation’ as “the process or result of forming a noun from a word belonging to another word class”. This definition reveals that nominalisation is not restricted to include only verb-based nominalisation (e.g., *belief* – *believe*), but it could be formed from other parts of speech such as adjectives (e.g., *intensive* – *intensity*). Researchers such as Thibault (1991: 282) and Ventola (1996) notice that nominalisation has a function of packing the lexical content of clauses into single noun

groups. They claim that by changing verbs and other parts of speech into nouns, the content of the text and its lexical density will increase. Halliday (1994) and Pueyo and Val (1996) add that an enormous amount of information in a text can be compacted by the use of nominalisation, which enables to present a continuous argument and helps to construct new concepts. This ability of nominalisation to condense meaning represents one of the major functions of SNs which make both categories overlap. Moreover, nominalisation represents one of the essential processes to produce derived repetition which is one kind of the lexical cohesive devices in the present study. Therefore, I need to identify the criteria of how to count SNs as a distinct category from nominalisation when they appear within the same text in adjacent pairs.

In his contrastive study of English and Spanish, Álvarez de Mon (2006) separates the class of retrospective and prospective nouns from nominalisations. Álvarez de Mon (2006: 25) argues that while the former class characterises a sentence through a noun that does not derive from any of the words in the preceding or succeeding sentence as in Example (17), the latter merely nominalises one of the verbs or adjectives in that sentence, as in Example (18). Álvarez de Mon (2006) follows in this view Francis (1994) who asserts this point when she claims that the verb or adjective from which the anaphoric noun is derived has not to be present in the previous sentence. Examples 17 and 18 illustrate this use of SNs (Álvarez de Mon 2006: 26):

- (17) Because gallium arsenide consumes less power, it produces less waste heat [...].
This *quality* is particularly valuable [...]

- (18) It sweeps the clusters along into an evacuated chamber, where the pressure differential causes the spray to expand supersonically. Collisions that take place during the *expansion* [...]

Benitez and Thompson (2015) comment on the use of SNs, as the one demonstrated in Example (18), that although the nominalised noun (i.e. *expansion*) encapsulates, it does not characterise the lexical content (indicated by the underlining) differently. Therefore, *expansion* in this case does not fulfil the criteria of shell nouns (SNs in the present study). In the present study, instances of nominalisation (e.g., *expand – expansion*) will not be counted as SNs but they will be picked up as instances of derived repetition. Consider Example (19) from ALEC.

- (19) **1** It is natural for individuals to perceive things that are more expensive to be of higher quality than things that are priced cheaply. // **2** However, it is only a *perception* after all. (ALEC)

In Example (19), the noun *perception* (T-unit 2) encapsulates the previous stretch of text, but it does not categorise it differently. It is just a nominalised form of the verb *perceive*. Thus, it will not be counted as an SN. However, if the SN (in a form of a nominalised entity) comes before the verb or the word from which it derives, I will count it. Consider the following example:

- (20) **1** We do make judgments about everything in our lives and our *perceptions* affect our opinions. // **2** It is natural for individuals to perceive things that are more expensive to be of higher quality than things that are priced cheaply. (ALEC)

In contrast to Example (19), the SN which is represented in the nominalised form *perceptions* precedes its derived verb *perceive*. Thus, it will be counted in this case. The initial occurrence gives the SN a feature of prospection in that the reader will look forward to what will be stated as *perceptions*. Consequently, I will assume here that order is a key factor in deciding the membership of SNs. This is only an analytical decision and has no theoretical basis.

2.7.1.3 Signalling nouns and synonyms

Benitez-Castro (2015: 184 – 185) posits that synonyms do not function as shell nouns (SNs in the present study), because the semantic meaning of the nominal label and the previous mention to which it refers is undetermined, and both nouns carry the same or similar meaning. Example (21) from Gray & Cortes (2011: 36) illustrates this point.

- (21) The next phase of research will consider ways in which teachers might best raise learner consciousness of the importance of theme in English information structure, and how *this awareness* may be activated to help learners produce fully coherent written discourse.

In Example (21), Gray & Cortes (2011) claim that the noun phrase in italics *this awareness* acts as shell whereas Benitez-Castro and Thompson (2015) contend that this phrase does not reveal any new characterisation of its antecedent because *awareness* and *consciousness* carry the same concept. Benitez-Castro and Thompson (2015) observe that although *this awareness* encapsulates the underlined noun phrase, it does not re-enter it differently. Instead, it merely operates as a synonym of *consciousness*. Thus, these researchers agree with López Samaniego (2011: 423) and postulate that *this awareness* is not a case of a shell noun, but of ‘encapsulating synonym’. They explain that this nominal phrase satisfies Schmid’s (2000)

function of temporary concept-formation (i.e. it encapsulates), but not that of characterisation. I could, therefore, establish my principle of counting which reports that an SN will not be counted if it is preceded by its synonym. In both ALEC and LOCNESS, no examples such as Example (21) are found. Instead, most of examples, if any, include cases where an SN is either used within the text (as Example 21) which falls outside the scope of the present study, or it acts as a cataphoric noun as in Example (22).

- (22) 1 Boxing, nowadays has a certain *aura* about it // 2 and the atmosphere is almost electric when there is a little fight. (LOCNESS)

In Example (22), the noun *aura* is an across-clause SN with a cataphoric function. The underlining text that follows it represents its semantic specific. This specific contains a synonym of *aura* which is *atmosphere*. In this case, the SN *aura* will be counted because the rule I mentioned above states that only an SN will not be counted if it is preceded by its synonym. In this example, however, the SN is followed by its synonym and hence it will be recorded. Francis (1994: 86) confirms this principle when she claims that “[t]he head noun of a retrospective label [...] does not have a ‘synonym’ in the preceding discourse, and [...] is actually a new lexical item”. In Example (22), *aura* is presented as a new lexical item.

Furthermore, the difference between true equivalents ‘true synonyms’ and ‘instantial equivalents’ (Francis 1988: 328-330) needs to be considered when analysing SNs. Benitez-Castro (2015) assumes that only true equivalents (cf. Example (21) above) will be excluded from the counts of shell nouns whereas instantial equivalents (see Examples 23 and 24) should be included. In my analysis, I will adopt Benitez-Castro’s (2015) strategy of counting.

(23) I called the ROV and explained my *situation* over the phone. A man said to bring the keys to ROV in Sin Ming Drive and that they would handle it. The following afternoon, I drove to ROV and explained my *dilemma* to a woman there.

(Francis 1988: 327)

(24) **1** Computers have negative *impact* on human health. // **2** For example, long hour gazing at the computer screen will negatively affect sight; // **3** further they produce radiations when they are operated with badly affect the heart. // **4** In addition to computers' negative impact on health, they have also negative *effect* on social rapports // [...] recently computers encourage human isolation. (ALEC)

Example (23) by Francis (1988) contains a pair of ‘instantial equivalents’ (*situation* – *dilemma*). According to Benitez-Castro’s (2015) principle, both of synonyms will be counted as SNs. Also, in Example (24) from ALEC, *impact* and *effect* are examples of ‘instantial equivalents’ not ‘true synonyms’ and hence they will be counted following the principle articulated above. The ALEC corpus contains a frequent number of these synonymous SNs and they are therefore worthy of inclusion in the frequency counts.

2.7.1.4 Signalling nouns and text nouns

Flowerdew and Forest (2015) explain that what differentiates SNs from other cohesive devices is that they represent an example of what Halliday and Hasan (1976: 52) describe as ‘extended reference’ compared to ‘text reference’. The ‘extended reference’ refers to “more than just a person or object, it is a process or sequence of processes (grammatically, a clause or string of clauses, not just a single nominal)” (Halliday and Hasan 1976: 52). By contrast, Halliday and Hasan (1976) indicate that the reference in ‘text reference’ is to a person or object. Flowerdew and Forest (2015: 48) support this view and argue that the “lexical

realisation of SNs must, at minimum, construe a process or group of processes in a semantic relation”. However, what about the class of ‘text nouns’, which is a subgroup of SNs? Some of these text nouns cannot label more than one discourse entity.

‘Text nouns’ are defined by Francis (1994: 93) as nouns which refer to the formal textual structure of discourse. She suggests that no interpretation is required for these nouns because they simply label stretches of preceding discourse whose boundaries are precise. She includes in this class nouns such as *phrase*, *question*, *sentence*, *words*, *excerpt*, *page*, *paragraph*, *passage*, *quotation*, *section*, *term* and *terminology*. Researchers such as Schmid (2000) exclude text nouns from SNs because they do not appear in in-clause patterns and thus do not represent a good example from a grammatical perspective. On the contrary, other researchers (e.g., Francis 1986, 1994; Ivanič 1991; Flowerdew 2002, 2003, 2006) whose focus are on semantic and discourse features include text nouns within SNs. Although Flowerdew and Forest (2015) describe text nouns as a type of nouns on the periphery of the SN class, they incorporate them in their analysis of SNs. They suggest that such nouns which refer to parts of the text, such as *section*, *paragraph*, *introduction*, *conclusion*, *chapter*, and *paper* satisfy the basic criterion for SN membership which is encapsulating specific sections of text. Nevertheless, these researchers do not explain how we will address text nouns that encapsulate or prospect a single discourse entity. Nouns such as *figure*, *table*, *word/s*, *term* and *terminology* usually refer to single units in the text as opposed to other text nouns such as *passage* or *paragraph*. The former set of nouns could also have the function of text nouns because they refer to small portions of the text. Some of them are mentioned in Francis’s (1994) list of text nouns (e.g., *term*). However, she did not indicate how these nouns fit in with the class of anaphoric (signalling) nouns regarding the length of their lexical realisation.

On the one hand, Francis (1994) claims that the key principle for identifying a retrospective label is that it does not need to refer back to single entities, but to stretches of text. On the other hand, she allows text nouns such as the noun *term* to be included within this class. This contradiction makes this category of SNs, namely those that encapsulate single textual stretches, questionable. Although my data contains a few examples of these nouns, it is still fundamental to decide how I will count them in both ALEC and LOCNESS. Consider Example (25) from ALEC.

(25) 1 These countries for example like Libya the fuel cost as little as 0.150 LYD a Litre which is about 7p: 10p GBP.// 2 That *figure* is very cheap. (ALEC)

In Example (25), the noun *figure* in T-unit (2) could be treated as an SN because it encapsulates a specific numeric value in the text: *0.150 LYD*. However, *figure* refers to a single entity not a stretch of text. Thus, the question at this point is: Does *figure* need to be counted as an SN or do I need to exclude it from the analysis? This is another example from the ALEC corpus.

(26) 1 It is hard to define “education” // 2 because the *word* is far too big to be consumed in a short definition. (ALEC)

In Example (26), the noun *word* in T-unit (2) can only be specific if we look backward to its realisation *education*. *Word* shares features of signposting and labelling with SNs. I therefore decided to include this class of SNs in my counts of SNs. These text nouns with a single realisation are very similar to the cohesive categories of ‘superordinates’ and ‘general nouns’ where the referent typically takes the form of a single unit. In this study, I could consider text

nouns such as *figure*, *term* and the like as a minor subcategory of SNs that serves as a borderline case between signalling nouns, superordinates and general nouns. All of these categories, however, have a similar function of compressing the text in a way that makes it hang together. Such an overlap leads some researchers such as Martin (1992) to combine these categories together.

2.7.1.5 Partitives

Another category which is subject to debate in terms of whether it functions like SNs or not is identified by Quirk et al. (1985: 249-251) as ‘partitives’, ‘species nouns’ by Biber et al. (1999: 255-7), or ‘enumeratives’ by Tadros (1985: 17). Examples of this group include *class*, *kind*, *part* and *type*. Schmid (2000: 118-119) focuses on the noun *part* and considers it as a shell noun that accomplishes part-whole relations. Nonetheless, he regards these nouns to be peripheral members of the SN class positing that they appear so frequently in non-SN uses. In contrast, Flowerdew and Forest (2015) suggest that partitives will be counted when they function as an SN. Example (27) illustrates how *part* functions as an SN.

- (27) The big difference is the organic systems have excited states and that’s a very important *part* of their radiolysis.

(Flowerdew & Forest 2015: 62)

The present study will follow Flowerdew and Forest’s (2015) approach and count partitives as SNs in both ALEC and LOCNESS. Example (28) is taken from ALEC and it illustrates the use of partitives:

- (28) **1** Education plays a very big *part* in human lives// **2** as it is the source of enlightenment. (ALEC)

In Example (28), the noun *part* in T-unit (1) refers forward to the following discourse in T-unit (2), which is illustrated by the underlining. Therefore, *part* here will be counted as an SN. In this regard, Flowerdew and Forest (2015) call the attention to a number of cases where partitives function as pre-modifiers to other nouns, including other SNs. (e.g., *kind of thing*, *part of the problem*, *type of problem*). Example (29) illustrates this case where *kind of argument* is ‘double-headed’ noun phrase and refers back to the preceding sentence.

If you ran evolution again on this planet you’d get photosynthesis. You’d get life because it’s downhill to chemistry. That’s one *kind of argument*.

(Flowerdew & Forest 2015: 62)

Flowerdew and Forest (2015) point out that the entire structure of the noun phrase *kind of argument* acts as an SN. These researchers prefer to label only one noun in double-headed noun phrases which is the canonical SN as the signalling item rather than the partitive. Thus, in Example (29), Flowerdew and Forest count *argument* rather than *kind of* as the SN. However, they admit that this choice does not indicate that partitives do not share characteristics with SNs. Instead, inclusion and exclusion of these items is a matter of the direction of the analysis. However, it would be more reasonable if we treat the double-headed phrase as one lexical unit rather than arguing whether to count the first noun or the second one in the noun phrase. In this I follow Sinclair’s (1991: 93) conception of semantic headedness in which he considers the constituents of the double-headed phrase as one head. He claims that “[n]either N1 nor N2 can easily be omitted”. Examples of double-headed phrase are very few in ALEC and LOCNESS. In Example (30) below, the nominal phrase *sort of scheme* functions as an SN. Therefore, as explained above, the constituents of the double-headed phrase will be counted as one SN.

- (29) **1** Cities such as Manchester have become largely pedestrianised and have seen the reintroduction of trams, // **2** this is the *sort of scheme* that can solve our problems.
(LOCNESS)

Overall, the overlap between lexical cohesive categories such as simple repetition, nominalisation (one type of derived repetition), synonyms and SNs indicates that semantic categories have fuzzy boundaries due to similarities in their characteristics. This fact makes the analysis of these categories not clear-cut, and there is no right or wrong in specifying the criteria of these categories or in how they are counted. The fundamental point about the counting process is to conduct a consistent analysis which follows certain principles that need to be applied systematically to both corpora in the present study. My counting technique does not posit a hierarchy of semantic categories but every category of lexical cohesion is counted separately. However, in some cases where the overlap between categories is clear, one category is preferred to another. This has been applied throughout the analysis of SNs where in certain cases simple repetition is recorded instead of an SN as a lexical cohesive device. Although this decision is not based on theoretical grounds, it will allow for each category to be distinct from the other, which will in turn facilitate the analysis.

2.8 Lexical cohesion in native and non-native English writing: A special reference to Arab speakers of English

Most studies on lexical cohesion in NNS writing have demonstrated that L2 learners differ from native language speakers in their use of lexical cohesion. This section focuses on three categories: simple repetition, derived repetition and signalling nouns, and how they are used and measured in native and non-native English writing. This will help to understand how

previous research into learner corpora analyse lexical cohesion in the light of text-linguistic approaches.

2.8.1 Simple and derived repetition in native and non-native writing

This section reviews a number of studies that have analysed learner writing in terms of cohesive categories including simple and derived repetition. I will classify these studies into two groups: Studies that analyse English writing of Arab L2 learners with no reference corpus; and studies that compare NNS writing (including Arab L2 writing) with NS writing. Also, it is important to clarify that studies that analysed lexical cohesion using Halliday and Hasan's (1976) model presented only the count for simple repetition or what Halliday and Hasan's (1976) labelled the reiteration of the same lexical item; this is because derived repetition is subsumed under simple repetition in their model. In contrast, studies that analysed derived repetition applied Hoey's (1991b) model of lexical repetition because it distinguishes it from simple repetition.

2.8.1.1 Studies on Arab speakers' English writing

A number of researchers (e.g., Kaplan 1966; Hamdan 1988; Khalil 1989; Shakir 1991; El-Gazzar 2006; Mohamed-Sayidina 2010) analysed cohesion in Arab speakers' English writing. Most of these studies agree that English writing by Arab L2 learners lacks cohesion and these learners are unable to compose adequate themes. Some of these studies are summarised in Table 4 below.

Studies	Lexical cohesion Categories	Focus	Learners' L1 background	No. Of NNS texts/Corpus size	Method of analysis	Unit of analysis
Khalil (1989)	Reiteration, Conjunction, Reference, Collocation, Substitution, Ellipsis	Relationship between cohesion and coherence	Arabic	20 descriptive college essays	Halliday & Hasan's (1976) model; Grice's (1975) method of coherence analysis	T-unit
El-Gazzar (2006)	Simple repetition, Reference, Derivation, Synonymy, Inclusion, Antonymy, Collocation	Quantitative analysis of lexical cohesion	Arabic	40 expository undergraduate essays	Stotsky's (1986) model; Hoey's (1991b) model; and Salkie's (1995) model	Sentence
Mohamed-Sayidina (2010)	Same noun repetition Reference, Substitution, Ellipsis	Contrastive rhetoric	Arabic	50 argumentative English compositions	Halliday & Hasan's (1976) model	Sentence

Table 4 Summary of lexical cohesion studies on Arab L2 English writing

Table 4 shows previous studies that analysed cohesion in the English writing of Arab L2 learners. The number of Arab L2 learners in these studies ranges from 20 to 50. Simple repetition is analysed by most studies whereas derived repetition has received little attention by researchers; it is only analysed when Hoey's (1991b) model is used. Furthermore, these studies do not compare their results with English native data. Gablasova et al. (2017: 131) state that "frequency information about features in L2 production is undoubtedly a valuable source of evidence about language development and use; however [...], it would be difficult to interpret this evidence without a reference point". Granger (2015) explains that a comparative research design enables researchers to study and examine L2 corpus evidence, and to interpret the findings more critically than if an L2 corpus is studied individually. However, the studies

that are shown in Table 4 are still relevant to the present study because they provide valuable information on the problems that Arab speakers of English have in their English writing. One of these problems is that Arab learners of English overuse simple repetition as a cohesive device, but this overuse does not make their texts coherent. For example, Khalil (1989) analysed cohesion in 20 English one-paragraph compositions written by Arab EFL college students. Based on Halliday and Hasan's (1976) model, the analysis showed that Arab students used simple repetition excessively. Khalil (1989) complemented the cohesion analysis by evaluating coherence. He measured the correlation between cohesion and coherence by carrying out multiple correlation statistics and found only a very weak correlation between the number of cohesive ties and the coherence score of the text ($r = 0.18$). Khalil (1989) explained that if we examine the essay with a highest score of cohesion, we observe that the writer developed the main topic by supporting his argument with explicit examples. In addition, the writer employed cohesive ties successfully by connecting the main topic with the subtopics. Conversely, the writer of the lowest coherent composition introduced two main topics without supporting them with a precise content. The coherence analysis proved that Arab students did not use cohesive devices to present adequate information about the specified topic, and that the kind of information that was contained in their writing includes conceptual redundancy.

The heavy use of lexical repetition is further justified by El-Gazzar (2006) who observed that Arab L2 learners' knowledge of vocabulary was not sufficient to allow them to communicate their ideas effectively. El-Gazzar (2006) conducted a quantitative analysis of lexical cohesion in the expository writing by 40 undergraduate Arab students registered in advanced academic writing classes. El-Gazzar's (2006) study did not follow Halliday and Hasan's (1976) model

because she claims that their framework is developed for analysing conversation and literary discourse only. Consequently, El-Gazzar (2006) applied a mixture of models suggested by Stotsky (1986), Hoey (1991b), and Salkie (1995). El-Gazzar (2006) found that the frequency of lexical cohesive devices varies depending on the category used. She ordered these categories according to their frequency from highest to lowest. Simple repetition was the most frequently used lexical cohesive form in the written products forming 45% of the total lexical types. However, derived repetition ranked 4th accounting for 11.1%. This category was relatively frequent compared to synonyms that established 4.3% or antonyms that represented only 1.4%. El-Gazzar observed that most of derivations that appeared in the English writing of Arab participants were derivational suffixes (such as *nation*, *national*, *nationalism*). However, this focus on forms of cohesion rather than on its textual function is still not enough to reveal how repetition functions in text to create lexical cohesion. Other researchers such as Mohamed-Sayidina (2010) relate overuse of word repetition to the problem of L1 transfer, which is the influence of a learner's first language (L1) on their second language (L2). Kohn (1986: 21) points out that "transfer is one of the major factors shaping the learner's interlanguage competence and performance". Mohamed-Sayidina (2010) analysed 50 English compositions written by native speakers of Arabic, studying an Academic English course. She adopted Halliday and Hasan's (1976) taxonomy of lexical cohesion and analysed repetition of the same noun, and grammatical cohesion. She compared the use of these cohesive devices in the English compositions with those used in native Arabic texts. She conducted a t-test to determine whether the means of repetition of the same noun and grammatical cohesion are statistically significant. She found that, as in Arabic native texts, the mean of simple repetition was twice the mean of grammatical cohesion, and the means were significantly different. She observed that Arab English writers repeated the same noun several

times in adjacent sentences. She explained that these writers transferred the rhetorical patterns from their native Arabic texts to their English written essays, and the cohesive forms that these writers selected were also influenced by cultural and literacy conventions in the Arabic-speaking countries.

2.8.1.2 Studies that compare NNS and NS writing

To my knowledge, studies on lexical cohesion that analyse English writing of Arab speakers using a control corpus of English native speakers are very few. Table 5 below outlines some of these few studies that are related to the present research.

Studies	Lexical cohesion Categories	Focus	Learners L1 background	Number of NNS & NS texts/Corpus size	Method of analysis	Unit of analysis
Connor (1984)	Same item, Synonym, Superordinate/ General word, Collocation	Relationship between cohesion & coherence	Spanish Japanese	2 NNS argumentative essays/ 2 NS argumentative essays	Halliday & Hasan's (1976) model	T-unit
Parsons (1991)	Co-reference, Co-classification, Co-existence	Relationship between cohesion & coherence	Arabic	8 NNS descriptive essays/ 8 NS post-graduate descriptive essays	Hasan's (1984) model	Sentence
Reynolds (1995)	Simple/complex repetition, Simple/complex paraphrase, Co-reference, Hyponymy, Superordinate	Cohesion is quality more than quantity	Arabic Japanese Romance Korean Taiwanese	26 NNS expository essays/ 16 NS expository essays	Hoey's (1991b) model	T-unit
Kai (2008)	Simple/complex repetition, Simple/complex paraphrase, Co-reference, Hyponymy, Superordinate Substitution Ellipsis Deixis	Similarities & differences in lexical cohesion between NNSs & NSs	Chinese	15 NNS/ dissertation abstracts 15 NS dissertation abstracts	Hoey's (1991b) model	Sentence

Table 5 Summary of cohesion studies on NS and NNS English writing

Table 5 shows that comparative studies of cohesion in NNS/NS writing differ from each other in a number of variables. These variables include what cohesive categories are counted, learners' L1 background, the text type, which model of cohesion analysis is adopted, and the unit of analysis – whether a sentence or a T-unit. The present study shares some of these variables with these studies. For example, most studies analysed simple and/or derived repetition, which are among the categories that this study analyses. The present study is also

similar to Connor's (1984) study in that it focuses on the text type of argumentative essays. Other researchers analysed different text types such as expository, descriptive or dissertation abstracts. Lexical cohesion is possible in all text types. However, cohesive categories vary with text genre, as Harnett (1986) observes. Thus, studies on cohesion should take into account the specific discourse being analysed. Harnett (1986) adds that researchers should not simply select cohesive taxonomies that have been delineated for descriptive linguistic purposes (e.g. Halliday and Hasan's 1976). Instead, they need to take into account the methods of analysing cohesion based on its function in the written products. For the unit of analysis, the present study resembles Connor's (1984) and Reynolds's (1995) selection of the T-unit as a measure to analyse lexical cohesion (a justification for this selection is presented in Chapter 4).

The present study, however, is different from the studies presented in Table 5 in that it does not apply traditional models to analyse lexical cohesion, but it uses a different method of analysis based on wordlists. This method will not be presented at this point of discussion because an evaluation of the traditional models is required (cf. Section 2.9). Based on this evaluation, this study will suggest a method of lexical cohesion analysis that fits in with type of data analysed in this study (see Chapter 5 for an introduction of the lexical repetition network model). A further observation in Table 5 is that most studies include non-native speakers of English who are from linguistic backgrounds other than Arabic, or they include Arab speakers of English as a single group among other groups of non-native speakers of English. For example, Arab L2 learners in Reynolds's (1995) study were represented by only 6 participants. Only Parsons (1991) includes in his study 8 participants who constitute a separate group of Arab speakers of English, but this number is still small. Connor's (1984)

study is also based on two NNS essays and two NS essays, and it has been criticised by Granger and Tyson (1996) for this small-size corpus. Nevertheless, Connor (1984) argues that this limitation in corpus size becomes reasonable when we consider the time needed to identify relations between cohesion and coherence. Although the number of Arab participants in the present study is also not large (28 participants), it is still adequate because the present study analyses another set of English native data which is composed of 30 participants (see Chapter 4 for corpora used in this study).

The studies summarised in Table 5 include Arab L2 learners as a small group. However, the findings of these studies reveal important points that address how cohesion is analysed, and how differences in the use of lexical cohesion can be spotted between NNSs and NSs. Connor (1984), for example, used Halliday and Hasan's (1976) method to compare the use of lexical cohesive categories in argumentative essays written by native-speakers of English and ESL students. The native-speaker compositions were drawn from a PhD dissertation, whereas ESL students were instructed to write about seven topics and the total number of essays for both groups was four: two native-speaker texts and two ESL students' texts. In contrast to Halliday and Hasan's (1976) model, Connor (1984) used the T-unit as a unit of analysis. She provided frequencies of the sub-categories of lexical cohesion as percentages, but did not demonstrate which lexical items were counted or how she operationalised definitions of lexical cohesive categories, such as collocation, for instance. She presented figures that represent collocation for both groups without explaining how this category was measured, or how many words constituted the collocation. Such an explanation is necessary as it will help any analyst to replicate the counting method to any study on lexical cohesion that involves comparing learner writing with native data.

Connor (1984) demonstrated that the two participants in the NNS group used a high percentage of simple repetition representing 84% and 73% of the total lexical cohesive ties. In contrast, the two native speakers displayed less simple repetition accounting for 32% and 61% correspondingly. These results, as Connor observed, indicate that ESL writers have a narrow range of lexical cohesive devices that allow them to extend concepts they introduce. Rather, their writing has more lexical and conceptual redundancy. Connor (1984) investigated specific lexical ties in both groups of students' essays to examine what kind of word range students used in their writing when they wanted to develop their argument of the discourse topic in focus. For example, she analysed the ways by which one of the native speakers expressed 'tests' and 'testing situations' in their essay about 'testing'. Connor (1984: 307) found that the student used a variety of synonyms referring to *tests* such as *methods of measuring, a set of questions, a satisfactory means of measuring a student's achievement, means of testing, gauge a student's mastery, and administering examinations*. Conversely, when she studied an essay written by one of the ESL students, she noticed that the student restricted their writing to few words that refer to tests such as repeating the word *tests*, or using phrases like *put scores*. Also, Connor (1984: 307) found that the native speaker described the negative effects of testing as *psychological tension, attitude of apprehension, and fear*, whereas the ESL student expressed this meaning by repeating these phrases: *getting nervous and feeling very, very bad*.

Although Connor (1984) found that one of the ESL students used a rich variety of cohesive devices including 'repetitions, synonyms and collocation', this increase did not correlate with a better composition, which means that the student's writing would not be rated as a good essay. Hence, she affirmed that cohesion alone does not lead to a coherent text. To distinguish

the quality of writing for both groups of students, she conducted an analysis of coherence. She examined the communicative functions of T-units in the argumentative essays by applying Tirkonnen-Condit's (1984) model of text argument structure to her data. This model explains that "a sequence of two units in an argumentative text is used to assert a claim, then to introduce observations to justify the claim, and finally to induce, by virtue of the observation, the original claim" Tirkonnen-Condit's (1984: 4). A particular advantage of this method of analysis is to identify whether students are successful in tying their T-units in a coherent way through interlinking the three basic parts of the argumentative structure adequately. This method is also used to examine whether there is a development of the main discourse topic through a progression of the subtopics throughout the text. Connor (1984) divided each essay into its T-units and determined the interactive relationship between the T-units following the basic structure of the argumentative text. She then assigned the three functions of text argument structure to all T-units in each text in her study. Afterwards, she identified 'subtopics' in each text. The 'subtopics' consist of a sequence of sentences that share the same topic sentence relating to the main topic (Lautammatti 1978). Connor (1984) reported that her ESL students, as compared to NS students, did not justify their claim adequately in their argumentative essays, and did not connect their concluding inductive statements with the previous subtopics of the problem. This shows that even when ESL students use a wide range of cohesive ties, their essays might lack coherence. The type of cohesive devices that non-native students used was not utilised effectively, but these devices were 'tautological', as Winter (1974) describes. This means that NNS writers repeated lexical items more than it is required to give the sense intended. But this repetition does not have any cohesive function, which is the provision of new information that is essential to the development of text argument. This repetition is therefore tautological according to the description given.

The view that is taken by Connor (1984) is consistent with those who acknowledge the positive contribution of cohesion to coherence in texts such as Parsons (1991). Parsons, applying Hasan's (1984) model of cohesion, found that NSs wrote better organised texts than Arab non-native speakers of English. Parsons (1991) observed that NSs used cohesive devices to form lexical chains that interact with each other across the text building up a network of lexical cohesion. Likewise, Reynolds (1995) examined how sentences are strung in non-linear ways to create cohesion and coherence. To achieve this aim, he applied Hoey's (1991b) model to investigate whether this model could reveal any significant differences between native and non-native learners. Reynolds (1995) inspected whether the theory of lexical cohesion requires simply the analysis of the general use of repetition, or it also needs to consider the proficiency of particular uses. He included Arab speakers of English as a small group (6 participants) with other ESL participants. He compared expository essays written by 26 intermediate ESL non-native learners of English, and 16 native English speakers. Both groups wrote about the same essay but under different conditions. The NNSs composed the essays at the final exam of their English programme while the NSs wrote their essays during normal class.

Reynolds (1995) identified categories of lexical repetition in each essay in his corpora and divided the essays into T-units. Then, Reynolds (1995: 191) provided basic quantitative measures of each essay such as the 'number of words per essay', 'T-units per essay', 'words per T-unit'. He suggested that such information gives a preliminary idea of the length of the essays and their discourse structure. With the help of the index of the T-units per essay, Reynolds (1995) identified the different types of lexical repetition that connect each T-unit in each essay in both corpora. He compared the first T-unit in each essay to all subsequent T-

units, and then the second T-unit to all subsequent T-units, and continued this procedure until the end of that essay. In Figure 3, I visualise Reynolds's (1995) net-like process of lexical cohesion analysis. This Figure displays the way of comparing T-units across each other in each essay in the corpus. The arrows indicate a continuous process that lasts up to the end of the essay under scrutiny.

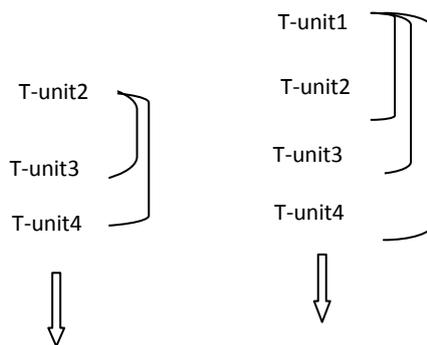


Figure 3 The connection between T-unit 1 and all successive T-units

Afterwards, Reynolds (1995) recorded all marked repetition links and put them into a matrix. He built a matrix for every essay in his corpus which contains 42 essays consisting of 13,170 words. He did not, however, explain how far Hoey's (1991b) repetition matrix could be replicated when dealing with a large-scale corpus for comparative purposes. Reynolds presented only an analysis of the first two paragraphs of an essay from his data and provided a simple matrix in his paper. Figure 4 illustrates a sample of this coding matrix of an essay. The codes that are used in the matrix below are explained by Reynolds as follows (cpa: complex paraphrase-antonym; dx: discourse external; sr: simple repetition; sub: substitution).

	1		
2	<i>dx: I – I</i> <i>sr: like – like</i> <i>sr: Sunday – Sunday</i>	2	
3	<i>dx: /- /</i> <i>sr: Sunday – Sunday</i>	<i>dx: /- /</i> <i>sr: Sunday – Sunday</i>	3
4	<i>dx: I – I</i> <i>sub: Sunday – Mother</i> <i>sr: day – days</i> <i>sr: week – week</i>	<i>dx: I – I</i> <i>sub: Sunday – other</i>	<i>dx: /- /</i> <i>sub: Sunday – other</i> <i>cpa: sleep – wake up</i> <i>sr: morning – morning</i>

Figure 4 A sample of Hoey’s (1991b) coding matrix of repetition links applied by Reynolds (1995)

(adapted from Reynolds 1995: 192)

Based on this matrix, Reynolds (1995: 191) quantified lexical cohesive categories for each essay and created four quantitative measures – ‘frequency count of repetition categories’, ‘ratio of repetition to paraphrase’, ‘density of links and bonds’, and ‘use of bonds at paragraph boundaries’. He postulated that such measures help compare the essays of both groups of NS and NNS from different dimensions. Only the first measure is relevant here because it focuses on frequency which the present study aims to quantify. For the first measure, Reynolds (1995) initially counted the frequency of repetition types for each essay in both groups. Nevertheless, the breakdown of these frequencies for each essay was not shown. Instead, he quantified the overall mean frequencies of lexical repetition types for each group, but did not provide formulas that he used. Therefore, I attempted to deduce how these figures were derived to understand how lexical cohesion was measured. For instance, he applied the following formula to measure the mean frequency of a particular repetition category in one group of learners:

$$\frac{\text{Total quantity of a repetition category in one corpus}}{\text{Average number of T-units per essay}}$$

Reynolds (1995) points out that in order to allow for an accurate comparison between the two corpora that contain essays of different lengths, the raw frequency needs to be divided by the number of T-units. This normalisation of frequency is crucial when comparing different length of texts. Focusing on simple and derived repetition, Reynolds (1995) applied the above formula and found that the most frequent type of cohesive forms for both groups was ‘simple repetition’ with an overall mean score of 7.43 for NS and 6.12 for NNS. On the contrary, complex repetition (derived repetition in the present study) constituted 0.25 for NS and 0.12 for NNS. Both simple and derived repetitions were higher in NS’s writing. However, according to Reynolds (1995), a two-tailed t-test comparing the two mean frequencies for each type of repetition was not significant. Comparable to Connor’s (1984) findings, Reynolds (1995) realised that these quantitative measures of repetition did not pinpoint any significant differences including the measure of the density of bonds that is suggested by Hoey (1991b). This measure does not count single lexical repetition instances, but it quantifies ‘bonds’ which need to contain three or more repetition links to allow significant connections between sentences. It is these sentences which are likely to be more useful in organising text and developing the theme of the text (Hoey 1991b). This implies that ‘bonds’ give a measure of saliency by shedding light on sentences that share multiple occurrences of repetitions, whereas the ‘frequency count’ measure provides only frequency. But does saliency always suggest significance? Does the measure of the density of repetition bonds always mark the text as significant? Does the existence of multiple repetitions across sentences suggest that these sentences are of salient features?

An initial answer to the above questions appears in Reynolds's (1995) study, which interestingly found that Hoey's (1991b) measure of density of bonds does not always work. According to Reynolds (1995), employing this measure did not show significant differences between NNS and NS groups in producing cohesive texts. The NNS learners revealed cases where multiple repetitions highlight information that is not central to the text argument and cases where such bonds of repetition are strung together without any meaningful focus. The issue of lexical cohesion is then more than simply counting the frequency of repeated words. And even when there are two T-units/sentences or more that are linked with more than three repetition links, this does not guarantee that these sentences are central to the argument of the text, as Hoey (1991b) claims (at least in the case of learners of English). Therefore, although the bonds measure is a useful tool for guiding researchers to significant sentences in a text that are likely a source of textual cohesion, researchers should not focus intensely on the number of repeated lexical items that are shared by two sentences. Rather, they need to examine differences in word usage of these repeated items when comparing texts to have an accurate picture of how repeated words function in texts. This means that saliency does not necessarily lead to significance – an issue which is similar to the keywords concept in corpus linguistics in which words that are repeated more often are expected to be key in the text which might not always be true.

One of the problems of the keyness procedure in corpus linguistics relates to frequency and salience. For example, as Wynne (2005) points out, words can show up as keywords because they are a type of high frequency words, which may indicate style, rather than aboutness. Or words can appear as keywords because they are topic-related. Baker (2004) adds that it is also possible for a word to be considered as key if it is frequently used in a particular section of a

corpus, so it is a matter of dispersion or word range issue. A further problem of keyness is that “key words only focus on lexical differences, rather than semantic, grammatical, or functional differences” (Baker 2004: 354). To relate the keyness procedure to Hoey’s (1991b) method of bonds density, Hoey (1991b) relies on identifying where there is salient repetition to come up with keyness, and this keyness to him is what makes a sentence prominent. However, as Reynolds (1995) observes, Hoey’s (1991b) method does not work in learner writing. In the case of L2 learner writing, words that appear as key words in a text are often a result of an unnecessary repetition of lexical items, and not because they are the most significant in a text. On the one hand, keywords, either from a corpus-linguistic perspective or from a traditional viewpoint, suggest an approach that helps to be aware of the topic or style of a text. However, they cannot explain a text, and the analyst therefore has to be cautious about interpreting the results. Furthermore, the degree of salience of lexical items should only be considered by studying the function of lexical items that are unusually frequent.

So what is the possible measure of lexical cohesion then? How can differences between students’ writings in terms of the use of lexical repetitions be spotted? Reynolds (1995) ascribed the failure of quantitative measures to distinguish the writing of native and non-native groups not to the language proficiency of students, but to the large degree of variation in the use of lexical cohesive devices amongst language users of all sorts. Such variation indicates that a specific NS norm, which represents the right or wrong quantity of repetition, does not exist. This means that there is no native corpus which could be used as a typical reference point to rely on. Therefore, Reynolds (1995) posits that differences between NNSs and NSs can be noticed by examining in what way repetition is used and how the two groups use repetition to communicate. Reynolds refers implicitly to a crucial point that can be related

to ‘semantic prosody’ (Sinclair 2004: 34), which is an essential constituent of a lexical item without which “the string of words just ‘means’ – it is not put to use in a viable communication”. This suggests that a lexical item has to carry out a discourse function when it is repeated and not to appear many times without serving a communicative purpose. This means that when L2 writers repeat a lexical item across text, they should not repeat single words. Rather, they also need to repeat lexical, grammatical, semantic, and pragmatic patterning that the repeated lexical item establishes with its context of occurrence. This repetition, as Mahlberg (2006: 371) points out, will create “units of meaning [that] fit in with each other in the creation of text”. Lexical repetition will therefore have a communicative function by contributing incrementally to the creation of cohesion (see Chapter 7 for further details on how semantic prosody plays a part in the creation of cohesion).

For this reason, Reynolds (1995) examined some repeated lexical items and checked how students used them to develop their argument. For example, he studied the word *day* in an NNS essay and noticed that the student referred to *day* in 18 out of 29 T-units (62%), which is related to the subject matter of the essay (i.e. what is your favourite day?). But this word does not present a thesis nor develop it. Furthermore, this repetition leads to ‘unintelligibility’, and to avoid this, Winter (1979) suggests that there should be a kind of lexical exchange to understand why the word is repeated. In contrast, the NS student referred to the same word *day* in 46 out of 64 T-units (72%), but made it central to the thesis of the essay and used it to develop the idea that is being repeated through lexical replacement and an effective use of bonds. Even where NSs used a low level of repetition, as Reynolds noticed, NSs compensated this lack by using other means to develop the discourse topic and employing repetition to tie the text argument across the text.

From Reynolds's (1995) study, we can deduce that both learners tend to repeat with different degrees but the difference lies in why and how they repeat. Does their repetition serve a function in their essays? Or perhaps the repetition they use is just a type of padding. All these questions are relevant to the current study (see Chapter 6 for a qualitative analysis of lexical cohesion). However, at this point I would argue that when we consider why and how lexical items are repeated, we move from a level of analysing surface features of lexical items (i.e. cohesion) to a level of analysing deep phenomenon, which is 'coherence'. This is also achieved in Connor's (1984) study, as mentioned earlier. Connor's (1984) study implies that cohesion has to create coherence, which is a by-product of cohesion. It is this connection between cohesion and coherence in a text that uncovers striking differences between students' writing. When this connection is broken, cohesion alone will not be able to show significant differences in students' writing.

The last study that Table 5 illustrates is by Kai (2008), who looked at whether the dissertation abstracts produced by NSs and NNSs show any similarities and differences in the use of lexical cohesive forms. While the non-native group comes from a Chinese background, not Arabic, Kai's results are still insightful. Based on Hoey's (1991b) model of lexical repetition, Kai (2008) found that simple repetition was used frequently by both groups. However, she observed that NNS abstracts contained more simple repetition than NS ones. The data obtained was analysed with t-tests, but the difference in the use of simple repetition was not significant ($t = 1.114$, d.f. = 24.273, $p = .276$). Kai interpreted these results by positing that the limited vocabulary and the inability to think in English creatively led NNS learners to overuse simple repetition, whereas NSs preferred lexical variety to simple repetition. With respect to complex repetition, Kai found that the difference between the two groups is

statistically significant ($t = 2.682$, $d.f. = 23.320$, $p = .013$) where the NS group used higher rate of complex repetition compared to the NNS group. She claimed that the influence of Chinese mother-tongue and the students' vocabulary learning habits might be the main reason for using less complex repetition.

2.8.2 Signalling nouns in native and non-native writing

A number of studies analysed SNs in English writing of different non-native speakers of English. Nevertheless, studies that compare the use of SNs in the writing of Arab Speakers of English with native writing are less frequently dealt with than others, or there are no such comparative studies at all as far as I know. Table 6 summarises six studies that will be reviewed in this section.

Studies	Category of lexical cohesion	Focus	Learners L1 background	No. of NNS & NS texts/corpus size	Function of SNs	Method of analysis
Francis (1988)	Advance & retrospective labels	NNS-NS Comparison: a) frequency & types; b) appropriate use of labels; c) appropriate use of collocation with these type of nouns	Singaporean	45 NNS letters of complaint (student corpus); 45 NS letters of complaint (Press)	Across clauses (Anaphoric)	Manual counting
Flowerdew (2010)	SNs	NNS-NS Comparison: a) overall frequency; b) frequency of different functions; c) selection of specific SNs; d) range of SNs selected for use	Cantonese	217 NNS essays (111,558 words) (ICLE sub-corpus); & 93 NS essays (110,537 words) (American-LOCNESS)	Across clauses (cataphoric and Anaphoric); Within the clause	Manual counting of frequency of SNs & patterns with the aid of a word frequency counter

Forutan & Nasiri (2011)	SNs	NNS-NS Comparison: a) frequency & types	Persian	10 NNS texts (30,000 words); 10 NS texts (30,000 words);	Across Clauses (cataphoric and anaphoric); Within the Clause; Exophoric Function	Manual counting of frequency of SNs with the aid of a word frequency counter
Petch-Tyson (2000)	Anaphoric demonstrative expressions	NNS-NS Comparison: a) frequency & types of demonstratives; b) the use of demonstratives as organisational and developmental devices c) L1 influence	Dutch, French Finnish Swedish	50,000 words by NNSs (ICLE); 50,000 words by NSs (American English component-LOCNESS)	Across clauses (Anaphoric)	Manual counting of frequency with the aid of a concordancer
Aktas & Cortes (2008)	Shell nouns	NNS-NS Comparison: a) frequency of 35 shell nouns; b) different lexico-grammatical patterns; c) functions associated with these patterns	Chinese, Korean, Spanish, Turkish	28 NNS texts (66,459 words); 166 NS texts (721,553 words)	Across Clauses (cataphoric and anaphoric); Within the Clause	Manual counting of frequency with the aid of a concordancer
Hasselgård (2012)	Abstract nouns in recurrent word combinations	NNS-NS & NNS-NNS Comparisons: a) recurrent word combinations of 5 shell nouns; b) use the word combinations in appropriate contexts and with appropriate discourse functions	Norwegian French German	ICLE-NO (213,940 words) ICLE-FR (206,194 words) ICLE-GE (240,917 words); LOCNESS (326,089 words)	Across Clauses (cataphoric and anaphoric); Within the Clause	Cluster analysis

Table 6 Overview of research into SNs in NS and NNS English writing

Among the comparative studies that examined the use of SNs by non-native writers other than Arabs is a study by Francis (1988) who, in a small-scale empirical study, compared English writing of Singaporean students to the writing produced by the British and Australian press. She found that SNs (labelled by her as advance and retrospective labels) were less frequent in the Singaporean student writing compared to the press native writing. The student writing also contained a narrower range of this type of nouns, and less use of modifiers. Additionally, she observed that the students used retrospective labels inappropriately to package previous discourse. An example of this inappropriate use is when one student used *this human factor* to encapsulate an earlier complaint about the overcrowded conditions in a university canteen (Francis 1988: 333). She further noticed an inappropriate use of collocation between *intolerable* and *phenomenon* when a student used this label: *this intolerable phenomenon* (Francis 1988: 333). She highlighted that this example showed a case of unnatural collocation, which was not used by native writers. The lower use of SNs in terms of frequency and type by non-native speakers of English has been also confirmed by Flowerdew (2010: 43), who compared two corpora of argumentative writing which are the Cantonese ICLE sub-corpus and the American component of LOCNESS. Both corpora are part of the International Corpus of Learner English (ICLE). The learner sub-corpus contained 217 essays consisting of 111,558 words, and the average essay length was 514 words. The L1 English sub-corpus, on the other hand, was smaller than the learner corpus. It consisted of 110,537 words and 93 essays in the corpus, with an average length of 1,188 words. The essays in both corpora were written on different topics, but all have an argumentative structure. Both groups of writers were allowed to consult reference materials but the essays were written under different conditions. The learner essays were written under timed test conditions whereas native essays were not written under timed test conditions. Granger (2004: 126) contends that

“there are so many variables that influence learner output that one cannot realistically expect ready-made learner corpora to contain all the variables for which one may want to control”. Huang (2011) adds that compiling two comparable corpora while assuring the same method of data collection, genre and context for a particular use of language is a complex task. Due to such diversity in the variables, Huang suggests interpreting the frequency information cautiously and further examination of the data is required.

Flowerdew (2010) identified manually the SNs with their patterns in both corpora. He points out that this manual tagging is time-consuming but achievable because the corpora are small sized. This manual coding is important because SNs are textual relations whose semantic features require a human investigation compared to lexical relations such as simple repetition, for instance. Flowerdew then used automated corpus techniques, which is frequency lists to obtain frequency data on all the tagged patterns. Subsequently, Flowerdew (2010: 38) compared the use of SNs in the two corpora according to overall frequency, frequency of different functions (across-clause, cataphoric and anaphoric, and in-clause), selection of specific SNs and range of SNs selected for use. This combination of various features is imperative to identify all possible differences between NS and NNS in their use of SNs. Moreover, Flowerdew’s study did not rely simply on counting general frequencies; rather, he computed individual text-based frequencies of SNs. Granger et al. (2015) argue that despite the suitability of learner corpora for studying entire groups of learners and the benefits they have from a teaching perspective, the great variability between and within learners can lead to obtain inaccurate results if aggregate data is used alone (see Chapter 5 for more detail on aggregate frequencies and text-based frequencies). In line with Francis’s (1988) finding, Flowerdew’s (2010) study revealed that Cantonese-speaking learners of English used a

smaller number of SNs in terms of the overall frequency and the range of use than did their L1 peers. Also, the frequencies of anaphoric across-clause SN and in-clause SN in the native-speakers' corpus were over twice as high as those in the learner corpus. On the other hand, Cantonese-speaking learners of English used a higher rate of cataphoric across-clause SNs. However, a qualitative examination of the data showed greater idiosyncratic variations in the learners' use of SNs, particularly in cataphoric across-clause SN realisations. Such stereotypical use by learners of certain linguistic features was attainable when individual variation was considered.

Similar findings relating to the use of an insufficient amount of SNs by NNSs have been confirmed in another study, but the focus this time is on academic writing rather than argumentative writing. In this study, Forutan and Nasiri (2011) conducted a comparison of SNs that function across and within clauses in native English linguistic texts and non-native English linguistic texts that were written by Iranian writers. Their analysis showed that the frequency of SNs in the native English texts was higher than that in the Iranian non-native English texts. The frequency of SNs in the native English texts was 18 SNs per thousand words, whereas it was 15 SNs in the non-native English texts. These researchers examined SNs more qualitatively by looking at specific types of SNs used in both corpora. For example, their study demonstrated that the SN *way* occurred 50 times in the NS English texts, whereas only 20 times in the NNS English texts. On the contrary, the SN *question* was quite frequent in both corpora (30 times in the native English texts and 25 times in non-native English texts). However, these researchers did not highlight types of SNs that were merely distinctive to the writing of the non-native group; this would be useful because it might reflect a specific linguistic style in the interlanguage cohesion of these L2 writers. Forutan and Nasiri (2011)

interpreted the difference between NS and NNS writing in the use of SNs by claiming that lexical cohesion in English-speaking countries is consciously taught in writing classes such as paragraph writing. However, such writing principles are not taught at all in Iran; this as a result makes Iranian learners of English unaware of how lexical cohesion is used in English writing.

The quantitative observations sometimes tease out several possible lines of inquiry, which have to be examined qualitatively. For example, Petch-Tyson (2000) examined the use of anaphoric demonstrative expressions (demonstrative pronoun + noun) in NS and NNS writing. These expressions function as a ‘retrospective label’ (Francis 1994: 83) and behaves like SNs by having a textual function. Petch-Tyson (2000) analysed a 100,000-word sample of argumentative writing from the American English component of the Louvain Corpus of Native English Essays (LOCNESS) and from the Dutch, French, Finnish and Swedish sections of the ICLE. She found a frequent usage of anaphoric demonstrative expressions that have a signalling function among native writers, whereas non-native students preferred reference to other noun phrases. However, Petch-Tyson (2000: 54) found a large amount of demonstrative noun phrases in the French texts compared to other non-native sub-corpora. She observed that this high frequency, however, did not indicate a near-native use but non-native texts used demonstrative nominal anaphors differently from native writers. More specifically, non-native texts contained a number of demonstrative nominal anaphors that were a nominalised form of a verb that had been previously mentioned in discourse (*escape – this escaping*). Such use of demonstrative expressions that have a signalling function is not counted as a SN, as Benitz-Castro and Thompson (2015) suggest. That is because the main function of SNs is to encapsulate previous stretch of discourse in a new label that has not been

repeated in the text before (see Chapter 6 for more detail on using SNs as nominalised forms). Petch-Tyson (2000) highlighted that frequency data should not be taken at face value but have to be re-tested through more robust qualitative analysis.

Similarly, Aktas & Cortes (2008) analysed shell nouns in native and non-native writing. Their study investigated shell nouns in two corpora, one of published research articles (166 texts, 721,553 words) and another of research articles by non-native MA and PhD students (28 texts, 66,459 words). These researchers addressed six shell nouns, which were the most frequent shell nouns in the published writing corpus. These nouns included *effect*, *result*, *fact*, *system*, *process* and *problem*. Their findings showed that the student corpus contained more of these shell nouns than professional writing did. By conducting a qualitative analysis, they noticed that the high frequency of some of the nouns in student corpus was only by one or two writers, which could be the reason for increasing the frequency of shell nouns in the student corpus. Aktas & Cortes (2008: 10) found further differences between the two groups in the type of the functional patterns which contain these nouns. For example, the shell noun *fact* occurred very frequently in the N-clause pattern with cataphoric function (i.e. *the fact that*) in the corpus of published writing, and in the anaphoric demonstrative-N pattern (i.e. *this fact is*) in the student corpus. Such a difference in the use of functional patterns of SNs has been further studied by Hasselgård (2012) who examined the distribution of recurring patterns of five shell nouns (*fact*, *idea*, *question*, *problem* and *issue*) in native (LOCNESS corpus) and non-native argumentative writing (ICLE). She found that the French texts in the non-native group shared common features with the native standards compared to the other non-native corpora (Norwegian and German). Such features as Hasselgård (2012: 34, 43) indicated, include the similar distribution of the N-of and N-that patterns for *idea* and *problem* in the

LOCNESS and French-ICLE texts. With a closer qualitative analysis, Hasselgård (2012: 51) observed that this resemblance was due to the high frequency of N-of and N-that patterns in French (e.g. *l'idée de/que*). The existence of these English patterns in French resulted in L1 positive transfer leading to correct language production.

In summary, the research studies on SNs described so far provide valuable insights into the study of SNs quantitatively and qualitatively. Nevertheless, none of them shed light on the use of SNs in the English written products by Arab speakers of English as compared to English native writing. This area of research is worth studying because English and Arabic belong to completely different linguistic systems, and this might reveal interesting results about the inter-language cohesion of Arab speakers of English. It is therefore one of the aims of this research study to address this gap.

2.9 Applying classic models of lexical cohesion to text analysis

Classic models of lexical cohesion, such as Halliday and Hasan's (1976) model and Hoey's (1991) model, provide specimen analyses, whereas the present study is concerned with analysing lexical cohesion in complete essays by NNSs and NSs. Section 2.8 showed how different studies on cohesion in NNS writing applied the traditional models of lexical cohesion to learner writing. However, these studies did not identify any issues that may arise in applying these models to analyse lexical cohesion in NS and NNS writing. Therefore, it is central to examine the general principles of analysis that underlie classic models in order to evaluate their applicability to text analysis. This evaluation will help understand how lexical cohesion is described and measured, and will further enable to suggest a systematic and rigorous approach to lexical cohesion. This will also allow us to consider the value of other

approaches to analysing cohesion for the present study. This section will firstly focus on the notion of a tie and directionality of lexical cohesion within the framework of classic models (cf. Section 2.9.1). Section 2.9.2 will discuss the boundary of the lexical item. Next, Section 2.9.3 will highlight the issue of a cross-categorisation of lexical cohesion. This will be followed by Section 2.9.4, which highlights the different types of lexical cohesion that are selected from the open-set words, and it further comments on the frequency factor in the analysis of cohesion. These sections illustrate with examples the different methodological principles that classic models apply to analyse lexical cohesion.

2.9.1 The notion of a tie and directionality of lexical cohesion

The typical property of cohesion is the connectedness of lexical items. The traditional models of lexical cohesion describe this feature differently. The key notion that is applied in analysing cohesion of a text in Halliday and Hasan's (1976) model of cohesion is that of a 'tie', already discussed in Section 2.2. Their model identifies that a tie is a relation between two items, the second of which depends on the first one for its interpretation. This indicates that a tie, as Halliday and Hasan (1976: 329) describe, is a "relational concept" and also "directional". The direction of a tie can be 'anaphoric', which means that the cohesive item refers back to another item in a text, or cataphoric, which indicates that the cohesive item refers forward to another item for its meaning. Halliday and Hasan (1976) observe that the typical direction is the anaphoric one whereas cataphoric ties are relatively infrequent.

The resolution of the anaphoric tie according to Halliday and Hasan's (1976) principles of analysis depends on the distance, which separates the presupposing item from the presupposed. For example, in their analysis of a text for cohesion, when cohesion is lexical and the same lexical item occurs twice or more, then a tie takes its interpretation from the

nearest preceding lexical item. This means that the lexical item does not establish a cohesive relationship with all presupposed items which relate to it but with the nearest item that resolves its meaning. Although this principle of analysis is not explicitly described by Halliday and Hasan (1976), the examples they provided in Chapter 8 show that. For example, in the sonnet *The bad thing* by John Wain that Halliday and Hasan (1976) analysed (Example 31 below), there is a phrase *the bad thing* that appears five times. Halliday and Hasan (1976) relate the final instance of this phrase (line 11) to that which immediately precedes it (line 10). They do not link it back to lines (8), (5) and (2). Example (31) illustrates such a principle. Only lines that contain *the bad thing* are extracted from the sonnet; therefore, the numbers of lines are not in sequence but are kept in the same order that is presented by Halliday and Hasan (1976). Also, lexical cohesive devices will be italicised in all examples illustrated below.

- (31) Sometimes just being alone seems the *bad thing* (2).
You think; this is the *bad thing*, it is here (5).
Then you think: the *bad thing* inhabits yourself (8).
Escape, into poem, into pub, wanting a friend
Is not avoiding the *bad thing* (10). The high shelf
Where you stacked the *bad thing*, hoping for calm, broke (11).

(Wain, cited in Halliday & Hasan 1976: 344)

From Example (31), it seems that Halliday and Hasan (1976) view lexical cohesion as a local phenomenon between words in a sentence and the preceding one. They prioritise short-distance lexical cohesive relationships over long-distance ones. Halliday and Hasan (1976: 290) point out that “if [two lexical items] occur in adjacent sentences, they exert a very strong cohesive force; this would be progressively weaker the greater the textual distance between

them”. Nevertheless, they also observe that a tie is not always resolved in the immediate sentence that preceding. The tie sometimes spans large numbers of intervening sentences until it is resolved. In connected texts, applying this theoretical concept of a tie makes it not simple to move around the text and keep back-tracking to find the proper ties. Such a difficulty arises because the analyst has to remember a substantial amount of the text within their active memory to be able to spot lexical ties and relate them to what comes after until they process the whole text. This, as Fligelstone (1992) describes, means that the mental resources of the analyst are not free to concentrate on the main target of the analysis which is the sense and the cohesion of the text.

Halliday and Hasan (1976)’s analysis of their sample texts also refers implicitly to how cohesive ties are counted. Example (32) below illustrates that sentence 2 includes two lexical items (*door ... door*) which repeat another item *doors* in the preceding sentence (sentence 1). So, in such a case, does Halliday and Hasan (1976)’s method of analysis count one tie or two ties? Halliday and Hasan (1976) do not make it clear how they count ties in their discussion of the principles of analysis. However, Example (32) demonstrates that when two lexical items occur in the same sentence, and they repeat an earlier item, Halliday and Hasan (1976)’s method of analysis counts two cohesive ties. Accordingly, in sentence 2, the two lexical words *door* and *door* enter into a lexical cohesive relationship with *doors* in sentence 1, and hence two cohesive ties are identified. This principle of analysis is illustrated below.

- (32) ...It was about thirty years old, on stone pillars, with a long stone staircase up and folding *doors* back on to a verandah (1). And I came through the *door* from the kitchen, and a thief carrying my handbag emerged through my bedroom *door* into the living room at the same moment (2). (Halliday& Hasan 1976: 341)

In presenting a model for the analysis of lexical cohesion, Hoey (1991b: 76) stresses that the method of analysis should be replicable “so that others can follow the same path”. Hoey (1991b), therefore, emphasises the need to spell out precisely what should count as repetition and the way repetition organises texts. Hoey (1991b: 52), unlike Halliday and Hasan (1976), prefers to describe the connection between lexical items as a ‘link’ not a ‘tie’ claiming that “a tie implies directionality more than link”. He suggests that links are formed in other than a chaining manner. He defines a link without regard to its distance or exact direction. As Hoey (1991b: 72) puts it, “[t]he direction of linkage is unimportant”. Sardinha (2001) explains that the concept of a link that is suggested by Hoey (1991b) indicates multi-directionality, which leads to the formation of a network of lexical relationships between lexical items rather than creating a string-like fashion of ties. Hoey (1991b) describes that this multi-directionality accepts items that spread over long distance to be included into a cohesive relationship. This is not compatible with what Halliday and Hasan (1976) apply in their model, as illustrated in Example (31) above. Hoey (1991b: 83) identifies that each lexical item which is part of a cohesive relationship is allowed to form a lexical relationship not only with its immediate predecessor in the text, but with all previous occurrences in earlier sentences which relate to it. This creates a network of links in a text. Hoey (1991b) observes that these earlier instances provide much of the textual context necessary to interpret the final occurrence. His model conceptualises lexical cohesion as “the only type of cohesion that regularly forms multiple relationships” (Hoey 1991b: 10).

Figure 5 illustrates how repetition links are identified according to Hoey’s (1991b) principle of analysis. If there is a lexical item, say, in sentence 16 of a text and this item repeats another item in sentence 14 and more other items in sentences 1, 7, 10 and 12 as well. Therefore, in

such a case, the lexical item that occurs in sentence 16 will enter into multiple lexical cohesive links with all its predecessors in sentences 1, 7, 10, 12 and 14. Similarly, each previous instance of the lexical item repeats its predecessors. Consequently, a network, or net of links is created; see Figure 5:

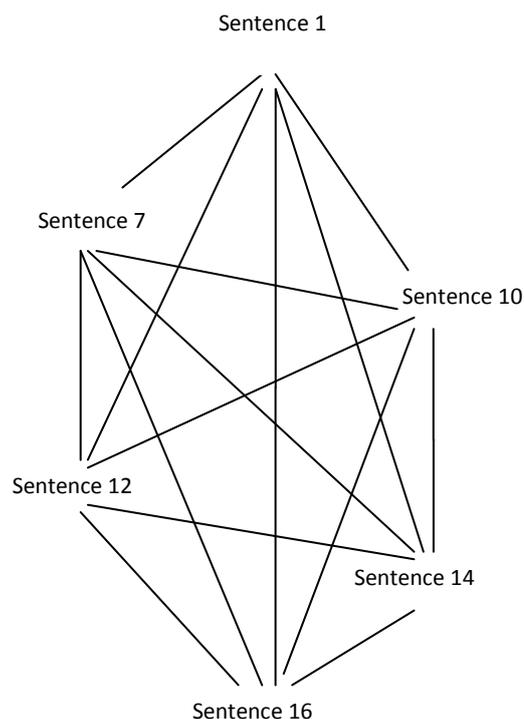


Figure 5 Hoey's (1991b) identification of a network of repetition links between lexical items across sentences

(adapted from Hoey 1991b: 81)

Although Hoey (1991b) attempts to develop Halliday and Hasan's (1976) method of analysis, his concept of a link and how he applies it to a text is still complicated. Figure 5 indicates this complexity. The figure shows the links between sentences formed by one lexical item. So, what kind of diagram would be needed to reflect the links made by all lexical items in a long text?

With regard to calculating the number of repetition links between a pair of sentences, Hoey's (1991b: 36) method of analysis suggests that "an item in one sentence that connects with two in another will be deemed to make only one connection not two". Example (33) illustrates this criterion of analysis which shows that despite the existence of two lexical items of *bear* and *bears* in sentence 5, *bears* in sentence 1 links with only one instance of *bears*. As a result, only one link is counted. (Sentences 1 and 5 are an extract from a short text presented by Hoey (1991b), and as the example shows the sentences are not adjacent).

(33) 1 A drug known to produce violent reactions in humans has been used for sedating grizzly *bears* *Ursus arctos* in Montana, USA, according to a report in The New York Times.

5 Although some biologists deny that the mind-altering drug was responsible for uncharacteristic behaviour of this particular *bear*, no research has been done into the effects of giving grizzly *bears* or other mammals repeated doses of phencyclidine.

(Hoey 1991b: 37)

The method that Hoey (1991b) adopts in Example (33) to count the number of links between sentences is different from the method that is suggested by Halliday and Hasan (1976), as illustrated in Example (32). This indicates that every approach to lexical cohesion is applied to a text differently which might be related to the fuzzy nature of the concept of cohesion itself. Fligelstone (1992: 162) comments that "[n]o serious attempt is made to deal with what Halliday and Hasan refer to as 'lexical cohesion'". This is because, as Fligelstone (1992) explains, the application of backward dependency of cohesive ties to the analysis of lexical cohesion makes it difficult to achieve consistency among analysts, or even by a single analyst.

These aspects of texts did not only cause inconsistency, but also “a considerable loss of time through ‘agonising’ on the part of analysts” Fligelstone (1992: 161).

Overall, both theoretical concepts of a tie or a link are complex notions, and their application to text analysis might pose a challenge to analysts. This argument is especially true if the text is long, or if the purpose of analysis is to compare and count instances of cohesive types in two data sets, as it is the case in the present study. In this regard, Sinclair (2004: 13) asks, “[do] we actually need all the linguistic detail of backward reference that we find in text description?” On the one hand, he mentions that drawing links from one element of the text to another in written text is common because the written text is accessible to readers who can examine it, and rely on retrospective reference for its interpretation. On the other hand, Sinclair (2004) suggests an alternative approach to describing discourse by claiming that:

The most important thing is what is happening in the current sentence, and the meaning of any word is got from the state of the discourse and not from where it came from. [...] The state of the discourse is identified with the sentence which is currently being processed. No other sentence is presumed to be available. The previous text is part of the immediately previous experience of the reader or listener. [...] It will normally have lost the features which were used to recognize the meaning and to shape the text into a unique communicative instrument.

(Sinclair 2004: 13)

From this perspective, Sinclair (2004) advises that there is no advantage to be gained in tracing the references back in the text. Chapter 3 will introduce the theoretical concept of a ‘lexical item’ suggested by Sinclair (2004), and will provide a further detail on how lexical cohesion is viewed in the light of corpus linguistics.

2.9.2 The boundary of the lexical item

A methodological problem which is concerned with identifying repetition ties or links relates to how a lexical item is defined by each researcher. Lieber (1979: 139) points out that “repetition may be of an item, a phrase or a bigger textual structure”. The concept of a ‘tie’ overlooks the fact that some categories of lexical cohesion create vagueness and their borders are not definite, as Carter and McCarthy (1997) observe. For example, Carter and McCarthy (1997) regard general words (e.g., *thing*, *stuff*) as a subcategory of vague language. They explain that this category of nouns is very frequent in speech and “enable a speaker to express attitudes and feelings without needing to locate an exact or precise referent” (Carter and McCarthy 1997: 16). This confirms that the meaning of a lexical item does not sometimes reside in one item but it spreads over a long stretch of discourse. This feature of vagueness that is associated with a number of lexical items is clear evidence which goes against the classic models of lexical cohesion that claim that cohesion could be identified clearly through referring back to precise single items.

Although Halliday and Hasan (1976: 292) admit that “the concept of a ‘lexical item’ is not totally clear-cut and it presents much uncertainty in application to actual instances”, they do not provide a working definition for the term. They also do not describe clear criteria for deciding how far a lexical item can be extended. However, following their sample analysis in Chapter 8, it seems that Halliday and Hasan (1976) include single words and phrases. Examples of these phrases that Halliday and Hasan (1976) provide in their analysis include: *carry knife ... carry knives* (Halliday & Hasan 1976: 342); *public men ... public man* (Halliday & Hasan 1976: 348); *fields of rice ... fields of rice* (Halliday & Hasan 1976: 346); *not far out of ... about three miles* (Halliday & Hasan 1976: 351). All these examples illustrate two or

more word phrases that represent reiteration of same item in their text. Halliday and Hasan's (1976) method of analysis for cohesion treats these phrases as comprising a single lexical item instead of two separate lexical items. Example (34) illustrates how a cohesive tie between *public men* (sentence 3) and *public man* (sentence 2) is formed.

(34) Birling: No, for being so offensive about it (1). I'm a *public man* (2).

Inspector [massively]: *Public men*, Mr Birling, have responsibilities as well as privileges (3).

(Halliday & Hasan 1976: 348)

In Example (34), Halliday and Hasan's (1976) method of analysis treats *public men* in sentence (3) a single item, not two items. As a result, their method counts one cohesive tie of same item repetition between *public men* (sentence 3) and *public man* (sentence 2). Hoey (1991b: 271) agrees with Halliday and Hasan (1976) that identifying the boundaries of a lexical item has no precise answer, because these boundaries are far from straightforward. However, unlike Halliday and Hasan (1976), Hoey's (1991b) strategy is to treat the orthographic boundary as strong grounds for assuming a lexical boundary. For example, he treats the phrase *grizzly bears* as two items when he marks the repetition link of this phrase with the same phrase in a later sentence in a short text of factual reporting named *Drug-Crazed Grizzlies*. Károly (2002) criticises Hoey (1991b) for dealing with isolated occurrences of lexical repetition patterns. She insists on the role of the context because the analytical tool that Hoey (1991b) himself designs is to examine how repetition pervades within text and not isolated, context-independent sentences. This might indicate a gap and inconsistency between

the theory of text organisation that Hoey (1991b) explains, and the application of it in practice when his data of repetition links are viewed.

Even when researchers extend the lexical item to include phrases and longer expressions, they tend not to look at a lexical item as a unit of meaning as it is looked at in corpus linguistics. Chapter 3 will revisit the lexical item and define it in the light of corpus linguistics. The corpus linguistic definition will then be used in the present study to analyse lexical cohesion, but only in some part (see Chapter 7).

2.9.3 The cross-categorisation of lexical cohesion

The process of simultaneously notating different categories of lexical cohesion within the same text is not entirely straightforward and raises important questions for counting. A phenomenon such as this has not been covered adequately by previous work on cohesion. The categories of lexical cohesion intersect in terms of their functional features when connections between lexical cohesive items are formed across units of text. This cross-categorisation is unavoidable in any natural linguistic system. As demonstrated in Section 2.7.1, signalling nouns, for example, overlap in some cases with simple repetition. Example (35), which is taken from the ALEC corpus, illustrates this case.

- (35) **1** For example, when I was doing my Bachelor degree at the University of Derby, I had to write up various technical reports about some complex mechanical *problems*.
// 2 Instead of trying to solve these *problems* using the power of my own brains, I instead relied on my computer and the access to the internet to identify projects that are relevant or closely related to the ones I was working on. (ALEC)

In T-unit 2 in the example above, the lexical item *problems* repeats another item *problems* in T-unit 1. However, *problems* in T-unit 2 could also function as a signalling noun in that it encapsulates the clause underlined in T-unit 1: *had to write up various technical reports about some complex mechanical problems*. In such a case, *problems* in T-unit 2 could either form simple repetition with *problems* in T-unit 1, or could function as an SN. Thus, the methodological question here is: In the analysis of cohesion, do we need to mark one cohesive function only of the lexical item *problems* – or do we need to register both functions of lexical cohesion?

In this respect, Flowerdew and Forest (2015) observe this issue of the overlap between signalling nouns and repetition. Section 2.7.1 showed that Flowerdew and Forest's (2015) method of analysis is based on putting restrictions to count the frequency of SNs when they are repeated in the same text. They propose that both cohesive functions whether an SN or repetition are possible but they prefer to control the number of cohesive functions that the item has for methodological reasons only. Halliday and Hasan (1976) accept all cohesive functions a lexical item has. Although this principle is not described explicitly in their final chapter of analysis, their analysis suggests that when a lexical item has more than one cohesive function; both functions that the item carries out are identified. Example (36) is a short extract from a dialogue that Halliday and Hasan (1976) provided in their sample texts in Chapter 8.

- (36) Birling [angrily, to Inspector]: look here, I'm not going to have this,
Inspector (1). You'll *apologize* at one (2).
Inspector: *Apologize for what* – doing my duty (3)?

(Halliday & Hasan 1976: 348)

The phrase *apologize for what* in sentence 3 creates a cohesive function of the type Ellipsis with the word *apologize* in sentence 2 (no attempt is made here to define Ellipsis because this category is beyond the scope of the present study). At the same time, *apologize* in sentence 3 forms a cohesive function of same item repetition with *apologize* in sentence 2. This means that Halliday and Hasan (1976) count two cohesive ties in this example, but they did not give attention to this issue when cohesive ties have multi-functions in a text.

In contrast, Hoey's (1991b) method for the analysis of lexical cohesion identifies that when a lexical item enters into a lexical cohesive relationship with more than one item in a previous sentence, one cohesive function only is recorded. Hoey (1991b: 83) provides a hypothetical example to illustrate this principle. He describes that if, for instance, there are two sentences; the first one contains a lexical item *writer*, while the second one includes a complex repetition *writings*, a simple paraphrase *author*, and a pronoun *he*. In this case, although the second sentence contains three separate items, they would be dealt with as creating only one type of link. But which type of link out of the three that Hoey's (1991b) method of analysis does record? In such circumstances, Hoey's (1991b: 83) general principle is to give a high placing for simple repetition to be recorded, followed by complex repetition, and then other cohesive types. Hoey (1991b: 83) assumes that repetition links are different in terms of their importance in a text. He, therefore, orders these links by starting with simple repetition as the most important link and ending with ellipsis as the least important link:

1. Simple lexical repetition
2. Complex lexical repetition
3. Simple mutual paraphrase
4. Simple partial paraphrase

5. Antonymous complex paraphrase
6. Other complex paraphrase
7. Substitution
8. Co-reference
9. Ellipsis

The four first categories of lexical cohesion have already been defined in this chapter (cf. Sections 2.4; 2.6). Other types of lexical cohesion are only listed as they are irrelevant to the purpose of the present study. Hoey (1991b) does not provide a theoretical basis for this hierarchy, which gives priority to lexical links over grammatical links. Hoey (1991b: 83) points out that “the arguments for a low placing are in fact practical, rather than theoretical”. So, in Hoey’s (1991b) hypothetical example, his criterion of analysis marks one cohesive function between *writer* and *writings*, which is complex repetition. Other types of links are considered by Hoey as internal links to the sentence and are not recorded.

In a similar vein, Garside et al. (1997: 2) comments that there is no an absolute and objective method that can be used to determine what label or labels need be assigned to a certain linguistic feature. There are many words which would be more contentious and portmanteau than other words. For example, the word *computer* in *computer technology*: Does it function as a noun or an adjective? This question might seem irrelevant here to the study of lexical cohesion and more pertinent to grammatical tagging. However, since one of the categories of lexical cohesion in the present study is derived repetition, the answer to this question becomes important. To put it more clearly, words that occupy a pre-nominal position such as *computer* in *computer technology* could function as a noun, and in this case it will form simple

repetition with their repeated nouns (e.g., *computer* (n)) in previous sentences. Simultaneously, such a type of words might function as derived repetition. That is, *computer* in *computer technology* can function as an adjective, and will, as a result, form derived repetition with the noun *computer* if it occurs in earlier mentions across the text. So, the category of derived repetition is formed between *computer* as an adjective in (e.g., *computer technology*), and the noun *computer/s* (n)) if it is repeated throughout the text. This latter case of repetition is called ‘zero derived repetition’, and represents one type of derived repetition as has already been referred to in Section 2.6.2.

To sum up, the discussion above on the cross-categorisation of lexical cohesion implies that there is no right or wrong in deciding whether the method of analysis records only one function or all the functions that the lexical cohesive item creates. However, what matters is that the analyst has to base their annotation scheme on well-defined principles or rules that allow other analysts for replication.

2.9.4 The selection of open-set words and the frequency factor

Setting up selection criteria is a practical solution in any text analysis where we are faced with two options – either we consider the text as a bag of words in which we need to analyse the whole set of the items available, or whether we need to design selection criteria in order to make the analysis more focused. Such a need of exclusion criteria in analysing cohesion does not mean that the excluded items do not participate in cohesion but this is only a practical analytical decision. In other words, it is a matter of which kind of lexical cohesion every researcher decides to analyse. The focus of the present study is on lexical cohesion. Accordingly, only open-class words (i.e., content words) will be examined in the current study. Content words include nouns, adjectives, verbs and adverbs. But what should be

selected from this range of open-set words? For example, will nouns only be marked in the analysis for cohesion? Besides, what about lexical words that have weak denotations and they are barely referring to anything, such as general verbs? Furthermore, are high-frequency words good candidates for studying?

Researchers who work on cohesion have different views regarding the potential words that could be candidates of lexical cohesive relationships. Some researchers such as Morris and Hirst (2006) and Stührenberg et al. (2007) consider only nouns and noun compounds for membership in lexical cohesion, and they assume that only nouns can participate in cohesion. However, other researchers include other content words such as adjectives and verbs, arguing that adjectives and verbs could also participate in lexical chains even though they are less common than common nouns. This confirms what Palmer et al. (2012) arrive at that nouns exhibit more cohesion with their context or at least participate in more easily identifiable cohesive relationships than other word classes such as adjectives and verbs.

However, Halliday and Hasan (1976: 288) maintain that “every lexical item may enter into a cohesive relation”, which is only possible to be formed by reference to the text (the underlining is added by me for emphasis). Their view turns every lexical item into a possibility of inclusion in cohesion, which in turn leads us to consider the complete set of lexical words in the text. They add, though, that the cohesive force between two lexical items in a text is affected by their overall frequency in the language system. They claim that the frequency factor controls the number of ties that need to be recorded, positing that the high frequency of a lexical item reduces its possibility of creating lexical cohesion in texts. They illustrate their claim with this list of general words: e.g., *go, man, know, way, take, do* and

good (Halliday and Hasan (1976: 290-291). They posit that these words can hardly be said to contract significant cohesive relations, because they go with anything at all. Although the general word *man* that has been mentioned in the list is one of the general nouns that Halliday and Hasan (1976) classify under ‘human noun’, they make it clear that not all general words are used cohesively. One condition for a general word to be cohesive is to share the same referent with the other word that it precedes or follows. Therefore, they suggest that in identifying the lexical cohesion in a text, repeated instances of lexical items whose frequency is very high can be discarded. However, Halliday and Hasan (1976) claim that when words of high frequency carry particular meanings with restricted patterns of collocation, significant cohesion could be established; for example *takings* in the meaning of *earnings*, or *good* in a moral context when it means *virtue*. Gutwinski (1976) supports Halliday and Hasan’s (1976) claim and agrees that high-frequency items such as *get*, *put*, or *say* will need to be eliminated from the analysis if they are not strengthened by other cohesive factors. He argues, on the other hand, that low-frequency items such as *ice-rink*, *excavate*, *prisoner* and *hermit* will be considered as cohesive, if repeated (Gutwinski 1976: 81).

In contrast to the above argument about the influence of frequency on the strength of cohesion in text, there are a number of researchers who consider frequency an important factor in their selection criteria for cohesive types. Mahlberg (2005: 161), for example, analyses ‘general nouns’, which she defines as “nouns which are frequent”. She suggests that frequency is part of the meaning of the nouns. She provides an example of *time*, which is the most frequent noun in the Bank of English and the British National Corpus. Mahlberg observes that despite the high frequency of *time* nouns, they contribute to the continuity in texts. In her discussion of the *time* orientation group, for example, it becomes clear how *time* nouns link to other text

elements. Likewise, Flowerdew analyses signalling nouns which represent a sub-class of general nouns. He stresses the significant role that these nouns play in creating cohesion in text despite the fact that these nouns are very frequent. If we look now at this issue in Hoey's (1991b) study, from his analysis it seems that Hoey (1991b) includes a different range of open-set lexical words (nouns, adjectives, verbs and adverbs) regardless of their lexical class or frequency. The crucial point in the analysis, as Hoey (1991b) points out, is that these words have to establish three-link connections between sentences. In his analysis of a short passage, for instance, Hoey (1991b) takes account of lexical items of very high frequency. The item *use*, for example, creates a complex repetition with *user* even though *use* is a very high frequent item in the language.

In a different study by Hoey (2005), he claims that the cohesive behaviour of lexical items in a text depends on whether or not these items are primed for cohesion. By 'priming', Hoey (2005: 8) means, "as a word (lexical item) is acquired through encounters with it in speech and writing, it becomes cumulatively loaded with the contexts and co-texts in which it is encountered". This mental accumulation of the lexical item encounters represents part of our knowledge of the item so that all kinds of patterns, including collocational patterns, are accessible and available for use in certain types of context. With reference to cohesion, Hoey (2005) claims that there are particular lexical items that are primed to establish cohesive relationships in a particular text, whereas other items are primed to avoid such relationships. This cohesive behaviour of some lexical items is because "many of the characteristics of the text are latent in the lexical items from the moment of their selection." (Hoey 2004: 401). Hoey (2005) provides a list, which encompasses a range of items that are primed for cohesion. This list includes common nouns (e.g., *army*, *planet*), names (e.g., *Blair*) and descriptive

adjectives (e.g., *political*). He remarks that these items are non-evaluative and with clear denotations. Nevertheless, Hoey (2005) argues that it is not always the case that non-evaluative and readily-defined items are so primed. For example, lexical items such as *thirty* and *hour* are not primed for cohesion even though they are non-evaluative and have clear denotations. Furthermore, he also observes that sometimes there are few items that are evaluative or have weak denotations (e.g., *ridiculous*, *racist*, *make*, *action*), but they are also primed for cohesion when they are examined in their texts. If we examine what the above list includes, we find that it contains evaluative adjectives *ridiculous*, *racist*, weak denotation verbs *make*, and abstract nouns *action*. Hoey (2005) provides an additional list of words – *asinine*, *blink*, *crossroads*, *elusive*, *particularly* and *wobble* – which he posits are no less frequent than the items mentioned in the previous lists above at least in his travel writing corpus. However, these words, as Hoey (2005) affirms, appear to avoid participating in cohesion. Hoey (2005: 120) concludes that “one might have predicted that a word’s infrequency in the language would make it less available for participation in cohesion, but frequency does not seem to be an important factor”. This observation may be attributed to the claim that cohesive properties of the item are already built into the item itself. Therefore, frequency might not be an influencing factor on the cohesive force of the lexical item, as Halliday and Hasan (1976) remark.

We could infer from the aforementioned discussion that every lexical item that belongs to an open-set class is a possible type of lexical cohesion but with variable strength – whether these items are descriptive or evaluative adjectives, common, proper or abstract nouns, or items with weak denotations such as light verbs (e.g., *make*), for instance. The selection of any of these types of lexical cohesion depends on every researcher’s purposes. Besides, the argument

which shows the relationship of lexical cohesion with frequency shapes fluid situation. If we decide, for example, to exclude lexical items of very high frequency from our analysis, we might end up with few items to count, particularly that we are dealing with learner corpora in the present study, which means that a large proportion of students' writing contains high frequency items. Thus, if we rule out all high frequent verbs and general nouns, for example, such as *make, use, people, problem...etc.*, we will miss out a lot of linguistic features that represent one of the characteristics of non-native speakers' writing. If we, however, keep all lexical items irrespective of their frequency, the analysis will become unwieldy. Thus, it seems plausible to take into account Halliday and Hasan's (1976) view that frequency controls the strength or weakness of cohesive relationships between lexical items. However, such a treatment, as Halliday and Hasan's (1976) acknowledge, demands a separate study for full interpretation of lexical cohesion relations and the different degrees of their strength. Chapter 4, Section 4.8 will suggest inclusion and exclusion criteria that will be applied in the actual analysis.

2.10 Conclusion

This chapter reviewed the key text-linguistic models of lexical cohesion which differ from each other in their classification of lexical cohesion and in the way they analyse it. The chapter also has introduced the working definitions of simple repetition, derived repetition and signalling nouns, which are the main focus of the present study. A number of studies on lexical cohesion in NNS writing used the traditional models of lexical cohesion to analyse these categories in NS and NNS writing. These studies showed that frequency counts of lexical cohesion are not a necessary indicator of coherence of a text, and demonstrated few if any significant differences between NS and NNS writing. These studies, therefore, interpreted

their quantitative results qualitatively. On the one hand, their analysis was useful in showing the NNSs' tendency of using a small range of lexical cohesive devices which made their writing redundant and less informative compared to NSs who preferred a lexical variety. However, the qualitative analysis in these studies was based on a little number of individual examples that were not sufficient to show how lexical cohesion with its different categories functions in learner writing to create continuity across the text. Also, these studies did not present clear analytical criteria describing how cohesive forms were identified and counted within the framework of the traditional approaches. Therefore, this chapter highlighted some of the issues and challenges involved in the application of classic models of lexical cohesion to text analysis. It seems that these challenges are fundamental obstacles that make the analysis of lexical cohesion not straightforward or easy whether it is performed manually or using a computer.

The central problem with the classic models of cohesion is that lexical cohesion is directional and relational. This retrospective character (looking back) of cohesion requires the reader to look back at an earlier item or items. Applying this theoretical concept of either a tie or a link to text analysis takes into account only the tenuous connection between isolated constituents of sentences and takes it for granted that the lexical unit has clear-cut boundaries. However, patterns of language are more complex and their meanings overlap. Sinclair (2004) suggests that a lexical item does not need to be retrieved from a previous text to have meaning. One approach is to look at cohesion from corpus linguistic perspectives. There are not yet many studies that apply or extend corpus linguistic concepts to a realm that has traditionally been occupied by text linguistics. The next chapter introduces different studies that analyse lexical cohesion in a corpus linguistic context.

Chapter 3

Lexical cohesion and corpus linguistics

3.1 Introduction

Lexical cohesion is fairly simple to compute by relying on surface cues and statistics rather than on deep linguistic processing (Palmer et al. 2012). Hoey (1991b: 12) also maintains that lexical repetition is suitable for objective analysis, and can therefore be counted and identified automatically. In addition, Morris and Hirst (1991: 23, 29) used chains of word relations (such as synonymy, antonymy and part-whole) taken from a thesaurus as a means to identify thematic units in texts. They applied a manual method that could “delineate portions of text that have a strong unity of meaning,” but claimed that the method could be computed. However, all these researchers highlighted the technical aspect of lexical cohesion analysis, which is only part of the problem. The theoretical part of the analysis aims to find an analytical unit that can describe cohesion qualitatively. As mentioned in the previous chapter, classic models of lexical cohesion do not take into account the extended meanings of lexical cohesion, whereas a corpus linguistic approach to lexical cohesion emphasises that meaning is spread over many items, which collectively provide extended text cohesion. Scott (1997) points out that a corpus-linguistic research on how texts are organised is very little. This chapter first provides a definition of the lexical item from a corpus-linguistic viewpoint and how researchers use it as an analytical unit to the analysis of lexical cohesion (cf. Section 3.2). Then, the subsequent sections review a number of studies that have addressed lexical cohesion in non-learner corpora (for example, newspapers; see Section 3.3), and studies that are concerned with lexical cohesion analysis in learner corpora (see Section 3.4). This will be followed by a conclusion.

3.2 The definition of a lexical item according to corpus linguistics

As discussed in Chapter 2, lexical cohesive relations according to the traditional text-linguistic models of lexical cohesion are tied to orthographic word boundaries in that each meaning of a lexical item is associated with one distinctive meaning of another item. This indicates that these approaches distinguish between form and meaning and do not analyse a lexical item as an extended unit of meaning, but they still emphasise the inter-clausal nature of lexical cohesion, as Mahlberg (2009) highlights. On the other hand, a corpus linguistic approach puts more emphasis on combining lexis and grammar in the description of language. A number of researchers have developed corpus-linguistic concepts, which can present evidence for the lexico-grammatical patterning. For example, Hunston and Francis (2000) suggest the ‘pattern’ concept, whereas Sinclair (2004) introduces his concept of the ‘lexical item’. These concepts, as Mastropierro and Mahlberg (2017: 80) claim, “have implications for the description of cohesion”. They clarify that when language is described as patterns; the lexico-grammatical connections in single texts can be explained in terms of a ‘pattern flow’. This pattern flow occurs “whenever a word that occurs as part of the pattern of another word has a pattern of its own” (Hunston and Francis 2000: 121); thus, creating linear relationships that connect each pattern with the one that follows. Likewise, the concept of the lexical item is useful for the description of cohesive relationships.

McEnery et al. (2006: 147) explain that the corpus linguistic definition of a ‘word’ is usually “a sequence of characters bounded by spaces”. This definition of a word is based on the idea of how words are displayed in a wordlist, which is a list of all words in a given text. However, McEnery et al. (2006) point out that this definition constitutes only a starting point for a corpus linguistic analysis, because it will allow us later to proceed in a reasonably objective

manner. This implies that these word-forms (in the wordlist) will then be co-selected and joined together; for example, by using a concordance analysis, we can observe the textual behaviour of the word-form by examining a number of examples of this word together with some context from the original text. Sinclair's (2004: 122) notion of a lexical item is "characteristically phrasal although it can be realized in a single word". He stresses that a lexical item provides a lexical description of items in texts.

Sinclair's (2004) lexical item is made up of five categories of co-selection. He claims that two of these categories are obligatory, which are the 'core' and the 'semantic prosody', whereas the other three categories are optional, which are 'collocation', 'colligation' and 'semantic preference'. Sinclair (2004: 35-36) illustrated these categories with the lexical item *true feelings*, which was found in the Bank of English. He identified that *true feelings* is the core, which is invariable and shows the evidence of the occurrence of the item as a whole. The core *true feelings* is itself a collocation. Sinclair (2004: 141) defines collocation as "the co-occurrence of words with no more than four intervening words". In terms of colligation, *true feelings* colligates with a possessive determiner (for example, *her true feelings*). Colligation is defined as "the co-occurrence of grammatical choices" (Sinclair 2004: 142). Semantic preference is "the restriction of regular co-occurrence to items which share a semantic feature, e.g., about sport or suffering" (Sinclair 2004: 142). Sinclair (2004: 35) showed that *true feelings* has a semantic preference for 'expression', which is recognised by verbs such as *express* and *reveal*. Mahlberg (2009: 113) observes that semantic preference is a technique for interpreting co-occurring item in the form of 'lexical sets'. The last category that represents one of the key components of a lexical item is 'semantic prosody', which is "the determiner of the meaning of the whole" lexical item (Sinclair 2004: 142), and displays a "subtle element of

attitudinal, often pragmatic meaning” (Sinclair 2004: 145). Sinclair (2004: 32, 35) found ‘reluctance’ or ‘inability’ to express the function of *true feelings*; he also noted that the semantic preference and semantic prosody can occasionally be ‘fused’ as a result of the speaker’s choice of co-selected words. For example, the adjective *invisible* in the example *invisible to the naked eye* shows a semantic preference for visibility and a semantic prosody of difficulty that are ‘fused’ in the same adjective. Stubbs (2001b) supports this view, and suggests that the border between these two categories is not always clearly defined.

With reference to lexical cohesion, Mahlberg (2009) points out that the description of a lexical item in terms of the five categories of co-selection makes it possible for lexical choices in texts to vary; “[i]t is this variability that is central to a corpus theoretical approach to cohesion” (Mahlberg 2009: 112). She explains that lexical items in texts consist of cores, and around each core, there are collocations. The patterns of the lexical item will vary considerably and connect with other items nearer to the borderline of the lexical item. Mahlberg (2009: 116-117) observes that lexical items create cohesion that operates on two dimensions. The first dimension is represented by ‘linear links’ between lexical items that contribute to the identification of a lexical item. Linear links define patterns of co-occurrence that surround the core of a lexical item and can be interpreted in terms of collocational patterns, which, as Stubbs (2001a) points out, create cohesion. The second dimension is illustrated by the non-linear links, which are distributed over larger stretches of text and show how different lexical items connect with each other away from their occurrence in a linear string. These links can be described as different forms of lexical repetition that connect the items in the immediate context around the core, and the items that spread throughout the text (Mahlberg 2009). Mahlberg (2009: 115) exemplified the two dimensions of cohesion by

analysing the lexical item *true feelings* that Sinclair (2004) already discussed. In the following example, Mahlberg (2009) selected an article from the *Guardian*, which demonstrates that when true feelings are expressed, they can turn into political insults. This article has the headline ‘I thought the mike was switched off...’, and it is about Dr Richard Simpson, who is said to have expressed his true feelings as the extract below shows. *True feelings* which illustrates the core of the lexical item is italicised as in Mahlberg’s (2009) illustration whereas the bold which shows the semantic preference of this core and the non-linear repetition links is emphasised by me.

- (1) Few may have heard of the junior Labour minister at Holyrood with responsibility for the fire service, if he had not **told** dinner party guests his *true feelings* of the striking firefighters. He **said** of them: “These people aren’t socialists, they’re protectionalists, and they’re fascists the kind of people who supported Mussolini. We must not give in too these bastards.

Despite insisting he was merely **repeating other people’s views** on the dispute; Dr Simpson fell on his sword yesterday.

(adapted from Mahlberg 2009: 115-116 – Guardian Newspapers Limited 2002)

In this example, Mahlberg (2009) shows that the semantic preference of *true feelings*, which was labelled by Sinclair (2004) ‘expression’, is recognised by verbs. For example, *tell* in *if he had not told dinner party guests his true feelings* carries out this preference. Mahlberg (2009) explains that this is a kind of a linear link. Then, she observes that lexical cohesion also works on a non-linear level in the example above. For example, *say* and *repeat* are lexical repetitions of *tell* as simple paraphrases. Therefore, cohesive links between *tell*, *say* and *repeat* have an added dimension of cohesion through the link with the core *true feelings*. This role of

collocation in creating text cohesion is also highlighted by Stubbs (2001a). Stubbs (2001a: 307) selected a number of individual items and phrases that occur in a text fragment in a book on the environment. These phrases were taken in the same order where they take place in the text to examine their contribution to text cohesion. One of these phrases is *cause untold environmental damage*. Stubbs's (2001a) approach to conducting this analysis is to compare the frequency and linguistic behaviour of some words and phrases in a single text with their frequency and linguistic behaviour in a large collection of texts, which represents norms of usage in the language in general. Based on this approach, he found that CAUSE (as a verb), as documented in large corpora, co-occurs frequently with collocates that are overwhelmingly unpleasant. It is also most frequently followed by abstract nouns:

CAUSE <*blindness, damage, danger, depletion, harm, loss, ozone, problems, radiation, warming*>

(Stubbs 2001a: 307)

Stubbs (2001a: 308) then selected another lexical pattern from the same text fragment, *the scale of the*. He observed that this pattern is succeeded by abstract nouns (such as *challenge, problem*), and is often used to refer to something very large, and usually to something bad (e.g., *underestimated the scale of the destruction* and *cannot cope with the scale of the fraud*). Stubbs (2001a) stressed that these examples of words and patterns and the referential meaning generated by them create a cohesive harmony together with other patterns from the same text, and share discourse prosody distributed across the text. Likewise, Mahlberg (2006: 374) underlines that, “[a]s semantic prosodies are attitudinal they do not only add to the connectedness of text but they also play a part in what is sometimes called the ‘tone’ of a text”.

Overall, this section emphasises that the meaning of a lexical item is not a property of a word itself but, as Cheng (2009) describes, but meaning in text is made through cumulative recurrences of the same item or of semantically associated items within and across texts that contribute to textual and intertextual meanings. Such overlapping of co-occurrence patterns is cohesive (Stubbs 2001b: 109; Mahlberg 2005: Chapter 7). Therefore, the present study will start the analysis of lexical cohesion by identifying the frequency of single word-forms through wordlists in both corpora (ALEC and LOCNESS). However, this analysis will be complemented by further analyses that focus on the meaning and function of lexical cohesion. That is, I will analyse lexical cohesion by applying the theoretical concept of a lexical item to individual essays from both corpora but this analysis will only be in some part (see Chapter 7). The following two sections will present studies that used corpus linguistic tools and concepts to analyse lexical cohesion.

3.3 Lexical cohesion in non-learner corpora from a corpus-linguistic perspective

A study by Morley (2009) highlights the importance of supporting the classic models of lexical cohesion with a corpus-linguistic analysis. Morley investigated how different types of lexical cohesion could connect newspaper headlines to the articles that followed them, and how such devices could enhance the rhetorical structure of a text. He first applied Halliday and Hasan's (1976) model to classify semantic fields of words in individual articles. For example, Morley explained that words such as *prison*, *trial* and *investigation* belong to the same semantic field of law in the article that he analysed, *The prison door*. He claims that such a semantic field does not merely form lexical cohesive links, but it also has a rhetorical effect in that its density shows the flow of the argument and the discourse structure. Morley considers this analysis of discourse using Halliday and Hasan's (1976) model to be an

example of a qualitative approach which is suitable to identify the rhetorical development of the discourse. However, he argues that cohesive ties between lexical items may not always produce meanings that are easily detectable on the surface of the discourse. In such a case, he suggests the use of a corpus-linguistic methodology to examine words in a text that are not immediately apparent to the naked eye.

Based on a corpus called Papers93, Morley (2009) analysed the lexical relationship between the headline ‘A moment for truth’ from the *Guardian* (2003) and the final paragraph of the article. He demonstrated that, when first looking at the paragraph, there did not seem to be anything particularly surprising about the lexical cohesion apart from the repetition of the phrase *moment for*, which created a cohesive link with the headline. He found that what was not visible on the surface of the text was that the expression *moment for* has its own strong semantic prosody. Semantic prosody is “the evaluative meaning – the good or bad ‘aura’ associated with the lexical item – which is not exhausted within the orthographic item but spread over the surrounding co-text” (Morley 2009: 11) – see also Sinclair (1996, 1998), Hunston (2002), Stubbs (2001) and Whitsitt (2005) for definitions and discussions of the concept. Morley (2009) collected adjectives that co-occurred with *moment for*. These adjectives imply that something important – either negative or positive – occurred at the moment mentioned (for example, a *critical moment* or a *sweet moment*). Morley’s (2009) approach to analysing his data, as indicated in the glossary notes in his research paper, referred to the use of corpus-assisted discourse analysis (see Partington 2004b). However, Morley did not provide any details about or discuss the advantages of this approach to the analysis of lexical cohesion. In addition, even though his study is relevant to the current research, it did not use a learner corpus and it examined the relationship between cohesion

and rhetorical structure, which is not the focus of the present study. Morley admitted that the results of his study only applied to the kinds of discourse found in newspapers.

Mahlberg (2005) analysed one category of cohesion, which is ‘general nouns’, that has been studied in the traditional models of lexical cohesion. With the aim of re-examining the characteristics of this category, she adopted a corpus-driven theoretical approach (i.e. the compilation of language data and observations) to describe the meaning and functions of this category. Her study builds on Halliday and Hasan’s (1976) concept of ‘general nouns’ in the sense that general nouns are the general meaning of the nouns and have cohesive functions. Mahlberg (2005) integrated the corpus-driven approach with language descriptive theories that help to identify the meaning of general nouns. The main descriptive theories underlying Mahlberg’s (2005) work are:

- (i) the theoretical concept of a lexical item as a unit of meaning via a lexical approach (Sinclair 1998),
- (ii) the Pattern Grammar approach, which suggests that the patterns and meanings of words are strongly related (Hunston and Francis 2000),
- (iii) the theory of lexical priming, which is based on the assumption that there is always a link between lexis and text through the lexical priming of a word (Hoey 2004), and
- (iv) the concept of local textual functions, which characterises the functions and the way in which a noun links to other items in a text (Mahlberg 2005: 171).

Mahlberg (2005: 19) asserts that “what is observed in the corpus needs to be named and described”, as these descriptive tools will have an impact on the results. Mahlberg’s methodology of analysing general nouns entails high-frequency words as a selection criterion. The selected nouns (such as *time* and *place*) were then analysed using concordance software according to the functional groups of the nouns. She analysed 100 concordance lines per noun. These concordance lines were observed via the lens of a corpus theoretical framework in that the concordance analysis focuses mainly on qualitative aspects, which prioritise lexis in the analysis of text cohesion. By combining quantitative and qualitative methodologies, her analysis showed descriptions of subtypes and functional categories for each noun type. Mahlberg’s (2005) results are comparable to Halliday & Hasan’s (1976) list of general nouns to some extent. For example, in Halliday and Hasan’s (1976) overview, nouns that are used to refer to people formed the largest subgroup; of the nouns they listed, *people*, *man* and *woman* also played a major part in Mahlberg’s (2005) study. She stresses that the functions of lexical items and properties of texts are closely interrelated, which has an implication for the analysis of cohesion. This indicates that cohesion is established through the overlapping patterns of multi-word units, and through connections between the units over passages of text. This observation is relevant to the present study because I will examine, but only in some part, how a corpus linguistic approach to lexical cohesion can help describe how individual lexical items connect with each other throughout the text creating a textual cohesion (see Chapter 7). However, Mahlberg’s (2005) results need to be generalised with caution. Voss (2008) points out that although the corpus which Mahlberg (2005) used in her pilot study included books (fiction, non-fiction and academic), the study relied extensively on journalistic texts. This may have affected the frequency count and the qualitative interpretation of categories that appear in this genre.

Another study that analysed lexical cohesion from a corpus-linguistic perspective was that of Cheng (2009), who examined a selection of spoken discourse events. She first identified forms of lexical cohesion using ConcGram (Greaves 2009) to generate a word frequency list. She then selected the ten most frequently occurring lexical items and suggested that they are possible sources of lexical cohesion. Subsequently, Cheng (2009) applied Sinclair's (2004) descriptive model of the lexical item as a way of analysing text qualitatively; this model is based on Sinclair's (2004: 148) remark that "the word is not the best starting point for a description of meaning, because meaning arises from words in particular combinations". This approach to characterising extended units of meaning is also shared by Stubbs (2001: 62-63), who states that linguistics lack "a descriptive theory of meaning". Cheng's (2009) findings showed that patterns of co-selection using the lexical item are able to provide a more comprehensive picture of textual and inter-textual coherence than when the focus is merely on lexical cohesion. Cheng's (2009) study showed how a combination of quantitative and qualitative research approaches can offer more detailed language descriptions, which can have useful pedagogical implications. Similar to Cheng's (2009) study, Mastropierro and Mahlberg (2017) applied the corpus-linguistic concept of a lexical item proposed by Sinclair (2004). They focused exclusively on semantic preference and semantic prosody to examine how cohesion can be investigated as an accumulative textual feature that plays a part in determining literary meanings in a single text. Unlike Cheng (2009), who employed a frequency wordlist, these researchers used a keyword analysis approach as the starting point for identifying cohesive networks in a short novel and its translated Italian version. Scott and Tribble (2006) point out that the keyword method can be used to identify the 'aboutness' of texts. Mastropierro and Mahlberg (2017) were primarily interested in examining the function of repetition as an essential feature in creating cohesive networks between lexical items.

Using WordSmith Tools to compare the short novel to a reference corpus, Mastropierro and Mahlberg (2017) generated the top 30 key words and selected three key words that were frequent, and which were relevant to the themes in the short novel. They used concordance data to identify the semantic preference and semantic prosody of these key words. Their approach to identifying these semantic fields is much more text-specific than it is generic. More specifically, they did not aim to identify the textual behaviour of these words in general. Instead, they were merely concerned with analysing the specific patterns in an individual text (the novel) that contributed to its literary meaning; that is, the 'local textual functions' of the key words (Mahlberg 2005; 2006). After they identified semantic preference and semantic prosody in the novel, Mastropierro and Mahlberg (2017) used them to search for cohesive networks that inter-connected the key words in question and their semantic fields in the source text. They then examined how these networks were reproduced in the translated text. Their findings showed that the key words exhibited similar patterns in the translated text to those identified in the original work. However, they observed that the cohesive network in the translated text appeared to be different from that in the source text when the key words under analysis were translated using a variety of different words. Such differences in the translated version, as Mastropierro and Mahlberg (2017) explained, were because the semantic preference and prosody words were connected with many nodes, which involved variation in the cohesive network, as well as different patterns of reiteration and collocation. In this regard, Baker (2011: 222) points out that the process of altering individual items can generate a chain reaction that interrupts the cohesive harmony of a text. Mastropierro and Mahlberg (2017) maintain that local alterations could affect the cohesion of the text extensively, particularly when the lexical item under investigation is relevant to the theme of the text.

3.4 Lexical cohesion in learner corpora from a corpus-linguistic perspective

There is an obvious lack of corpus-based studies that address lexical cohesion in learner corpora, with the exception of a few researchers (e.g., Flowerdew 2009; Thornbury 2010), who provided some insights into how lexical cohesion can be analysed from the perspective of corpus linguistics. On the contrary, a wealth of corpus-based research (e.g., Bolton et al. 2002) has focused on the study of conjunctions or connectors, as cohesive devices, rather than on lexical cohesion. In general, these studies have calculated the raw frequency and the percentage of use of a list of connectors in both NS and NNS corpora, and have compared the numbers with those in professional academic writing. These studies have also produced concordance lists to examine the patterns in which these connectors occurred. Such large-scale research in the area of conjunctions might be because these devices are more amenable to corpus analysis than is lexical cohesion, which requires a more subtle method of analysis. One of the potential complications of the analysis of lexical cohesion is due to the ability of corpus tools to capture the different categories of lexical cohesion. In this regard, Adorjan (2013) investigated how concordancing software applications could assist in examining the text-organising function of lexical repetition patterns using a corpus that contained a collection of summaries and comparison-contrast essays written by Hungarian EFL university students. Her aim was to observe which features of text that require an interpretation by a human, and which those that could be analysed via a computer. Based on these objectives, Adorjan (2013) analysed the summaries manually by employing Hoey's (1991b) model. By contrast, the comparison-contrast essays were partially computerised using *Concordance* (version 3.3) (Watt 1999). Each text in the comparison-contrast corpus was loaded into *Concordance* and was annotated for basic structural features, such as titles and paragraphs. Furthermore, the tagging included a description of paragraphs such as introductory

paragraphs, paragraphs describing similarities and paragraphs describing differences. Adorjan (2013) points out that concordance programmes are helpful in terms of organising the large quantity of data drawn from the corpus, particularly in the early stages of analysis when there is a need to count the number of words, sentences and paragraphs in the corpus. This claim is also supported by Collier (1994), who stresses the advantage of utilising frequency analysis software for counting the number of times particular words or pairs are repeated in a text, and to identify sentences in which repetitions occur. Concordance programmes, however, are unable to interpret qualitative data without human assistance in processing information (Hunston 2002). Furthermore, the results of Adorjan's (2013) study indicated that *Concordance* was only able to count simple repetition frequency lists, but it could not recognise other categories of lexical repetition. Instead, other types of lexical repetition (such as synonyms) are identified via the creation of Hoey's (1991b) traditional coding matrix, which consists of cells that contain potential cohesive links in the text. Moreover, *Concordance* was unable to establish bonds between sentences. Accordingly, Adorjan (2013) suggests that there is a need for the development of software that can incorporate more complex categories of lexical repetition than just simple repetition in order to allow research to use large corpora.

Despite the inability of concordance applications to analyse the different categories of lexical cohesion, Thornbury (2010) used the concordance software Wordsmith tools effectively to identify the potential chains of lexical cohesion. With the aim of analysing grammatical and lexical cohesion, Thornbury (2010) compiled a small corpus (10,000 words) of narratives written by teenagers (the Cringe Text Corpus, or CTC). Even though this corpus might not be a precise representation of a learner corpus, I consider it to be a type of learner writing;

therefore, I have listed it in this section instead of in Section 3.3, in which more professional texts are used. The part of Thornbury's (2010) study that is relevant to this discussion is his analysis of lexical cohesion. Using WordSmith Tools, Thornbury (2010) performed a keyword analysis measured against the BNC (the British National Corpus, World Edition 2001). He points out that a keyword analysis helps to identify words that are unusually frequent in the corpus. After excluding function words, he selected the top thirty keywords in the CTC corpus. He observed that these thirty words provided a clear signal of the main theme of the narratives that constitute the CTC. When examining this list, Thornbury noticed that a number of the key words were connected semantically because they belonged to the same lexical set (e.g., *walked, ran*) or to the same word family (e.g., *friend, friends, friend's*); because they were synonyms (e.g., *boyfriend, crush*), or they were collocates (*fell + butt*). Such relationships, as Thornbury asserts, are an indication that the texts in the CTC corpus have what Hasan (1989: 71) called texture: "The texture of a text is manifested by certain kinds of semantic relations between its individual messages". Thornbury (2010) further observed that some of these key words formed a type of what Hasan (1989) called a 'similarity chain', in which lexical items "belong to the same general field of meaning, referring to (related/similar) actions, events, and objects and their attributes" Hasan (1989: 85). An example of this chain in the keyword list, as Thornbury (2010) illustrates, is *tripped – fell – butt*. However, Thornbury points out that these relationships require a more fine-grained search to verify whether they were certainly inter-connected in their texts.

In order to examine more extended texts to find evidence of the semantic relationships in the CTC corpus, Thornbury supported the key word analysis with a cluster analysis (Scott 1997), also known as 'lexical bundles' (Biber et al. 1999) or 'n-grams' (Fletcher 2003; 2008). This

type of analysis involves the identification of a short (usually two to six) chain of words that are related simply because they occur together repeatedly. Nesi and Basturkmen (2006) argue that these sequences contribute to establishing cohesion. Thornbury (2010) generated four-word clusters that occurred five times or more in the CTC corpus. As Thornbury demonstrates, this analysis helps to identify typical contexts in which certain keywords are repeated. It also shows that certain patterns (and themes) are repeated regularly, and that they are discussed in the same way. This repetition, particularly of the patterns that contained keywords, makes these patterns generic, and confirms that the texts in the CTC corpus are inter-linked. What makes Thornbury's (2010) approach to the analysis of lexical cohesion valuable is his use of a combination of corpus tools that extract different levels of textual meanings. Furthermore, Thornbury's analytical procedure is in line with Schiffrin's (1987: 66) observation that "quantitative analyses [...] depend on a great deal of qualitative description prior to counting (in order to empirically ground ones' categories) as well as after counting (statistical tendencies have to be interpreted as to what they reveal about causal relations)". Schiffrin (1987) emphasises the importance of the complementarity of quantitative and qualitative approaches in discourse studies. Thornbury (2010) concludes that this cyclical interchange between counting and interpreting characterises the application of corpus analysis to discourse accurately.

One of the key concepts in corpus linguistics is the relationship between form and meaning. Flowerdew (2009) analysed the use of signalling nouns such as, *attitude* and *difficulty* across and within clauses, and observed that such nouns contributed to the creation of lexical cohesion in a text. Even though Flowerdew's (2009) study is a corpus-based error analysis, it offers an insightful way to analyse lexical cohesion that could be still relevant to the present

study. To determine whether the use of SNs correlated with high marks received by students, Flowerdew (2009) used a corpus of argumentative essays produced by Cantonese L1 first-year university students (drawn from ICLE). Based on this corpus, he categorised errors as paradigmatic (inappropriate choice of a word), or syntagmatic (inappropriate choice of collocation/colligation). He then provided the frequency of different types of errors in the use of SNs. Flowerdew's grouping of errors into paradigmatic and syntagmatic levels is beneficial for explaining the forms and functions of SNs; furthermore, he stresses the importance of corpora and corpus tools to study the context of use and the co-text of specific errors. Nevertheless, he did not give a clear and adequate explanation of how he used corpus tools and concepts that can show, for example, how he judged the appropriateness of the selected SNs with respect to colligation and collocation. Flowerdew (2009: 351) mentioned briefly that he checked the students' writing for the appropriateness of colligation (for example, SN + preposition) against the BNC using *Phrases in English* (PIE) (<http://phrasesinenglish.org/>). However, what is the procedure that he adopted to check the appropriateness of collocations in the use of SNs? Overall, Flowerdew's (2009) results revealed that learners with higher marks tended to use more SNs and with greater accuracy than learners with low marks. However, he admitted that these findings would have been more useful if a comparison between non-native data and native data had been made because he contends that this comparison would demonstrate whether students were under- or over-using SNs, and would indicate the extent to which students used different types in comparison to native speakers.

Although not addressing the concept of lexical cohesion explicitly, there is other work in corpus linguistics that emphasises the importance of analysing lexical items at the syntagmatic level, particularly through collocation, semantic preference and semantic

prosody. In order to identify similarities or differences between native and non-native writers, Xiao and McEnery (2006), using comparable corpora, contrasted the collocations of near synonyms and their semantic prosodies in English and Chinese. This type of comparison was not between native speakers' and learners' languages (as is the case in my study), but between speakers of different L1s, which Granger (1998a: 12-13, 2002: 12-13) termed 'Contrastive Interlanguage Analysis' (CIA). Nonetheless, introducing such work in this section still adds some thoughts to my analysis of lexical cohesion. Xiao and McEnery (2006: 104) examined three groups of near synonyms in English and their close translation equivalents in Chinese, namely the 'consequence group', the 'CAUSE' group and the 'price/cost' group. They first discussed collocation and semantic prosody for each group of near synonyms in English, and then they repeated the same procedure for each group of synonyms in the Chinese corpus. Their findings showed that although English and Chinese are markedly distinct languages, they were somewhat similar in terms of the behaviour of collocations and semantic prosodies of near synonyms. This observation complies with the findings of previous research that has examined language pairs that are related to each other, such as English versus Portuguese (Sardinha 2000), English versus Italian (Tognini-Bonelli 2001: 131–56) and English versus German (Dodd 2000).

Despite the similarity in the use of semantic prosodies in English and Chinese, Xiao and McEnery (2006) affirmed that it was generally common that Chinese learners of English, as compared to native speakers, selected inappropriate words because they were not aware of semantic prosody. For example, Xiao and McEnery (2006) demonstrated this argument using an example taken from the Chinese data. The students used the verb *cause* inappropriately with a positive semantic prosody; for example, 'the city *caused* him great interest, *caused* all

citizens to grasp time and chances, to work for a better life' (Xiao and McEnery 2006: 126). The authors maintained that an English speaker would be highly likely express the same example more naturally by selecting *lead to* or *bring about* rather than *cause*. They alleged that the intuition of native-speakers could mark the unnatural usage of a word with its semantic prosody, whereas the intuition of L2 learners is not as reliable as their L1 intuition. Consequently, the authors suggest that teachers need to raise learners' awareness of the differences between L1 and L2, by comparing the collocational behaviour and semantic prosody and preference of near synonyms in the L1 and their close translation equivalents in the L2. This should decrease the type of errors that result from differences in semantic prosodies between L1 and L2.

3.5 Conclusion

This chapter has emphasised that lexical cohesion is about meaning and this meaning is not characterised in terms of individual lexical items but it is created through the collocational patterns of lexical items which inter-connect in a text creating textual cohesion. The chapter also reviewed studies that used non-learner corpora and those that used learner corpora. Both types of studies highlighted valuable techniques for the study of lexical cohesion from a corpus-linguistic perspective, starting with frequency lists and progressing to a keyword list and cluster analysis. These studies further stressed the importance of descriptive tools for corpus linguistics that include concepts such as semantic preference and prosody. Nevertheless, the emphasis of most of this work has been monolingual or from a cross-linguistic perspective (contrastive analysis). To the best of my knowledge, no study to date has taken a corpus-based comparative approach to the analysis of lexical cohesion in English writing by native and non-native speakers of English, namely Arab speakers of English.

Nonetheless, this is an area that requires further exploration. Chapter 7 will explore the potential of a corpus-linguistic approach to the analysis of lexical cohesion by identifying differences or similarities between English writing by native speakers and by Arab speakers of English. The following chapter describes the two corpora that were used in this study.

Chapter 4

Methodology

4.1 Introduction

This study aims to analyse argumentative writing by native and non-native speakers of English to examine the use of lexical cohesion. Argumentative writing means the use of language to promote support for or against certain ideas. Such a process of argumentation, as Károly (2002) points out, requires persuasive skills and logical organisation of information. Connor (1987) suggests that meaningful usage of cohesive devices can contribute to the flow of text argument and the development of its patterns. Therefore, argumentative essays would provide appropriate basis for the analysis of lexical cohesion in the present study. Surprisingly no corpus of argumentative writing by Arab learner writers exists. Consequently, I needed to compile an Arab learner corpus of English writing to make this study achievable. The first part of this chapter will describe the compilation of the Arab Learner English Corpus (ALEC) defining its design criteria and the type of Arab participants (cf. Section 4.2). This will be followed by Section 4.3, which summarises the ethical considerations in this study. Afterwards, the chapter will discuss which native speaker corpus is appropriate to serve as control corpus in this study – whether an expert native corpus or a student native corpus (cf. Section 4.4). Section 4.5 then suggests the Louvain Corpus of Native English Essays (LOCNESS) as reference corpus for a comparative analysis. Subsequently, in Section 4.6 I will highlight comparability considerations between both corpora.

The second part of the chapter focuses on the framework of analysis that will be adopted in the present study. The chapter stresses how integrating quantitative and qualitative methods is

recommended for a comprehensive analysis of lexical cohesion (Section 4.7). Then, the chapter highlights the issue of punctuation and run-on sentences in learner writing, which can mask a number of important cohesive relationships in text. Therefore, the T-unit is suggested as a discourse unit in the analysis of lexical cohesion when addressing learner writing (cf. Section 4.8). Section 4.9 then identifies which lexical items to include or exclude in the analysis of simple and derived repetition. The section that follows discusses whether co-reference is a necessary criterion in the analysis of lexical cohesion (cf. Section 4.10). Finally, a conclusion is provided.

4.2 Building the Arab Learner English Corpus (ALEC)

A number of learner corpora are available. One of the most widely used learner corpora that include a type of argumentative writing is the International Corpus of Learner English (ICLE) (Granger et al. 2009). It contains over 3 million words of EFL writing from 16 categories of learners: Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Tswana, and Turkish. However, ICLE was not used because it does not contain English essays that are written by Arab speakers. Consequently, for the current study I compiled a small corpus of Arab L2 learners. I will refer to this corpus as ALEC (Arab Learner English Corpus) throughout the thesis. I followed the guidelines of ICLE in collecting the ALEC corpus. For example, as in ICLE, I requested participants to fill in a learner profile, and I collected the right type of material, which is argumentative essay writing. ALEC can then be extended in size in a future study and added to the ICLE corpus if it will be accepted by Sylviane Granger, the former and coordinator of the ICLE project.

Yet it is worth mentioning that the attributes of the subjects who contributed data to ALEC are slightly different from those in ICLE. For example, the subjects in the ALEC corpus are ESL learners not EFL learners (For definitions of ESL and EFL, refer to Chapter 1, Section 1.4). A further difference is that participants' educational level in ALEC is Master and PhD, which is unlike ICLE that consists of undergraduate university students. However, both corpora still share some attributes such as medium, genre and technicality. This means that both corpora consist of written productions, which all represent the same genre of argumentative essay writing. Furthermore, even though the essays in ICLE and ALEC cover a variety of topics, the content is similar since the topics are all non-technical, and about general issues.

To create the ALEC corpus, I distributed argumentative essay prompts to 28 Arab participants in the period between July and September 2015. The writing task required them to write about general topics in an argumentative style (see Appendix 1 for a template of the writing task instructions and essay titles of the selected topics). Participants had to select one of the topics presented in the writing task and write about. The writing task was completed at home (untimed) for participants' convenience. The essays had to be at least 500 words long. Then, I requested the Arab participants to fill in a learner profile (see Appendix 2 for a learner profile template). This profile contains information about different variables for each participant in the corpus, which can help any researcher who wants to use this corpus in future to interpret their results meaningfully. This information was stored for each participant in an excel file with an ID number. Table 7 summarises the basic information that was elicited from learner profiles for the 28 participants.

No. of Participants	L1 language background	Gender	Age range	Level of study	Proficiency mean average	Postgraduate programme	Learning context
28	Libyan (n=21) Iraqi (n=4) Saudi Arabia (n=3) Egypt (n=1)	5 F 23 M	25 to 45	MA (n=2) MSc (n=1) PhD (n=25)	IELTS 6.5	Economics Pharmacy Medicine Business Engineering Management Linguistics English Literature	University of Nottingham University of Derby University of Hull

Table 7 Summary of the basic variables in the learner profiles by Arab L2 learners

As Table 7 indicates, participants came from four Arabic-speaking countries and both genders were included. Their age ranged from 25 to 45. Most participants were Ph.D. candidates and they were all expected to be advanced English learners. The mean average IELTS score for the Arab participants was 6.5. This average was calculated from the IELTS scores that were provided by participants in their learner profiles. The participants studied at different universities in postgraduate programmes from various fields.

Although the expected total corpus size of ALEC was 50,000 words, I only managed to collect 28 argumentative essays with 17,564 words. This difficulty of collecting a large number of essays was because most of participants were doing their PhD, and they were therefore extremely busy, and not willing to allocate some time to complete the writing task. Also, due to the limited time frame for any PhD study, the collection of these essays was completed over a short period of time. In this regard, researchers (e.g., McEnery and Wilson 2001, McEnery et al. 2006) comment that the corpus size is determined by the language

variety that the corpus is expected to represent, and by the research aims. For the language variety, the present study proved that learner data is not easy to acquire. Granger (1998b: 10) supports this view and argues that “one can hardly expect learner corpora to reach the gigantic sizes of native corpora”. With reference to the purpose of the research, the current study analyses lexical cohesion which requires a detailed analysis. As Connor (1984) points out, a small number of essays are a common practice in detailed, exploratory analyses such as that of cohesion analysis. What is more, I am also using another set of data by native speakers, which means another corpus needed to be analysed.

Consequently, I would argue that the ALEC corpus is large enough for the research purpose I undertake in this study. What is more important is that there are existing studies which used small corpora. For example, Thornbury (2010) identifies grammatical and lexical cohesion in a small corpus of teenage written narratives which contains 10,000 words. Also, Reynolds (1995) compares lexical cohesion between NNSs and NSs using a small corpus which includes 13,170 words of expository essays. Furthermore, McEnery and Kifle (2002) use a corpus of 22,000 words to investigate epistemic modality in argumentative essays written by Eritrean and British students. McEnery et al. (2006: 72) stress that the sheer volume of language data in many larger corpora defies a close linguistic examination and makes it at the very least impractical. Likewise, Flowerdew (2004) suggests that very large corpora tend to rely more on quantitative methods for analysis, while smaller corpora are more conducive to qualitative analyses which, although not as generalisable, can yield richer, more detailed insights into language in particular contexts (McCarthy and Handford 2004).

4.3 Ethical considerations

I am aware that any study that involves human participants needs an ethical approval. Therefore, I submitted the ethics form to the specialised Ethics Committee at the University of Nottingham and I was granted an approval to start collecting data. In addition, I further applied for a financial support fund to the School of English at the University of Nottingham to allocate an incentive for each participant as an appreciation of their participation and time. The school agreed to provide me with a £15 electronic Amazon voucher for each participant.

Participants were informed that their personal information would be kept confidential and would be used for research purposes only. Thus, participants were asked for their permission to complete a consent form (see Appendix 3 for a consent form template). They were also informed from the beginning that the participation in this study would be voluntary and they had the right to withdraw from the study whenever they wanted. Participants sent their essays to my email and I created a folder in my computer to save all these essays along with the students' learner profiles and signed consent forms.

4.4 Finding a reference corpus of native data in this study

The ALEC corpus that I collected represents pieces of opinion and argumentative essays which are to some extent non-technical and semi-academic. Accordingly, I needed to find a corpus that is as close as possible to ALEC in terms of genre and other dimensions. As pointed out by Barlow (2005: 345), “a variety of issues arise when a learner corpus is to be contrasted with an NS corpus”. A major issue in this type of corpus comparison as stressed by Granger (2002: 12) includes variables such as the field of corpus, its level of formality, the level of proficiency of native speakers and many other factors. A native speaker corpus could

be a corpus of native students or a corpus of native professional writers. Meunier et al. (2011) suggest that if the purpose of comparing the NN and N corpora is to describe areas of argumentative or academic writing that learners need to improve, a native speaker corpus such as press editorials or academic articles may be more advisable than a native student corpus. Ädel (2006: 206) further supports this view and states that professional native-speaker writing would be more recommended than native-speaker student writing as it represents the norm that advanced foreign learner writers endeavour to achieve and their teachers seek to promote. In cases where a reference corpus is based on native students' writing, however, Ädel (2006) demonstrates that caution should be taken before any recommendations to learners are provided because language usage by native students might not be the proper target.

However, a number of researchers do not recommend the use of an expert native corpus as a reference corpus against which to compare non-native writing. Barlow (2005), for example, discusses the issue of genre and argues that the mixture of genres in the general corpus makes it inappropriate reference corpus for the learner corpus, which regularly incorporates one genre. Granger and Paquot (2009), who selected BNC (the British National Corpus) as reference corpus in their study of ICLE, point out that the two corpora are not comparable. They interpret such a difference to the fact that expert texts are expository in that they are topic-based (cf. Britton 1994) and depend on the comprehension of general concepts (cf. Werlich 1976). Besides, Granger and Paquot (2009) explain that expert texts are discipline-specific (i.e. humanities, technology, engineering), while learners' essays discuss a variety of general topics such as 'feminism', 'the impact of television', 'drugs', etc. These researchers, therefore, suggest that we need to take caution when we interpret results because a number of

differences between learner essays and expert texts may purely indicate differences in their communicative goals and settings (cf. Neff et al. 2004). For example, as Sugiura (2007) points out, it is not easy to compare an academic essay about the global economic depression produced by a native speaker and a short informal essay that describes a summer holiday by a non-native speaker, and then we discuss differences between these learners, which might be only due to the difference in the given topic.

Barlow (2005) maintains that such variability in genre could be decreased when another reference corpus is used such as a newspaper, which is based on a single text type. Nevertheless, he explains that the newspaper can also contain different genres, because it consists of different sections, which are written by different writers who vary in their writing style, and this might lead to variability in language use. But such variability, he claims, is not as great as the one in a general corpus. Barlow's (2005) possible solution of genre inconsistency is to find a corpus that is closely matched in terms of genre and other variables to a learner corpus. He suggests a corpus such as LOCNESS as a point of reference when a learner corpus such as ICLE is studied claiming that both corpora represent one text type (i.e., argumentative writing). McEnery and Kifle (2002) support the idea of comparing learner corpora with native student writing instead of professional native writing. McEnery and Kifle (2002: 182) claim that "the comparison of learner English against that of journalists, authors, and the like hardly seems relevant". Researchers such as Lorenz (1999) and Hyland & Milton (1997) support this argument and criticise the use of professional writing in learner corpus research. They claim that it is "both unfair and descriptively inadequate" (Lorenz 1999: 14) and argue against the "unrealistic standard of 'expert writer' models" (Hyland & Milton 1997:

184). I nevertheless agree with De Cock's (2003: 196) observation that "argumentative essay writing has no exact equivalent in professional writing".

Gilquin and Paquot (2008) add that the use of native student writing (e.g., LOCNESS) seems an appropriate reference corpus to EFL learner writing when the comparison seeks to describe and evaluate interlanguage(s) properly. Thus, LOCNESS is the closest, among other learner corpora, to be used as reference corpus to the ALEC corpus to examine features of non-nativesness at the level of lexical cohesion. Despite the availability of many other corpora of native-student English (e.g., BAWE; Cambridge Learner Corpus) that are also publicly available and easily-accessible, they are more academic and discipline-specific, compared to LOCNESS.

4.5 LOCNESS as reference corpus

LOCNESS is an example of a corpus of native students' English, which was specially built by Granger in order to make it as comparable as possible to ICLE. Granger and Tribble (1998: 204) describe the LOCNESS corpus as a small, highly specialised corpus of NS student essays, because it is a genre-specific in that all essays are argumentative essays. LOCNESS encompasses a wide range of topics (e.g., 'boxing', 'traffic') that concern different issues and could belong to multiple domains of 'technology', 'sport', 'environment', etc. Researchers such as Neff et al. (2004) who analyse the type of English in LOCNESS observe that novice L1 writing often appears to occupy an intermediate position between academic writing and EFL learner writing. This assessment shows that the LOCNESS student corpus contains linguistic features that are similar to non-native learner writing. These features of LOCNESS

make it more comparable to the English language in the ALEC corpus. Table 8 demonstrates the composition of LOCNESS.

LOCNESS sub-corpus	N. of texts	Tokens	Topics
British A-Level students	114	60,209	Argumentative
British university students	90	95,695	Literary
American university students	232	168,400	Argumentative

Table 8 The composition of LOCNESS

Access to the LOCNESS data (the sub-corpus of A-Level students) was kindly provided via email by Sylviane Granger, who created the corpus. Only the British sub-corpus of A-level students' essays was considered as reference to ALEC (see Appendix 4 for both ALEC and LOCNESS British A-level students' essays sub-corpus CD-ROM). This selection was due to the fact that most of the essays written by the British A-level students are argumentative essays about general topics, which make them analogous in their genre to the essays in the ALEC corpus. It would be, though, more sensible if the sub-corpus of British university students' essays were chosen to match the educational level of Arab L2 students in ALEC. Nevertheless, most of the British university students' essays are written on a range of literary topics that deal, for instance, with works by Camus, Sartre and Hugo. This issue of literary topics makes this sub-corpus of LOCNESS genre-specific and therefore it was left out.

I selected 30 essays from the British A-level students' sub-corpus. This sample contains 16,268 words out of 60,209 words. The corpus size was selected to be as close as possible to the corpus size of the ALEC corpus which is 17,564. The essay topics in the British A-level students' corpus were selected to match closely the topics in the ALEC corpus. Table 9

illustrates that the first three essay topics are matched (computers vs. human brain; transport /traffic jams; boxing) while the other topics are different. However, what matters is the genre of these texts, which is the same for all topics for both groups in that all essays are argumentative.

LOCNESS sample: British A-level essays Essay topics	No. of essays	No. of words	ALEC Essay topics	No. of essays	No. of words
Computers and. human brain	10	4,657	Computers and human brain	7	4,766
Transport (traffic jams)	10	6,162	Transport (traffic jams)	3	1,800
Boxing	10	5,449	Dangerous sports	3	1,496
			Free education	4	2,721
			Immigration	11	6,781
	Total: 30	Total 16,268		Total: 28	Total 17,564

Table 9 Topics of argumentative essays in LOCNESS sample and ALEC

The essay length in the A-Level corpus of LOCNESS ranges roughly from 300-900 words, whereas essays in ALEC are longer ranging from 500-1000. The average essay length for both corpora is 542 for the A-Level corpus of LOCNESS and 627 for ALEC. Essays in both corpora were saved as text files. In addition, errors were not modified because any change might alter the students' writing character. Subsequently, each essay in each corpus was segmented into T-units (Section 4. 8 provides a definition of the T-unit).

4.6 Comparability between ALEC and LOCNESS

The ALEC corpus is a new corpus and no native students' data exactly match it. This left LOCNESS as the most possible benchmark for the purpose of the present study. While both corpora are closely comparable for text type (i.e. argumentative writing), they differ in a number of dimensions. For example, the setting (e.g., time, place, level of formality) under which participants in both corpora wrote their essays is different. Participants in the LOCNESS corpus wrote their essays under formal timed-exam conditions, while my Arab L2 participants were requested to write their essays at home in untimed conditions. Writing setting, as Ädel (2008) points out, is important in learner writing, and task variables such as time profoundly influence linguistic output (Chafe 1986). However, such a difference in task variables when comparing two corpora is not easy to control in any study that involves learner corpora and thus results have to be interpreted carefully.

Other variables such as participants' age and their educational level were also different in ALEC and LOCNESS. As demonstrated in Section 4.2, the ALEC corpus contains MA/MSc and PhD students, whereas the LOCNESS corpus includes A-Level British undergraduates. Similar learner-based studies also compare NNS and NS corpora that are not comparable in terms of these variables. McEnery and Kifle (2002), for example, compare argumentative compositions written by Eritrean second-year university students, whose age are around 20, with argumentative essays produced by 16-year-old British school children. These researchers argue that in learner corpus-based studies an attempt to precisely match factors such as age groups and educational levels can be hard to achieve. McEnery and Kifle (2002: 185) observe that such a difference in variables is somewhat inescapable as the educational environments

being compared are so different that simply matching age groups and educational levels are trivial.

Furthermore, the fact that Arab L2 participants in this study are masters and doctoral students does not entail that their English writing is professional and near-native. This situation could be true if these participants were requested to write for academic purposes where they are put under pressure to complete their theses, and will then be assessed on them. However, the purpose of the writing task in the present study was to ask participants to express their personal opinion and defend it without relying on academic sources to quote from. Such a difference in register of writing between formal and informal might make the writing of the Arab L2 learners less academic and hence more comparable to that of the British A-Level students. We could conclude that language register can determine vocabulary, structure, and sometimes grammar in writing.

A further difference between ALEC and LOCNESS is that the discourse topics in the A-level sub-corpora of LOCNESS and ALEC are not equally well represented. For example, the British A-level essays on the topic of 'transport' include 6,162 words, while the Arab L2 essays contain only 1,800 words. Additionally, topics in both corpora are about a range of general issues in life such as traffic, immigration and free education. On the one hand, topic is an important consideration influencing the choice of vocabulary (Biber 2006) and the type of lexical cohesion established in return (Hoey 2005). On the other hand, this range of subject matters in both corpora deals with aspects of everyday life, discussing people, events, social problems, etc. This suggests that the corpora used in the current study would be unlikely to contain specialised words.

The issue of variability in topics (also called ‘domains’ or the ‘field’ of the corpus) is further discussed by Hoey (2005: 115), who argues that “priming (for cohesion) is genre and domain specific in the first instance, though there are many primings that apply across generic and domain boundaries”. The absence of specialised/technical vocabulary in LOCNESS and ALEC can be explained by the fact that participants were instructed to write their essays for a non-specialist reader although the subject matter of the essays may be technical. As a result, their writing would enclose more general vocabulary than technical giving that their essays are not addressed to specialists in the field (L. Flowerdew 2004: 132). The variability of topics in both corpora could be further justified because other researchers (e.g., Hinkel 2002; J. Flowerdew 2010), who have dealt with lexical cohesion, also used corpora that encompass a range of topics.

4.7 Combining quantitative and qualitative approaches in the analysis of lexical cohesion

Bell (2005) suggests that no research should be built on an individual method of analysis if the research intends to produce valid and reliable results. Therefore, some researchers (e.g., Johnson and Onwugbuzie 2004: 17) recommend conducting ‘mixed-methods research’, where the researcher combines quantitative and qualitative methods. Johnson et al. (2007) explain that such a combination includes the benefits of both methods and enables researchers to understand the breadth and depth of the research problem under investigation.

In the present study, a quantitative method will not be satisfactorily sufficient to describe lexical cohesion in native and non-native writing. I will, therefore, integrate quantitative and qualitative data analyses by combining text linguistics and corpus linguistic tools and concepts. Such a combination can offer different layers of analysis of the forms, frequency

and function of lexical cohesion in NNS and NN writing which can then have useful pedagogical implications. In Sections 4.7.1 and 4.7.2, I will establish the basic framework of the analysis that is used throughout the course of the forthcoming chapters.

4.7.1 The framework for the quantitative analysis

In the present study, I will count simple and derived repetition applying my model of ‘Lexical Repetition Network’ (LRNetM) that I will introduce in the next chapter. LRNetM is based on corpus linguistic tools, namely wordlists, which will be generated with the help of the software programme *WordSmith Tools* version 6.0 (Scott 2012). A wordlist reveals the number of times each individual word occurs in the corpus data (Hunston 2002: 67). For quantifying signalling nouns, I will use a manual text analysis method due to the textual nature of SNs and the small size of both corpora used in this study. Chapter 5 will provide more detail on the counting procedure for each cohesive category.

The quantitative analysis of simple repetition, derived repetition and signalling nouns in both corpora includes descriptive and inferential statistical analyses. The descriptive analyses include frequency counts, mean and standard deviation. Means of frequency counts of each cohesive device will be compared between both groups. Then, a t-test is used to verify whether the means of the two groups are statistically significantly different from each other. A p-value of below .05 is considered significant in this study due to the small size of each corpus. Chapter 5 will provide a thorough discussion on how frequency means are compared.

4.7.2 The framework for the qualitative analysis

The quantitative procedure provides a starting point for further qualitative-based analysis. The second part of the present study is to examine the functions of simple repetition, derived repetition and signalling nouns in both corpora. In order to attain this goal, the quantitative data is firstly analysed from a text-linguistic perspective. Such a qualitative analysis is principally concerned with the paradigmatic relationships between lexical items by indicating what frequently recurs and which words are likely to belong to the same category of lexical cohesion. This analysis will be introduced in Chapter 6.

The text-based analysis will be then complemented with a further qualitative analysis using corpus descriptive concepts, namely ‘semantic preference’ and ‘semantic prosody’. These two concepts will add further detail to the function of lexical cohesion at the syntagmatic level. This analysis will be carried out by selecting three lexical items from ALEC and LOCNESS and then examine their semantic preferences and semantic prosodies in a reference corpus (Chapter 7 describes this reference corpus in further detail). In this reference corpus, concordance lines of each lexical item in question will be obtained. Mahlberg (2005: 54) describes that a ‘concordance’ is a method that is used to demonstrate instances of a word. This demonstration is achieved by concordance software, which records each occurrence or particular occurrences of the search item or the ‘node’ in a text or corpus, with a certain amount of context on either side, left and right. These concordance lines, as Tognini-Bonelli (2001: 3) explains, will then be scanned vertically in order to look at the repeated patterns that take place in the co-text of the node. Tognini-Bonelli (1996) reports that a concordance makes repeated patterns visible; the syntagmatic relations can be observed on the horizontal axis, whereas the paradigmatic relations, which show what frequently recurs syntagmatically, can

be observed on the vertical axis. These repeated patterns, as Mahlberg (2005) points out, provide clear evidence for the close relationship between meaning and form, and they will help identify the meanings and textual functions, namely semantic prosody of the word being examined. Once semantic preferences and prosodies of the three lexical items under investigation have been identified in the reference corpus, they will be identified in individual texts from both ALEC and LOCNESS. Then the semantic preferences and prosodies of the items in question will be compared against the reference corpus to see whether they follow the typical use of language as it is indicated in the reference corpus. Chapter 7 will provide further detail on this procedure of analysis.

To conclude this section, the combination of quantitative and qualitative methods to analyse lexical cohesion demonstrates Partington's (1998: 106) observation of how the methodological combination of wordlists, concordance lines, and viewing the lexical word under investigation in its original context, can enable the researcher to examine patterns of textual cohesion.

4.8 The T-unit and the problem of punctuation in L2 writing

Researchers vary in selecting which discourse unit is the base for cohesion analysis. For example, Halliday and Hasan (1976) and Hoey (1991b) take the sentence as the unit of analysis and identify cohesive ties or links between sentence boundaries. Halliday and Hasan (1976: 9) claim that “cohesive ties between sentences stand out more clearly because they are the only source of texture, whereas within the sentence there are the structural relations as well that ensure that the parts go together to form a text anyway”. In contrast, other researchers (e.g., Halliday 1985) suggest the clause complex, or the clause (e.g., Halliday &

Matthiessen 2004), as opposed to the sentence, to analyse cohesion. Gutwinski (1976), on the other hand, analyses lexical cohesion both within the clause and among sentence boundaries. Tanskanen (2006) also analyses both intra-sentential (i.e. within sentences) and inter-sentential (i.e. between sentences) cohesive relations. He explains that written sentences can be considerably long, which can make cohesion within a sentence meaningful for the unity of a text.

However, most of these researchers, particularly those who take the sentence as the unit of cohesion analysis, have analysed expert writing, where the use of punctuation is correctly applied and the full stop can be trusted to mark the end of the sentence. In the current study, however, I analyse learner writing, in which a number of Arab speakers of English and sometimes native speakers of English use incorrect punctuation and tend to write very long sentences. Researchers such as Gutwiniski (1976: 105-106) stress the importance of punctuation when considering cohesion in a written text. Therefore, my analysis takes the ‘minimal terminable unit’ or what is called a ‘T-unit’ as a unit of analysis. Hunt (1970: 189) defines the ‘T-unit’ as “the shortest units into which a piece of discourse can be cut without leaving any sentence fragments as residue”. Witte and Faigley (1981) advocate the use of T-units and argue that to examine only cohesive ties that stretch over the boundaries of orthographic sentences would ignore cohesive markers between independent clauses. The following examples illustrate how sentences are divided up into T-units according to the development of their syntactic structures (i.e., simple, compound and complex sentences). The T-unit is marked with (/) in the examples below.

(a) **Simple:** There is a traffic jam today. (1 T-unit)

(b) **Compound:** There is a traffic jam today // so I will be late for work // and my boss will be mad. (3 T-units)

(c) **Complex:** If there is a traffic jam today, I'm going to find a different way to drive to work since I don't want to be late. (1 T-unit)

(adapted from Kaderavec 2015: 381- 382)

In a T-unit analysis, simple sentences are usually not divided up into T-units because they are normally an equivalent to a T-unit in their length. Generally, T-units have most of the characteristics of a complete sentence in that both of them need a subject and a verb. For example, if we look at the first sentence above which illustrates the simple sentence example, we find that this sentence can be classified as either a T-unit or a complete sentence. In the case of compound sentences, coordinating conjunctions (*and, but, or, so*) usually work as an indicator to separate combined sentences into multiple T-units (see Example b). In contrast, a complex sentence containing a subordinating clause cannot be separated because subordinating clauses cannot stand on their own. Subordination can be formed using words such as (*after, although, as, as if, because, before, even if, if, since, unless, until, when, whenever, wherever, whereas, and while*). The last sentence in (c) demonstrates a type of complex sentence, which contains two subordinating clauses: (*If there is a traffic jam today; since I don't want to be late*). These two subordinating clauses are not detached, and the whole complex sentence is deemed to form 1 T-unit.

A number of researchers who analysed second language writing support the use of the T-unit as a unit of lexical cohesion analysis. Reynolds (1995), for example, justifies his selection of the T-unit instead of the sentence by claiming that NNS writing, and sometimes NS writing, contains sentence fragments and run-on sentences, which make it difficult to mark out sentences for cohesion properly. In addition, the fact that Arab speakers of English tend to use a considerable amount of additive co-ordinators (e.g., *and*) (Williams 1989) can also disguise a number of cohesive forms that occur in these co-ordinating independent clauses. These reasons, therefore, have made the use of the T-unit more appropriate unit of analysis than the sentence. Examples 1 and 2, taken from ALEC and LOCNESS, indicate the use of the co-ordinator *and*, and the use of commas in students' writing.

(1) 1 People needs *computers* to help on their work and *computers* needs human brain to develop and integrate new functions based on the operation raised. 2 It is us who think for *computers* and it is us who embed and program the *computer* to run some certain operations. (ALEC)

(2) 1 It is the control centre of a body, can instruct it to move freely around its surroundings, it is able to interact with other humans and objects. 2 The *brain* can make decisions, the *brain* can communicate. (LOCNESS)

Examples (1) shows that the Arab non-native writer uses the co-ordinator *and*, which creates a kind of a parallel structure, while in Example (2), the native writer uses commas to separate their sentences, and uses also a sentence fragment (e.g., *can instruct it to move freely around its surroundings*). Both writers do not terminate their sentences where needed, which as a result affects the lexical cohesion count. A case in point is Example (1) in which the first sentence contains two instances of the word *computers*. If the sentence is taken as a unit of

cohesion analysis, no cohesive relationship in this case would be recorded because this study considers only cohesion across (not within) text unit boundaries. Nevertheless, if we consider the first sentence in Example (1) as consisting of two T-units, there will be a lexical cohesive relationship of simple repetition, which connects *computers* with *computers*. This also applies to the second sentence in the same example that contains two occurrences of *computer*.

The majority of essays in ALEC and LOCNESS have a good control of the sentence structure and punctuation. This facilitated the task of splitting up the essays into T-units because students wrote in complete sentences that are correspondent to a T-unit. This means that every sentence almost equals a T-unit because they express complete ideas. However, a number of essays were of low level, and thus the T-unit was selected to deal with such cases. Accordingly, this study identified lexical cohesive devices that occur between T-units in each essay in both corpora. The end of the T-unit marks the place of a possible full stop. The T-unit was demonstrated with two slashes. Appendix (5) shows an essay, taken from ALEC, which is segmented into T-units. This essay is a good example of the intensive use of commas.

4.9 Exclusion and inclusion criteria for determining simple and derived repetition

The automatically determined lexical items typically do not all contribute to lexical cohesion. Chapter 2 (Section 2.9.4) discussed the issue of which lexical items to include or exclude in the analysis of lexical cohesion. It showed that nouns may be more likely to participate in substantive cohesive relations than verbs. Also, special purpose vocabulary may be more likely to contract cohesive relations than general vocabulary. This allows, for example, the exclusion of very generic verbs such as *be* and *make*. This observation has led to issues relating to frequency. That is, which items need to be included in the analysis of lexical

cohesion regarding their frequency? This has also been discussed in Chapter 2 (Section 2.9.4) where I supported Halliday and Hasan's (1976) view that the high frequency of a lexical item in the language reduces its cohesive force. This support is based on my analysis of a number of lexical cohesive networks that connect high frequency items. I noticed that these items do not build up strong connections that contribute to the lexical cohesion of the text. Accordingly, I suggested that these items need to be removed in order to reduce the resulting list of potential lexical cohesion. In this regard, Teich and Fankhauser (2004) point out that we need to keep words that demonstrate what the text is about. Nonetheless, it is fundamental to stress that in any study that deals with lexis, it is not possible to put rigid exclusion and inclusion criteria that will cover all lexical items that need to be included or excluded in the analysis. However, the criteria that are set below are important to control what should be analysed, and they therefore were applied to each essay in both corpora. The main condition of the lexical items that were included in the analysis is to occur two or more times in the text and create lexical cohesion.

1. Nouns (except proper nouns), verbs, adjectives were included.
2. Adverbs of manner (e.g., *sadly*, *accurately*) were included, whereas most of adverbs of time (e.g., *yesterday*, *today*) and place (e.g., *nearby*, *upstairs*) were excluded. Besides, most frequency adverbs such as *sometimes*, *often*, *always* were excluded. However, adverbs such as *frequently* were considered when they had derived word-forms to connect with such as *frequent* and *frequency*. Furthermore, unmarked adverbs such as (*a lot*, *well*, *enough*, *far*) were excluded.

3. Function words such as determiners, pronouns, prepositions, auxiliaries, negatives, conjunctions were excluded.
4. Determiners and quantifiers such as *all, same, another, anybody, anything, enough, everybody, everyone, everything, other, others, somebody, something, few, little, much, more, many* were excluded.
5. Numbers: ordinals (e.g., *first, last, sixth*), and cardinals such as *one* were excluded.
6. Substitution links such as nominal substitutes (*one/ones, the same, so*), and verbal substitutes such as (*do, be, have, do the same, likewise, do so, do that*) were excluded.
7. Nouns that are part of a discourse marker phrase when they function as conjunctions in the text were eliminated, such as *hand in on the one hand*.
8. High frequency items such as *get, make, put, become, come, go, give, take, know, tell, say, think, feel, seem, want, like, try, find, show, keep, bring, day, good* were excluded.
9. Repetition within T-units was ignored, because it does not contribute to the organising function of lexical cohesion of texts.

4.10 Co-reference as criterion in lexical cohesion analysis

There is a debate among researchers as to whether co-reference criterion is indispensable in establishing lexical cohesion. The term ‘co-reference’ means that two or more linguistic items refer to the same entity and share the same referent in their text (e.g., *the pig...the creature*).

Kunz (2010) describes that co-reference constructs a relation between certain linguistic elements in different clauses, sentences or even paragraphs. Sometimes when the text is longer than a group of sentences, it contains more than one co-reference relations and this is called a ‘co-reference chain’ (e.g., *John...he...the lad...him*). The first mention of the linguistic element in this chain is called ‘antecedent’, while the same referent for the second or more times is called ‘anaphor’. The anaphors normally comprise of some lexico-grammatical items which establish a textual tie to their antecedent in the text. These items are called cohesive devices. Kunz (2010) explains that co-reference is used by writers to link textual parts in a way that the reader or the listener of the text can build a conceptually and semantically coherent unit.

The main point to be raised in this section is whether the two repeated items in a co-reference relation need to refer to the same entity. Halliday and Hasan (1976: 281) claim that the clearest way to know if the referent is the same or not is by using a reference item such as the definite article ‘the’ or a ‘demonstrative pronoun’ with the anaphor. Nevertheless, they find that cohesive devices are not restricted to establishing co-reference. Halliday and Hasan (1976: 283) admit that a “lexical item coheres with a preceding occurrence of the same item whether or not the two have the same referent, or indeed whether or not there is any referential relationship between them”. Halliday (1985: 310) illustrates this claim with the following example.

(4) Algy met a *bear*. *Bears* are bulgy.

In Example (4), *bears* in the second sentence means ‘all bears’, and no definite article or other reference item are attached to *bears*, but Halliday (1985) argues that *bears* would still

establish a lexical cohesive relationship with *bear*. Furthermore, Halliday and Hasan (1976) notice that there are forms of lexical cohesion, other than identical repetition (simple repetition in the present study), that are not dependent on identity of reference. Instead, such forms, as Halliday and Hasan (1976: 282) claim, are cohesively related on their own, giving an independent and purely lexical dimension of internal cohesion to a text. Halliday and Hasan (1976) illustrate this case of cohesion with two lexical items, which are (*boy*) and (*girl*), and they contest that these items have by no means the same referent, yet their proximity in a discourse being opposites contribute to cohesion. Likewise, McCarthy (1991) observes that it is not always the case where two lexical items share the same referent. He compares two examples where in the first one (Example 5) *commence* and *begin* both refer to the same thing in the actual world. In contrast, *commence* and *begin* refer to separate events in Example (6), but McCarthy (1991) argues that these items could still display a stylistic relationship, perhaps to create dry humour or irony. This suggests that lexical items would establish cohesion with each other irrespective of whether or not they have a common referent.

(5) The meeting *commenced* at six thirty. But from the moment it *began*, it was clear that all was not well.

(6) The meeting *commenced* at six thirty; the storm *began* at eight.

(McCarthy 1991: 65)

Hoey (1991b) further draws careful attention to the role of the referential factor in identifying lexical cohesive links. Hoey (1991b: 58) establishes a group of connected flowcharts for this purpose to test whether or not there is a co-reference link between lexical items. He suggests that it is significant to provide a justification for any repetition link between a pair of lexical

items even if the practical purpose of the analysis does not entail such justification. He claims that in some contexts, the presence of ‘common reference’ helps to differentiate text-forming repetition from ‘chance repetition’. Hoey (1991b: 56) defines ‘chance repetition’ as “repetition where the only common ground is the choice of the same lexical item”.

Hoey’s (1991b) definition of chance repetition refers presumably to ‘homonyms’. Murphy (2003: 97) defines ‘homonymy’ as a phonological (or orthographical) item that has two or more meanings, which are not related semantically. For example, *bank* refers to the edge of a river (e.g., *river bank*), and can also refer to a financial institution (e.g., *savings bank*), or, as a verb, to describe the movement of an aircraft (Mahlberg 2005). Hoey (1991b) proposes that common reference leads us to judge that a pair of lexical items does not establish a chance repetition if the items in question refer to the same ‘object’. Nevertheless, he observes that this does not apply to all lexical items, and it is not a standard for items to form a chance repetition if they have different referent(s) in their text.

Hoey (1991b: 56) points out that “lexical items with the grammar of verbs or adjectives cannot, after all, be conveniently said to refer at all”. Hoey (2005: 118) further explains that there are few items which are evaluative or having weak denotations (e.g., *asinine*, *ridiculous*, *racist*, *make*, *action*) and are still primed for cohesion and participating in cohesive chains. *Asinine*, for example, enters into a cohesive chain of near-synonymous (*stupid*, *asinine*, *inaneities*, *stupidity*, *imbecilities*, *inanity*) (Hoey 2005: 120). Kunz (2010) adds that some notions, such as *music*, *love*, *trust*, etc. do not seem to designate any physical object in reality compared to items like *chair*, for instance. She further refers to the nominalised forms of verbs and adjectives (e.g., *development*, *cleverness*), which hardly indicate to anything as

concrete as the reference of nouns. She claims that such items are treated as non-referential because they do not point to things in the world, but to concepts only, and as such, the co-reference criterion is hard to work all the time.

Hoey (1991b) adds another criterion to figure out text-forming from chance repetition which is the 'contextual criterion'. That is, as Hoey (1991b) defines, the context of the lexical items has to indicate that the repeated items have common or related contexts, or share common relationships with neighbouring lexical items. Example (7) illustrates this criterion. In this example, Hoey (1991b) observes that the two lexical items *give* and *giving* do not form a repetition link because they do not share a common context with each other and the paraphrase of one is not the paraphrase of the other. According to their context, the two items do not mean the same. The first item *gives* means 'provides with', whereas *giving* means 'administering to', so the items fail the criteria of repetition.

- (7) After one bear, known to be a peaceable animal, killed and ate a camper in an unprovoked attack, scientists discovered it had been tranquillized 11 times with phencyclidine, or 'angel dust', which causes hallucinations and sometimes *gives* the user an irrational feeling of destructive power. [...] Although some biologists deny that the mind-altering drug was responsible for uncharacteristic behaviour of this particular bear, no research has been done into the effects of *giving* grizzly bears or other mammals repeated doses of phencyclidine.

(Hoey 1991b: 52)

In a long text, however, Hoey (1991b) asserts that the contextual criterion is hard to operate principally if particular items are repeated with high frequency throughout the text. Hoey (1991b) admits that such a criterion might be useful in manual analysis, but it is not effective

for automatic analysis. However, he affirms that even in manual analysis where the texts are long, ensuring that co-reference repetition exists is not practical or even impossible.

According to the different views on the co-reference factor, I agree that it is sometimes not possible to judge if an individual item has the same referent or not particularly if the text is long, or as researchers mentioned above that some items are non-referential at all. Hoey (1991b) acknowledges that absolute identity of meaning is not mandatory and even the type of simple repetition may be partial but not total repetition. As Rimmon-Kenan (1980: 152-153) points out, “even when the whole sign is repeated, difference is introduced through the very fact of repetition, the accumulation of significance it entails, and the change effected by the different context in which it is placed”. Hoey (1991b: 56) argues that the notion of common reference is “neither a necessary property of lexical repetition nor particularly frequent, at least in non-narrative text”. Furthermore, he adds that the distinction between text-forming and chance repetition is not required since the latter will normally be excluded from the analysis at the stage of linking lexical items together to form bonds, and stray doubtful cases of lexical items will not be picked up once these bonds are identified. The following example is taken from an essay in the LOCNESS corpus, where the student repeats the lexical item *exist* more than once in their essay. T-units in (8 a-b) are not adjacent in the original text.

(8-a) The belief *exists* that the computer age [...] will remove the challenges from life.

(8-b) The *existence* of computers is very much dependent on the *existence* of humans.

In Example (8 a-b), the repeated items *exist* – *existence* – *existence* do not share the same referent in their context. In Example (8-a), *exists* refers to the item *belief*, whereas *existence* is an ‘abstract noun’ followed by a prepositional phrase in Example (8-b): ‘The *existence* of computers’. The prepositional phrase determines what abstract nouns relate to because abstract nouns have weak denotations and cannot stand by themselves. Therefore, the first mention of *existence* in (8-b) refers to ‘computers’. I will count the relationship between *exist* – *existence* as derived repetition. My data is full of cases such as this one and it is not sensible to discard these examples from the count of lexical cohesion. This decision has been satisfactorily justified according to the arguments that different researchers presented in this section (e.g., McCarthy 1991; Hoey 1991b, 2005). The analysis of lexical cohesion in the present study did not completely ignore the co-referentiality factor. After a wordlist was run, I paid attention, for example, to instances of homonymous lexical items. I removed items that have the same forms but carry unrelated meanings. Example (9 a-b) below is taken from LOCNESS and illustrates a case of homonymous items.

(9-a) **1** Sharing cars seems feasible, but is unpopular for various reasons. // **2** It removes the *element* of independence afforded by a car [...].

(9-b) **6** People walking or using a bike are not protected from the *elements* and, until cars and buses are banned, have to breathe everybody else’s exhaust fumes.

In Example (9), the lexical item *element* appears two times in the same essay. In (9-a), *element* means ‘factor or aspect’, whereas in (9-b), *elements* refers to ‘climate conditions’. Therefore, in this case both items were excluded from the analysis.

4.11 Conclusion

This chapter has introduced the Arab Learner English Corpus (ALEC), which is a new corpus of argumentative writing that I created for the present study. Among other learner corpora, I have argued that LOCNESS is the closest corpus to be used as reference point to ALEC. The comparison between these two corpora aims to highlight features of interlanguage cohesion between native and non-native writers. Although both corpora vary in a number of dimensions, they are matched in terms of text genre in that both corpora contain argumentative writing. The second part of this chapter has stressed the advantages of combining quantitative and qualitative approaches to the analysis of lexical cohesion, providing the basic framework of analysis that is adopted in the present study. This mixed-methods approach involves integrating a corpus-based approach with text analysis in order to capture the different types of lexical cohesion, and to analyse lexical cohesion on the paradigmatic and the syntagmatic axes. Such an integration will help understand how lexical cohesion functions in native and non-native writing. The chapter further advocates using the T-unit as a base for cohesion analysis in order to avoid the problem of punctuation errors in L2 writing. The chapter ends by arguing that a co-reference criterion is not a necessary feature for studying lexical cohesion. This is because some lexical items still establish cohesive relationships with other items despite the fact that they are not co-referents.

Chapter 5

Quantitative analysis and results: Corpus-based analysis and text analysis

5.1 Introduction

Chapter 4 showed that the quantitative analysis of lexical cohesion will combine corpus analysis and text analysis. This combination is important because categories of lexical cohesion range in their complexity between lexical and textual, which means that not all cohesive devices are amenable to automatic analysis. Lexical relations such as simple repetition and derivatives can be conveniently identified with the assistance of corpus tools (Thornbury 2010). However, cohesive relations that are textual such as signalling nouns require qualitative textual analysis. This observation is also evidenced in a project that was led by Hoey and Collier. This project aims to develop Hoey's (1991b) model of lexical repetition into a working software suite, which can be used to provide a summary for electronic texts. Collier (1994) admits that the software is able to incorporate the two simplest forms of repetition, simple and complex, but not other textual links between, for instance, *John Major* and *Prime Minister*. Instead, such relations and other simple thesaural links are added manually.

Based on the distinction between lexical and textual cohesive relationships, the present chapter will, firstly, deal with the analysis of simple and derived repetition applying Lexical Repetition Network Model (LRNetM), which I suggested as a method of analysis in this study. The chapter will start with an introduction to this model with reference to the classic models of lexical cohesion. In the LRNetM model, lexical cohesion is viewed as a 'network'. Section 5.3 will therefore provide a definition of a 'network' as it is used in the present study,

identifying its characteristics and comparing it with similar approaches to lexical networks in the literature. Section 5.4 explains the analytical steps that are applied to group the wordlist of each essay in both corpora into lexical repetition networks. Lexical repetition networks are of different types. Section 5.5 explains the three types that will be the focus of the present study. Each type contains simple and derived repetitions. Accordingly, Section 5.6 identifies three elements that will be used to quantify instances of simple and derived repetitions in lexical repetition networks in each essay in ALEC and LOCNESS. Section 5.7 illustrates a detailed analysis of one student essay applying LRNetM and the classic models of lexical cohesion. This section shows how frequency of lexical cohesion depends on the method of analysis that is applied to analyse the data. The second phase of the quantitative analysis which examines signalling nouns is described in Section 5.8. The chapter will subsequently present the quantitative results of the frequency information gathered from ALEC and LOCNESS (Section 5.9) to answer the first main research question on the frequency of simple repetition, derived repetition and signalling nouns.

5.2 The Lexical Repetition Network Model (LRNetM)

Traditional models of lexical cohesion are not principally designed to compare frequency counts of lexical cohesion in complete texts from learner writing. Rather, these models identify cohesion in individual texts of native English. These models describe lexical cohesion as directional and co-referential. This means, as discussed in Chapter 2, that the method of analysis that is applied for these models entails tracking each cohesive item with its antecedent and both lexical items need to share the same referent. What helps the application of these models to text analysis is the fact that they analyse short texts, which makes lexical cohesive relations span over adjacent sentences, and hence the co-referential factor is more

manageable than it tends to be when dealing with longer texts. However, the method of analysis that is based on these principles is not easy to replicate in practice to analyse a large-scale corpus of learner writing. Hoey's (1991b) suggestion of building a repetition matrix is not only time-consuming but still prone to subjectivity as the matrix is mainly a display format but not a solution of the analysis.

In Chapter 2, I argued that lexical cohesion needs to be analysed separately from the directionality that the conventional literature on lexical cohesion requires, and I therefore proposed my model of 'Lexical Repetition Network'. The LRNetM model does not describe lexical cohesion as lexical ties, but is based on a new perspective of cohesion derived from corpus linguistics. That is, "cohesive properties of the word are built into the word itself" Hoey (2005: 122). Hoey's assumption has been informed by previous work such as Emmott (1989) and Sinclair (1993, 2004). Sinclair (1993) claims that the whole of the previous text is encapsulated in the sentence currently being read. Emmott and Sinclair, as Hoey (2005: 122) points out, throw new light on the literature on cohesion by stressing that lexical cohesion is not created through back-reference, as the previous literature on cohesion has suggested. Hoey (2005) argues that the reason why we do not refer back is that we are primed to expect the cohesion of particular types for particular items and therefore anticipate its occurrence in advance of its appearance.

Hoey's (2005) proposition does not entail tracing back chains or networks of cohesive ties, but it involves concentrating on the lexical item in question and predicting its potential chain or network of similar items beforehand. If this new perspective on cohesion is taken into account, it becomes possible to track lexical cohesion using corpus tools. As demonstrated in

Chapter 3, some researchers (e.g., Cheng 2009) suggest using frequency lists as a starting point to reveal potential sources of lexical cohesion. My model of LRNetM is based on a frequency/wordlist (see Section 5.4), where each item represents a starting point for cohesive networks, or in Hoey's (2005) terms each item is primed to participate in cohesive chains (or links) or to avoid them. In the present study, the 'prime' item indicates the start item in the lexical repetition network that activates other items to be produced throughout the text creating a network of lexical cohesion. The principal idea of LRNetM is to group lexical items that are possible candidates of lexical cohesion together in one network in terms of their repetition category. Direction, according to LRNetM, is not a key factor in how these cohesive members of a network are grouped. Table 10 summarises a number of the main differences between the LRNetM model and the classic models of cohesion.

Model Characteristics	Halliday & Hasan's (1976) model	Hoey's (1991b) model	LRNetM
Purpose of lexical cohesion analysis	Classification of cohesive ties and presenting a new coding scheme of cohesion analysis	Producing readable abridgements of text	Comparing NS and NNS writing
Type of data analysed	NS literary texts (e.g., Alice in Wonderland) and conversation	NS Scientific articles	NS-NNS argumentative essays
Directionality of lexical cohesive relationships	Uni-directional: Chain-like (linear form of a text)	Multi-directional: Web or Net-like (non-linear form of a text with a systematic direction).	Non-directional: Network-like (non-linear form of a text which is displayed through a wordlist, but the direction of cohesive relationships is not important).
Type of analysis	Manual analysis	Manual analysis	automatic means along with a manual analysis

Table 10 A comparison between the LRNetM model and the classic models of lexical cohesion

5.3 The definition of ‘network’

Lexical cohesion has been viewed in the literature as forming a pattern of chains or a network. A number of researchers (e.g., Hasan 1984, Morris & Hirst 1991, Teich and Fankhauser 2004) analyse lexical cohesion as chains. Teich and Fankhauser (2004), for example, analyse lexical cohesion based on lexical chains by using WordNet, which is an electronic lexical database of English and functions as a thesaurus, because it groups lexical items with related meanings in one semantic set. They define ‘chains’ as consisting of words that are connected with each other through lexical cohesive ties. The term ‘chain’ entails sequence and directionality in that every lexical item connects to the previous one and so forth in a brick-to-brick system, as Hoey (1983) describes. However, cohesion is created in a more complex way than merely a sequence of lexical relations. Other researchers, therefore, prefer to analyse lexical cohesion as a ‘network’ to stress that cohesion is established by interlocking different patterns of lexical repetition. These researchers apply the term ‘network’ differently and for different purposes. For example, Hoey (1983) uses ‘network’ to describe the web-like interconnectedness of clause relations, and Trabasso et al. (1984) use the same term to describe webs of causal connections in narratives. Hoey (1991b) uses the term ‘net’ as opposed to ‘network’ to describe the non-linear complexity of connected sentences through lexical repetition patterns. But both terms imply the property of interconnectedness in lexical cohesion. The definition below describes a ‘network’ as it is used in my model (LRNetM).

A network is a set of two or more lexical items that are grouped together because they represent threads of continuity and signify repetitive units of text. This repetition is due to simple or derived repetition. The lexical items in the lexical repetition network move out from the prime item, and are interconnected with each other in an unpredictable direction, thus forming a network of lexical cohesion.

Sections 5.3.1 and 5.3.2 will describe the specific characteristics of a lexical repetition network, and compare this approach to others as described in the literature.

5.3.1 Characteristics of a lexical repetition network within LRNetM

Typically, any text consists of an array of lexical repetition networks. However, the focus of this study is only on networks that are made up of simple or derived repetitions. In the definition of a network, lexical items that a writer produces in their writing are an activation of a prime item. The writer might produce these lexical items consciously or unconsciously. These lexical items interconnect with each other forming a number of lexical repetition networks. Each network of these has a prime as one of its candidates. This idea of the prime item is inspired by Hoey (2005), who points out that ‘priming’ is a psychological concept but it is used slightly different from how ‘priming’ is used in psycholinguistic studies, where the focus is not on the priming item itself but on the associative relationship between the prime item and the target item. Hoey’s (2005) use of priming is concerned only with the characteristics of the prime item itself. He defines ‘priming’ as “a property of the word (item) and what is primed to occur is seen as shedding light upon the priming item rather than the other way round.” Hoey (2005: 8) (cf. Section 2.9.4).

With reference to cohesion, Hoey (2005) uses the notion of ‘priming for cohesion’, which explains how text is cohesively organised. He finds a psychological explanation for the operation of cohesion through the psychological concept of priming. As explained in Chapter 2, Hoey (2005) claims that the lexical item has cohesive properties that are inbuilt in it from the moment of their selection and we are primed to anticipate cohesion of specific types for specific items and thus predict its occurrence in advance of its appearance. Hoey (2005: 116)

introduces the concept of ‘textual collocation’ and points out that “textual collocation is what lexis is primed for and the effect of the activation of this priming is textual cohesion”. Hoey (2005) identifies many types of textual collocation but what is relevant to the present study is the type of textual collocation that Hoey (2005) describes as the collocation of a lexical item with itself which results in cohesion by repetition. Hoey (2005: 118) illustrates this kind of priming with the lexical item *army* when it occurs in *Guardian* news articles. He examined a sample of 65 different texts, which contained the item *army*. By consulting a concordance of *army* and looking at the cohesion of *army* in every text from which a concordance line is drawn. Hoey’s (2005) examination revealed that of the 65 texts examined, 23 used the instance of *army* within a cohesive chain which contained more than two instances of the same item, and 15 used *army* in a cohesive link (i.e. two instances), which means 58 per cent of the texts that contained the item *army* used it cohesively; with only a few exceptions a text that did not use *army* cohesively contained only one instance of the item. Hoey (2005) points out that the calculation can alternatively be done from the perspective of the item rather than the text. He observed that by focusing on the lexical item, 81 per cent of all instances of *army* examined were found to be used cohesively. The concordance evidence suggests therefore that *army* is typically primed for participation in cohesive chains of simple repetition in newspaper articles (e.g., *army ... army ... army ...army*). Hoey (2005: 122) provided another example of the lexical item *planet*, and he noticed that when *planet* is used in the domain of solar system; it is primed to occur in chains of simple repetition and hyponyms (e.g., *planet – Uranus – Saturn – Planets – Pluto*). He points out that everyone would expect in advance that a text that contains *planet* in the first sentence to contain a cohesive network stemming from the first occurrence made up of instances like the ones given above.

Even though the term ‘prime’ that I use in my definition of a network is a psychological concept and it might have a role in interpreting the cohesive behaviour of students’ writing in certain examples, it does not imply any psycholinguistic theories. The analysis in this study focuses primarily on surface features of text without going into any cognitive background. The main reason for selecting the term ‘prime’ is to indicate its relevance to Hoey’s (2005) work. In this context, it is also important to stress that the idea of a network in LRNetM does not presuppose the underlying coherence of a text. It is only concerned with capturing elements on the surface of a text without tracking the cognitive behaviour of students. The prime item is also not concerned with specifying the direction of a cohesive relationship. As pointed out in Section 5.2, the LRNetM model assumes that it is not easy to examine which item refers to which. However, studies on eye movements in psycholinguistics can check the direction of back-word references, but these studies are concerned with mental representations of the text while the present study considers only linguistic features at the discourse level. I therefore postulate that the analyst of a text cannot be certain whether the writer connects this lexical item to that unless there is a structural clue in the sentence that indicates this cohesive relationship. However, this clue is not always available. The analyst of a written text also cannot judge whether the writer intended to connect this item to the previous one or to another one in another sentence. This is particularly true if a cohesive relationship is established in non-adjacent pairs of sentences, or it is excessively created from simple or derived repetitions where the same item is repeated many times over the text – or in the case where cohesion is made up of non-referential items such as evaluative adjectives or verbs. With such types of cohesion, it is challenging to identify the direction of cohesion. Such a use of network in this sense is, therefore, in contrast with Halliday and Hasan (1976) and Hoey (1991b), who identify that lexical items refer to determined referents either forward or backward. We could

infer from these researchers through their directionality concept that they consider referential boundaries of lexical items are fully determined.

5.3.2 Lexical repetition networks compared to similar lexical networks in existing literature

Without explicitly focusing on the concept of cohesion, researchers such as Williams (1998: 156) utilised ‘network’ to study collocations in terms of collocational networks. He defined a ‘collocational network’ as a network that is used to “signify a web of interlocking conceptual clusters realised in the form of words linked through the process of collocation”. McEnery (2006: 18-19) also applied the term ‘network’ when he studied collocation and claimed that, “we will find words which establish networks of collocation. Within these networks, the items which attract most collocates to them [...] are called nuclear nodes”. Collocational networks as defined by these researchers are relatively close in their function to lexical repetition networks in the LRNetM model. In both networks, the text is grouped into a network of repeated lexical associations, which are key items in a text in terms of their frequency of occurrence. The purpose of these networks is to create the aboutness of a text and provide us with an insight into important lexical connections in discourse. At the level of cohesion, the items in the lexical repetition networks collocate with themselves and can therefore be considered repetitions rather than members of a collocational network. On the other hand, collocational networks achieve collocational cohesion through their regular co-occurrence with other lexical items.

There are also structural similarities; both networks consist of initial look-up forms: the ‘prime’ item in the lexical repetition networks, and the ‘node’ in collocational networks. However, the lexical repetition networks in LRNetM move out from the prime item and

consist of lexical repetitions that interlink together around the prime item, whereas collocational networks evolve from a central node and include significant item collocations of both node and collocate. Besides, the significance of lexical items in these networks is not measured similarly. For example, the significance between the node and its collocates in the collocational networks is measured on the basis of the strength of the association of collocate with a given node by employing the MI (mutual information) statistical test. Nevertheless, the significance of lexical items in lexical repetition networks does not depend on statistical measures. It is based on selecting items that occur with certain frequency in a wordlist. The collocational network is also more complex in its formation than the lexical repetition network in LRNetM. The collocational network is dynamic in that it is continuing to grow in size because it includes collocations not only of the node-item, but also of the collocates of a node, which are considered as nodes in their own right allowing the network to expand. In contrast, the lexical repetition network is fairly static. It reaches its end when there are no repeated items left in the text. However, these repeated items with their derived forms could be looked into more deeply if we study their collocates, which could complement the picture of lexical cohesion analysis. For example, if a text contains repeated items such as *calculate*, *calculation* and *calculators*. This is in itself a form of lexical repetition cohesion. These items could also create collocational cohesion if we extend the analysis and study collocates of these items, either in individual texts or in the complete corpus as one discourse community. This analysis has not been conducted in the present work, but it could be taken as a direction for future research.

5.4 Grouping a wordlist into lexical repetition networks

In the present study, the starting point for the analysis of lexical cohesion is to generate wordlists for each essay in both corpora (ALEC and LOCNESS). However, wordlists cannot be used straightaway to provide counts of the different categories of lexical cohesion. For example, if we have the lexical item *computerise* that is repeated 6 times in the wordlist for a particular text, we will record these instances as simple repetition. This is because wordlists are good at determining simple repetitions. However, the wordlist might have other derived forms of *computerise* (e.g., *computers*, *computation*) which need to be taken into account in the present study. This is where LRNetM becomes operative to group the wordlist into closely related lexical networks of repetition, including these word-forms of derived repetition. Therefore, I ordered the wordlist alphabetically, which facilitated the task of grouping lexical items with similar inflectional and derivational change into one network. As Thornbury (2010: 279) points out, a list of words alone could not be a semantic network, but that it presents the raw data that can assist in mapping out a potential cohesive network. The alphabetical listing also shows word frequency, which is a key factor in revealing a text's lexical repetitions, which are a primary concern of the analysis in the present study because they serve good indicators of lexical cohesion.

It is important to point out that the method of analysis which is adopted in this study analyses each essay in each corpus individually. This means that each corpus is not analysed as one entity (i.e. one discourse community) as is typically done in corpus-based studies. Nonetheless, the separate analyses of the single texts in each corpus will be integrated in order to be able to make general statements about the frequency use of lexical cohesion in each corpus. The following analytical steps explain how a wordlist of a single essay was grouped

into lexical repetition networks. Grouping a wordlist is fundamental for the quantitative analysis, which will be discussed in Section 5.6. These steps were applied to each individual essay in both corpora.

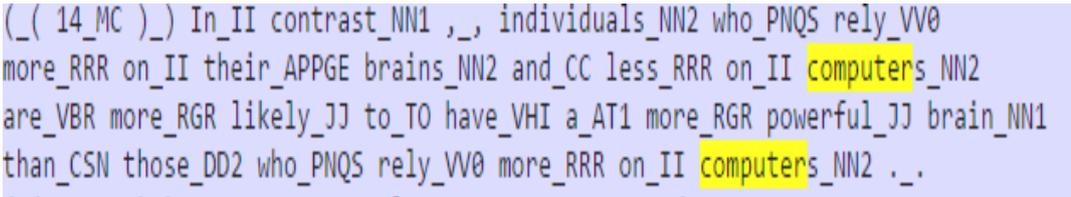
1. As mentioned in Chapter 4, I suggested dividing each essay into T-units and numbering them.
2. I used *WordSmith tools* (version 6.0) for this analysis. I created my own prepared stop-list as a linguistic filter to eliminate any unwanted items such as function words and high frequency words (see Chapter 4, Section 4.9 for the list of inclusion and exclusion criteria).
3. I then generated a wordlist for the essay under analysis and displayed the list alphabetically.
4. Giving that I am dealing with an open set of lexical items, the stop-list cannot exclude all items I need to delete. Therefore, I did more filtering for lexical items that were left in the wordlist and were unwanted for the analysis. For example, lexical items that appeared once and had no other lexical items to combine with were deleted. The alphabetical order of the wordlist helped to observe this clearly.

The frequency cut off point could have been set up at the beginning by adjusting it to 2 in the software before I generated the wordlist. However, doing this would have deleted items that occur once but have other items to relate with. For example, if I have the lexical item *calculate* that occurs only once in the wordlist, but has other related words to connect with

such as *calculated*, *calculator*, *calculation*, I would remove *calculate* from the list if I had set the minimum frequency at 2. Therefore, the frequency cut-off point was controlled manually by discarding items that occur once with no company. Other unwanted lexical items were further filtered from the wordlist (e.g., nouns which are part of a discourse organiser phrase e.g., *addition* in *in addition*).

5. A further step was to disambiguate lexical items in the wordlist in terms of their word classes, when necessary (e.g., *travel* (v) – *travel* (n)). This step is crucial in any lexical cohesion analysis to ensure that lexical items in the wordlist are disambiguated. This was accomplished through the CLAWS web speech tagger (Garside and Smith 1997). During tagging, I also highlighted lexical items that take place in the same T-unit and hence deleted them from the wordlist. This is because the analysis is only concerned with across T-unit cohesive relations.

Consider the example in Figure 6 below – it illustrates T-unit number 14, which is extracted from an Arab NNS essay. The essay was tagged in terms of word classes, and checked to exclude within T-unit cohesion. The item *computers* appears in the same T-unit, and therefore one of its occurrences has to be deleted from the wordlist. This is a screenshot of how lexical items were highlighted in the free CLAWS Web POS Tagger:



```
( ( 14_MC ) ) In_II contrast_NN1 , , individuals_NN2 who_PNQS rely_VV0  
more_RRR on_II their_APPGE brains_NN2 and_CC less_RRR on_II computers_NN2  
are_VBR more_RGR likely_JJ to_TO have_VHI a_AT1 more_RGR powerful_JJ brain_NN1  
than_CSN those_DD2 who_PNQS rely_VV0 more_RRR on_II computers_NN2 . .
```

Figure 6 A screenshot of the CLAWS Web POS Tagger

6. Once each wordlist was alphabetically ordered and filtered, I saved it as an Excel file and grouped it into lexical networks of simple and derived repetitions. Figure 7 illustrates an extract from the lexical repetition networks in an Arab NNS essay. At this point, the figure does not show the quantitative elements. These elements will be explained in Section 5.6.

Network N.	Lexical item	Freq.
1	ACCIDENT	3
	ACCIDENTS	1
2	ADVANTAGE	1
	ADVANTAGES	1
3	AWARENESS	2
4	BUS	1
	BUSES	1
5	CAR	3
	CARS	3
6	COUNTRIES	2
7	DRIVE	1
	DRIVERS	1
8	EASIER	1
	EASY	1
9	ENCOURAGE	2
10	ENVIRONMENT	1
	ENVIRONMENTAL	1

Figure 7 Sample of lexical repetition networks in an Arab NNS essay

5.5 Types of lexical repetition networks within the LRNetM framework

Different types of lexical repetition networks can be found within a text. For example, Hoey (2005) highlights that a cohesive network may be made up of simple repetition or a combination of simple repetition, co-referential expressions and pro-forms. It could also be made up of near-synonyms of the item but not of repetitions. However, the LRNetM model

focuses on simple and derived repetitions. Therefore, the following types of lexical repetition networks were analysed in the current work:

1. A lexical repetition network of simple repetition (e.g., *children* (6) – *child's* (1))
2. A lexical repetition network of derived repetition (e.g., *technologically* (1) – *technological* (1))
3. A lexical repetition network that is a combination of simple and derived repetition (e.g., *belief* (3) – *believe* (1))

The numbers in the brackets above are raw frequencies of lexical items as they appear in the wordlist. In the third type of the lexical repetition network (i.e. *belief* – *belief* – *belief* – *believe*), *belief* occurs 3 times in the essay under analysis. One of the occurrences of *belief* functions as a signalling noun, as it is illustrated in Example (1) (the SN is italicised, whereas its lexical realisation is underlined).

- (1) **1** Computers have also been blamed for a fall in education standards among primary school children. // **2** This *belief* is accentuated by the results of surveys which state that the vast majority of primary school children are unable to do simple arithmetic in their minds and rely on another computer, the calculator, to help them.
- (LOCNESS)

The item *belief* in T-unit 2 encapsulates the underlined lexical realisation in T-unit 1. Thus, it could be said that the third type of the repetition network consists of simple repetition, derived repetition and signalling nouns. However, for analytical reasons, I counted signalling nouns in

a separate stage (cf. Section 5.8) to avoid overlap with simple and derived repetitions, which can create complexity in counting. Consequently, it would be plausible to consider any network in the text such as the one that is discussed in Example (1) as a network that forms simple repetition and derived repetition while counting signalling nouns separately.

5.6 Counting frequencies of simple and derived repetitions in lexical repetition networks

Every essay from both ALEC and LOCNESS is composed of a certain number of the three types of lexical repetition networks that were presented in Section 5.5. Using wordlists, these networks have been grouped for each essay and also numbered. Figure 8 below shows an essay which has been analysed into its composing repetition networks.

Network N.	Lexical item	Freq.	Total number of tokens in each lexical rep.network (length)	Tokens (freq.) of Simple Rep.	Tokens (freq.) of Derived Rep.
1	ABILITY	3	5	2	3
	ABLE	1			
	UNABLE	1			
1	AREA	1	2	1	0
	AREAS	1			
2	BELIEF	3	4	2	2
	BEILEVE	1			
3	BRAIN	4	4	3	0
4	CALCULATE	1	2	0	2
	CALCULATOR	1			
5	CHALLENGED	1	2	0	2
	CHALLENGES	1			
6	CHILDREN	6	7	6	0
	CHILD'S	1			
7	COMPUTER	8	22	21	0
	COMPUTERS	14			
8	CONTROVERSY	2	2	1	0
9	CREATE	1	5	3	2
	CREATED	3			
	CREATION	1			
10	DEVELOPMENT	2	3	2	0
	DEVELOPMENTS	1			
11	EXISTENCE	2	3	1	2
	EXISTS	1			
12	FEAR (V)	1	2	0	2
	FEARS (N)	1			
13	GAMES	4	4	3	0
14	HUMAN (ADJ)	5	10	8	2
	HUMANS (N)	5			
15	INDUSTRY	2	2	1	0
16	INTERNET	5	5	4	0
17	INTRODUCED	1	2	0	2
	INTRODUCTION	1			
18	LEARN	2	2	1	0
19	LIFE	3	3	2	0
20	MAN	1	2	0	2
	MANKIND	1			
21	MINDS	2	2	1	0
22	PARENTS	2	2	1	0
23	PEOPLE	3	3	2	0
24	PROBLEM	3	3	2	0
25	PROGRAMME (V)	2	2	1	0
26	RELY	2	2	1	0
27	REPLACE	1	2	1	0
	REPLACED	1			
28	REQUIRE	2	2	1	0
29	ROLE	2	2	1	0
30	SCHOOLS	2	2	1	0
31	SOCIETY	1	3	0	3
	ANTISOCIAL	1			
	UNSOCIABLE	1			
32	TECHNOLOGICALLY	1	2	0	2
	TECHNOLOGY	1			
33	VAST	1	2	0	2
	VASTLY	1			
34	WIDESPREAD	2	2	1	0
35	WORK	1	3	1	2
	WORKPLACE	2			
			127	75	30

Figure 8 An illustration of lexical repetition networks in an NS essay from LOCNESS, ranked in alphabetical order

On the basis of this view in Figure 8, we see that each row represents a single lexical repetition network which is part of the essay, while there are 6 columns which include the following elements:

1. The lexical repetition network number (network No.): This number does not indicate any rank order or sequence, but it helps to count total number of networks in each essay.
2. Lexical item(s) that make up lexical repetition networks in the essay under analysis.
3. Frequency of each lexical item as it appears in the original wordlist.
4. Network length for each lexical item (i.e., total number of lexical items participating in each individual network).
5. Tokens (frequencies) of simple repetition in this network.
6. Tokens (frequencies) of derived repetition in this network.

Elements 4, 5 and 6 are the main quantitative measures that were used to count simple and derived repetitions in lexical repetition networks for each essay in both corpora. To illustrate the counting procedure for these three elements, Example (2) presents the same essay as in Figure 8. This essay is from LOCNESS, written by a British native speaker. The essay contains 544 words and 29 T-units. The T-units have been numbered and a mark of (//) has been added when two sentences has been cut into T-units.

- (2) **1** The continual development of computer technology has created a great deal of controversy in modern times.//
- 2** There are nowadays widespread uses for computers in all parts of society.//
- 3** One area which has drawn particular criticism is the computer games industry.//
- 4** Since the early 1980's, when computers such as the Spectrum 48K and the Commodore 64 were introduced into the homes of millions of people, controversy has followed about their effect on children.//
- 5** There is a widespread belief among parents that computer games hinder a child's ability to learn.//
- 6** The development of more sophisticated and technologically advanced computers, such as the Amiga and PC, has served to deepen the "problem".//
- 7** The growing realism of the games appeal strongly to children.//
- 8** However, parents feel that children should be broadening their minds by reading books and that the computer games industry is encouraging children not to learn.//
- 9** Computers have also been blamed for a fall in education standards among primary school children.//
- 10** This belief accentuated by the results of surveys which state that the vast majority of primary school children are unable to do simple arithmetic in their minds and rely on another computer, the calculator, to help them.//
- 11** Another moral dilemma that computers have created is their role in the workplace.//

- 12** There are many people who fear that computers will eventually replace man in the workplace.//
- 13** The role of computers is already significant in employment areas, such as accountancy, which require a specialist ability to be able to calculate and manipulate numbers.//
- 14** These computers have replaced humans //
- 15** and such changes have sparked fears that the level of unemployment could vastly increase as a result.//
- 16** As the computers are: more efficient than humans, do not require payment for their work, are less temperamental than humans and will never have a day off because it is ill, //
- 17** it seems mankind is faced with a great problem.//
- 18** Another claim made that suggest the human brain could become useless is that computers promote anti-social behaviour.//
- 19** The creation of the World Wide Web and the Internet have created concern that the human race will become unsociable.//
- 20** The Internet allows you to correspond with people all over the world //
- 21** and it also contains many other functions to, supposedly, make life easier for us.//
- 22** The Internet also has another fundamental problem according to many.//

- 23** The belief exists that the computer age and its developments, such as the Internet, will remove the challenges from life.//
- 24** The human brain is constantly in need of being challenged to maintain itself; //
- 25** but with the introduction of the Internet, one could conduct their life without ever needing to leave the house which many believe would menace society.//
- 26** However, at the moment, the existence of computers is very much dependent on the existence of humans.//
- 27** No matter how much artificial intelligence a computer may be able to show,//
- 28** it still has to rely on the human brain to programme it and to put it into operation.//
- 29** Although, there are plans to create computers which can programme themselves, (which I, personally, feel is a very dangerous idea) the human brain still very much controls the computer and still the ability to end the existence of computers at any given moment; thankfully, a power computers do not have over humans.

The following sub-sections will identify the counting procedure for the three types of lexical repetition networks that are presented in Section 5.5. As mentioned above, the counting procedure is based mainly on three elements: the length of the lexical repetition network (see Section 5.6.1), tokens (frequencies) of simple repetition in this network (see Section 5.6.2), and tokens (frequencies) of derived repetition in this network (see Section 5.6.3). These elements will be presented according to their sequential order in Figure 8. Section 5.6.4 will then explain the procedure of calculating the total frequencies of each measure of the three.

5.6.1 The length of the lexical repetition network

This measure helps show the dominant (i.e., the longest) networks throughout the text. Teich and Fankhauser (2004) observe that when we look at the concrete words that make up the dominant chains (networks in the present study) in a text, we can notice that they are good indicators of the text's topic. The length of a lexical repetition network is similar to the idea of keyness (Scott and Tribble 2006). Long repetition networks which contain many instances of repetitions work as aboutness indicators by reflecting what the text is about. This is in line with the method of identifying keywords which is also based on repetition. The basic principle is that a word-form which is repeated many times within the text will be more likely to be key in it (Scott and Tribble 2006: 58).

Hoey (2005: 118) also discusses the length factor of a chain/network and argues that the longer a chain is, the more it appears to be related to the topic, whereas cohesive links (i.e., only two members in the chain) tend to be less closely associated with the topic of the text. In this regard, Teich and Fankhauser (2004) further maintain that short chains (with few participating items) have a different function in that they do not signal the text's topic but they glue it together locally. For example in the essay in Figure 8, the dominant (longest) networks are built around *computers*, *humans*, *children*, *ability*, *internet*, *brain* and *games* giving a general idea of the text's topic, i.e. the relationship between computers and humans. We can also predict from these networks that they indicate a sub-topic, which is the effect of the internet and computer games on children's brain abilities. The shorter networks in this essay, made up of two candidates only, are built around, for example, groups of items such as *area*, *challenges*, *role* or *fear*. This confirms that shorter chains/networks are usually built around general vocabulary (Hoey 1991b).

The first column in Figure 8 shows that the NS essay has 35 individual repetition networks. Each one contains simple and/or derived repetition as members of lexical cohesion. The total count of these members in the essay is 127. Although the element of chain/network length is an important measure, which could add further detail to the way lexical cohesion works in native and non-native writing, I did not use it in the present study as a discriminator. But this measure was used to calculate the frequency of derived repetition, as we will see in Section (5.6.3). The network length is useful for a future study that investigates lexical cohesion across registers. Teich and Fankhauser (2005) find that we can classify registers into groups (e.g., LEARNED vs. PRESS) according to the average length of the longest chains in each register.

5.6.2 Frequency of simple repetition in a simple repetition network

Table 11 illustrates an example of a simple repetition network which is taken from Figure 8.

Network No.	Lexical item	Freq.	Total No. of tokens in each LRNet (length)	Tokens (freq.) of SR	Tokens (freq.) of DR
6	CHILDREN	6	7	6	0
	CHILD'S	1			

Table 11 An example of a simple repetition network in an NS essay from LOCNESS

(**LRNet**: Lexical repetition network; **SR**: simple repetition; **DR**: derived repetition)

In Table 11, the simple repetition network is made up of 7 participants (*children – children – children – children – children – children – child's*). This network does not have a derived repetition type as illustrated in the last column in Table 11. For counting the tokens (i.e.

frequency) of simple repetition that are candidates in the simple repetition network, I do not count the first mention of the lexical item in this network. I presume here that the first mention works only as a prime whose function is to activate other items to be produced throughout the text. In addition, Hasan (1984) points out that an element by itself cannot be its own repetition.

Therefore, the procedure that I took to count occurrences of simple repetition in the simple repetition network was to subtract 1 from the total number of tokens in the simple repetition network. It is vital to distinguish at this point between the ‘total number of tokens in the simple repetition network’ and the ‘network length’ (4th column in Table 11). This distinction is important because the network length was not used to calculate the frequency of simple repetition, as it was the case in counting derived repetition (see Section 5.6.3). The total number of tokens in the simple repetition network and the network length are equal in Table 11, only by chance. Therefore, the formula for counting instances of simple repetition in a simple repetition network is as follows:

$$\textit{Tokens (frequency) of simple repetition in a simple repetition network} = \textit{Total number of tokens that make up the simple repetition network} - 1$$

According to this formula, I deduct 1 from the total number of tokens in the whole network of simple repetition, which are 7 (cf. Table 11). This deduction results in having figure 6, which represents the frequency of simple repetition in the simple repetition network.

5.6.3 Frequency of derived repetition in a derived repetition network

The second type of the lexical repetition networks is to group closely related derived forms together in one network. What constitutes derived repetition has already been explained in Chapter 2, Section 2.6.2. It is vital to note that the alphabetical wordlist can only group morphological words that start with the same prefixes. Other derived forms such as the morphologically related antonyms (*legal – illegal*) or the less transparent derivatives (e.g., *surety, assure*) need to be observed manually throughout the wordlist in order to group them with the relevant networks.

The following lexical network, which consists of two lexical items *technologically* (1) – *technology* (1), is a type of derived repetition network. This network is also taken out from the NS essay, which is used in Figure 8. Table 12 presents how to provide counts of derived repetition in the derived repetition networks:

Network No.	Lexical item	Freq.	Total No. of tokens in each LRNet (length)	Tokens (freq.) of SR	Tokens (freq.) of DR
32	TECHNOLOGICALLY	1	2	0	2
	TECHNOLOGY	1			

Table 12 An example of a derived repetition network in an NS essay from LOCNESS

The *technology* network has 2 tokens of derived repetition, which characterise its length. It is a short cohesive network which, according to Hoey (2005: 117), represents a link which is created when two items are connected. It does not contain any type of simple repetition. The prime item in the derived repetition network is included in the count, and the network length

is used to count the frequency (tokens) of derived repetition in the derived repetition network.

The following formula demonstrates the counting procedure:

$$\text{Tokens (frequency) of derived repetition in a derived repetition network} = \text{Total number of tokens in a lexical repetition network (network length)} - \text{Tokens (frequency) of simple repetition in this network}$$

If I apply this formula to the *technology* network, tokens (frequency) of derived repetition will be calculated as follows:

$$\text{Tokens (frequency) of derived repetition} = 2 - 0 = 2$$

The third type of lexical repetition networks is the lexical network that combines simple and derived repetition simultaneously. Table 13 illustrates this type of a network.

Network No.	Lexical item	Freq.	Total No. of tokens in each LRNet (length)	Tokens (freq.) of SR	Tokens (freq.) of DR
2	BELIEF	3	4	2	2
	BELIEVE	1			

Table 13 An example of a lexical network that combines simple & derived repetition in an NS essay in LOCNESS

In Table 13, the lexical repetition network of the lexical item *belief* is composed of four candidates (*belief – belief – belief – believe*). The counting procedure of both simple and derived repetition in this network is the same as explained in Sections 5.6.2 and 5.6.3. For example, the lexical item *belief* occurs three times in the wordlist. Therefore, the frequency of simple repetition is calculated by subtracting 1 from the total number of the item *belief* in the

simple repetition network. Thus, the frequency of simple repetition is 2 because the first mention of the item, as explained in Section 5.6.2, is not counted. For counting the tokens of derived repetition in the *belief* network, the tokens of simple repetition are deducted from the length of the network (i.e. $4 - 2 = 2$). Therefore, the lexical item *believe* establishes with *belief* a derived repetition network (a link).

5.6.4 Calculating the total frequencies of simple and derived repetition in the lexical repetition networks in an individual essay

Next, I created wordlists for each essay in ALEC and LOCNESS, and each wordlist was grouped into lexical repetition networks. Then I applied the counting procedure, which is based on the three elements that were demonstrated in Sections 5.6.1, 5.6.2 and 5.6.3. This helped calculate the different tokens of both simple and derived repetition in each network that composes the essay under analysis. Based on this frequency information, I computed the sum for each element of the three. That is, the network length: the total number of tokens in all lexical networks that make up the essay (column 4 in Figure 8), the total number of tokens (frequencies) of simple repetition in these lexical networks (column 5), and finally the total number of tokens (frequencies) of derived repetition (column 6). The last row in Figure 8 demonstrates the total frequency of each element. This calculation of the overall frequency of the tokens of lexical repetitions in the lexical repetition networks that construct the essay gives overview of the repetition in each individual essay.

5.7 Comparing frequency counts of lexical cohesion between the classic models of lexical cohesion and the lexical repetition network model

In this section, I will show how the classic models of lexical cohesion differ from the LRNetM model in their method of counting. I will also provide textual examples which show how the method of counting by the classic models can allow for variation in the frequency counts of lexical cohesion when it is applied to the same essay. As such, I will analyse an essay which is taken from the ALEC corpus, first applying Halliday and Hasan's (1976) model and then Hoey's (1991b) model. Subsequently, I will apply the LRNetM model to the same essay and compare the frequency counts provided by each model. The principles of lexical cohesion analysis that underlie the two classic models of cohesion have been explained in Chapter 2. In the analysis, the term 'simple repetition' is used in Halliday and Hasan's (1976) model, which indicates the category of 'same item repetition' in their model. It is important to stress that Halliday and Hasan's (1976) model subsumes the category of derived repetition under the category of simple repetition, which is in contrast to Hoey's (1991b) model and LRNetM. Therefore, only the count of simple repetition is provided in the analysis that applies Halliday and Hasan's (1976) framework. Although Halliday and Hasan (1976) and Hoey (1991b) use the sentence as the unit of analysis, I will apply the T-unit to the analysis of lexical cohesion in the three analyses because this will enable a systematic approach when comparing frequency counts by each model. Example (3) introduces the complete essay which is numbered and divided into T-units (the (//) mark indicates the end of each T-unit).

- (3) **1** There are various reasons why I think that as we rely more on computers, our brains will become weaker and more reliant on computers to provide us with quick, efficient answers to most of the complex problems we used to use our brains to solve.//
- 2** In order to maintain the strength and the power of our brains, we need to keep exercising them.//
- 3** The exercises include involvement in solving complex problems, less reliance on calculators and computers to solve mathematical problems and engagement in numerical and non-numerical challenges.//
- 4** However, in this day and age, the computers provide us with easy access to vast amounts of data and information that can be used to easily solve some of our most complex problems.//
- 5** The easy access to the internet meant that people are becoming increasingly reliant on their computers to provide them with reliable and efficient answers rather than their own brains.//
- 6** For example, when I was doing my Bachelor degree at the University of Derby, I had to write up various technical reports about some complex mechanical problems.//
- 7** Instead of trying to solve these problems using the power of my own brain, I instead relied on my computer and the access to the internet to identify projects that are relevant or closely related to the ones I was working on.//
- 8** The accessibility to such information made it easier for me to identify reliable solutions to the problems I was trying to solve.//
- 9** If it was not for the computer and for the internet, I would have had to use the power of my own brain to solve those problems.//
- 10** To be honest, I think the computers are making us lazy.//

- 11 Everything we need and everything that we may want to do is often a click-of-a-button away.//
- 12 These days, I sometimes find it difficult to even conduct simple mathematical equations using the power of my brain.//
- 13 Instead, I rely mainly on the Excel spreadsheets to help me to swiftly calculate the multiples or the sums of the numbers I have in hand.//
- 14 In contrast, individuals who rely more on their brains and less on computers are more likely to have a more powerful brain than those who rely more on computers.//
- 15 For example, one day I was trying to make a simple mathematical calculation,//
- 16 but I struggled and took me awhile to come up with an accurate number, while my younger nephew was quicker in terms of finding the outcome of the multiples I was trying to calculate.//
- 17 This is not a matter of me being older or him being younger, //
- 18 but it is more of the fact that I rely more on computers and calculators in my life than he does.//
- 19 He had far less access to computers and to the internet than me.//
- 20 Therefore, he has to rely on his brain at all times to find solutions to all problems; numerical and non-numerical, complex and simple.//
- 21 This proves that the less to rely on computers, the more exercise your brain gets and the more powerful your brain becomes.//

The analysis of the essay progresses in three steps as follows:

1. A text analysis using Halliday and Hasan's (1976) method of analysis is provided (see Chapter 2 for an explanation of Halliday and Hasan's (1976) principles of analysis). Only the first six T-units are presented for an illustration of the counting method (see Appendix 6 for a coding of lexical repetitions for the full essay applying Halliday & Hasan's (1976) model).

T-unit No.	No. of ties	Cohesive item	Type	Distance	Presupposed item
2	1	brains	Simple Rep.	0	brains
3	8	exercises (n) complex problems X 2 reliance computers solving solve	Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep.	0 1 1 1 1 1 1	exercising (v) complex problems rely computers solve solve
4	5	computers provide solve complex problems	Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep.	0 2 0 0 0	computers provide solve complex problems
5	9	easy access reliant reliable computers provide efficient answers brains	Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep.	0 0 1 1 0 0 3 3 2	Easy access reliance reliance computers provide efficient answers brains
6	3	various complex problems	Simple Rep. Simple Rep. Simple Rep.	4 1 1	various complex problems

Table 14 An analysis of lexical cohesion in an NNS essay from ALEC applying Halliday & Hasan's (1976) model

2. Hoey's (1991b) repetition matrix is applied to analyse the same essay. Again, only the first six T-units of the essay are presented here. This is for the complexity of inserting a complete

matrix due to its giant size. A complete repetition matrix for the full essay can be found in the CD-ROM enclosed in Appendix 7.

		1		
2	<i>sr</i> : brain-brains		2	
	<i>sr</i> : solve-solve <i>sr</i> : complex-complex <i>sr</i> : problems-problems <i>dr</i> : rely-reliance			
3	<i>sr</i> : computers-computers	<i>dr</i> : exercising-exercises	3	
	<i>sr</i> : solve-solve <i>sr</i> : complex-complex <i>sr</i> : problems-problems <i>sr</i> : provide-provide		<i>sr</i> : solving-solve <i>sr</i> : complex-complx <i>sr</i> : problems-problems	
4	<i>sr</i> : computers-computers			4
	<i>sr</i> : reliant-reliant <i>sr</i> : computers-computers <i>sr</i> : provide-provide <i>sr</i> : efficient-effienct <i>sr</i> : answers-answers		<i>dr</i> : reliance-reliant <i>sr</i> : computers-computers	<i>sr</i> : computers-computers <i>sr</i> : easy-easy <i>sr</i> : access-access <i>sr</i> : provide-provide
5	<i>sr</i> : brains-brains	<i>sr</i> : brains-brains		
	<i>sr</i> : complex-complex <i>sr</i> : problems-problems		<i>sr</i> : complex-complx <i>sr</i> : problems-problems	<i>sr</i> : complex-complex <i>sr</i> : problems-problems
6	<i>sr</i> : various-various			

Figure 9 An analysis of lexical cohesion in an NNS essay from ALEC applying Hoey's (1991b) repetition matrix (Sample coding: *sr* = simple repetition; *dr* = derived repetition)

3. A text analysis using the LRNetM model is carried out to analyse the same essay. Figure 10 below displays the repetition network view of the essay.

Network No.	lexical item	Freq.	Total number of tokens in each lexical rep. nNetwork (length)	Tokens (freq.) of Simple Rep.	Tokens (freq.) of Derived Re
1	ACCESS	4	5	3	2
	ACCESSIBILITY	1			
2	BRAIN	6	11	10	0
	BRAINS	5			
3	CALCULATE	2	5	2	3
	CALCULATION	1			
	CALCULATORS	2			
4	COMPLEX	5	5	4	0
5	COMPUTER	4	14	13	0
	COMPUTERS	10			
6	EASIER	1	4	2	2
	EASY	2			
	EASILY	1			
7	EFFICIENT	2	2	1	0
8	EXERCISE (n)	1	3	1	2
	EXERCISES (n)	1			
	EXERCISING (gerund/v)	1			
9	IDENTIFY	2	2	1	0
10	INFORMATION	2	2	1	0
11	INTERNET	4	4	3	0
12	MATHEMATICAL	3	3	2	0
13	MULTIPLES	2	2	1	0
14	NUMBER	1	6	3	3
	NUMBERS	1			
	NUMERICAL	2			
	NON_NUMERICAL	2			
15	POWER	4	6	4	2
	POWERFUL	2			
16	PROBLEMS	9	9	8	0
17	PROVIDE	3	3	2	0
18	QUICK	1	2	1	0
	QUICKER	1			
19	RELIABLE	2	13	9	4
	RELIANCE	1			
	RELIANT	2			
	RELIED	1			
	RELY	7			
20	SIMPLE	3	3	2	0
21	SOLUTIONS	2	9	7	2
	SOLVE	6			
	SOLVING	1			
22	VARIOUS	2	2	1	0
23	YOUNGER	2	2	1	0
			117	82	20

Figure 10 An analysis of lexical cohesion in a complete essay from ALEC applying the LRNetM model

The following table presents the total frequency counts (i.e., tokens) of simple and derived repetition for the full essay that is presented in Example (3) above.

Category of lexical cohesion	Halliday & Hasan's (1976) model	Hoey's (1991b) model	LRNetM
	Freq.	Freq.	Freq.
Simple rep.	89	209	82
Derived rep.	/	47	20
Total count of lexical repetitions	89	256	102

Table 15 Comparing frequency counts of lexical cohesion applying the classic models of lexical cohesion & the LRNetM model in a complete essay from ALEC

Table 15 shows that the LRNetM model is closer to Halliday & Hasan's (1976) model in the frequency counts of simple repetition (82 vs. 89). However, Halliday & Hasan's (1976) model shows a slight increase in the frequency of simple repetition of around 8% compared to LRNetM. This percentage does not reveal a significant difference between the two models. It is perhaps a result of the fact that Halliday & Hasan (1976) count instances of derived repetition within the category of simple repetition, which in turn increases the total frequency counts of simple repetition. Example (4) below demonstrates how the word *calculate* creates a simple repetition network when Halliday & Hasan's (1976) model is used. I only selected T-units that contain the word *calculate*. Therefore, the sequence of T-units is not a complete essay. Instances of simple and derived repetition in all subsequent examples are in bold.

- (4) **3** The exercises include involvement in solving complex problems, less reliance on **calculators** and computers to solve mathematical problems and engagement in numerical and non-numerical challenges.//

13 Instead, I rely mainly on the Excel spreadsheets to help me to swiftly **calculate** the multiples or the sums of the numbers I have in hand.//

15 For example, one day I was trying to make a simple mathematical **calculation**.//

16 but I struggled and took me awhile to come up with an accurate number, while my younger nephew was quicker in terms of finding the outcome of the multiples I was trying to **calculate**.//

18 but it is more of the fact that I rely more on computers and **calculators** in my life than he does.//

According to Halliday & Hasan's (1976) model, *calculate* in the example above establishes 4 cohesive ties of simple repetition (i.e. same item). I could configure this relationship as follows:

calculate – *calculators* (1 tie – T-unit 13 with T-unit 3)

calculation – *calculate* (1 tie – T-unit 15 with T-unit 13)

calculate – *calculation* (1 tie – T-unit 16 with T-unit 15)

calculators – *calculate* (1 tie – T-unit 18 with T-unit 16)

These counts of derived repetition of the item *calculate* are included in the count of simple repetition, which as a result increased the frequency of simple repetition in Halliday & Hasan's (1976) model. This will not be, however, the case when the LRNetM model is applied to count the same lexical item. According to LRNetM, *calculate* builds up a relationship of derived repetition that will be calculated separately from the count of simple repetition. In Table 5, if I merge the frequency counts of derived repetition with the frequency of simple repetition in LRNetM, the total frequency of simple repetition will be 102 tokens,

which is now higher than that of Halliday & Hasan's (1976) model. However, again, such a difference is not substantial. The frequency counts of simple repetition along with derived repetition in the LRNetM model are higher than Halliday & Hasan's (1976) model around 15% (102 vs. 89).

One reason for this an increase in the frequency counts of the types of repetitions in LRNetM compared to Halliday & Hasan's (1976) model could be attributed to the method of counting adopted by each model. On the one hand, the method which is used to count simple repetition is very similar in both models in that Halliday & Hasan (1976) count lexical ties, which means two cohesive items are counted as one item establishing one tie. Likewise, in the LRNetM model, the first item in the lexical repetition network is excluded from the count. This means that if there are two lexical items in the simple repetition network, only one item is counted. This creates similarity in the count of simple repetition between the two models. On the other hand, the difference in the count between the two models seems to lie in the way of counting derived repetition. The following example demonstrates this difference in counting between the two models.

- (5) **2** In order to maintain the strength and the **power** of our brains, we need to keep exercising them. //
- 7** Instead of trying to solve these problems using the **power** of my own brain, I instead relied on my computer and the access to the internet to identify projects that are relevant or closely related to the ones I was working on.//
- 9** If it was not for the computer and for the internet, I would have had to use the **power** of my own brain to solve those problems.//

12 These days, I sometimes find it difficult to even conduct simple mathematical equations using the **power** of my brain. //

14 In contrast, individuals who rely more on their brains and less on computers are more likely to have a more **powerful** brain than those who rely more on computers.//

21 This proves that the less to rely on computers, the more exercise your brain gets and the more **powerful** your brain becomes.//

The LRNetM model considers the lexical repetition network that contains the lexical item *power* as having 4 tokens of simple repetition while it counts derived repetition by deducting tokens of simple repetition from the network length. The length of the network in Example (5) is 6 (i.e. *power*, *power*, *power*, *power*, *powerful* and *powerful*). Therefore, tokens of derived repetition are 2 (see Figure 10 above). If I assume that simple repetition and derived repetition are counted separately by Halliday & Hasan's (1976) model, the count of simple repetition in Example (5) is 4 ties which is the same count as LRNetM. Conversely, there will be one tie of derived repetition (*powerful* – *power*) according to Halliday & Hasan's (1976) principles of analysis, as compared to two tokens within LRNetM. These cohesive ties of the lexical item *power* are illustrated as follows:

power – *power* (1 tie – T-unit 7 with T-unit 2)

power – *power* (1 tie – T-unit 9 with T-unit 7)

power – *power* (1 tie – T-unit 12 with T-unit 9)

powerful – *power* (1 tie – T-unit 14 with T-unit 12)

powerful – *powerful* (1 tie – T-unit 21 with T-unit 14)

Table 15 above also shows that when Hoey's (1991b) model is applied to the same essay, the figures of both simple and derived repetition go up higher compared to both the LRNetM model and Halliday & Hasan's (1976) model. Hoey's (1991b) model is approximately 61% higher than LRNetM with respect to simple repetition (82 vs. 209), which seems a compelling difference between the two models. Likewise, derived repetition based on the counting method of Hoey's (1991b) model amounts 57% higher than LRNetM (20 vs. 47). This indicates that lexical cohesion according to Hoey's (1991b) method of analysis is more prominent in the essay compared to LRNetM and Halliday and Hasan's (1976) model. Hoey (1991b) believes that lexical cohesion is the dominant mode of creating texture and thus he counts lexical cohesion as multiple links. Sardinha (2001: 216) comments on Hoey's (1991b) study and points out that "by admitting of multiple links between lexical items, the number of ties proliferates, thus increasing the share of lexical ties". Analysing lexical cohesion as multiple links by Hoey (1991b) implies multi-directionality, which means that each lexical item which is part of a cohesive relationship can form a lexical relationship with all previous occurrences in earlier sentences which relate to it. This creates a complex network or web of links in a text, which Hoey (1991b) represented in his repetition matrix. The LRNetM model also analyses lexical cohesion as networks but as stressed in Section 5.2, the concept of a network in LRNetM does not apply the concept of directionality to the analysis of lexical cohesion. A network in LRNetM does not involve drawing lines and webs between lexical items, as in Hoey's (1991b) model. Rather, it requires grouping lexical items that share the same type of lexical cohesion together in one group irrespective of their direction in text.

The frequency counts are fairly close particularly between Halliday & Hasan's (1976) model and the LRNetM model. This indicates that LRNetM is not a replacement of the traditional

models but it suggests a refinement to them on the method of counting of lexical cohesion which allows the frequency of lexical cohesion to be measured systematically. Examples (4) and (5) that I presented above illustrated how the method of analysis that is applied by a model such as Halliday & Hasan's (1976) identifies lexical cohesion differently from the LRNetM model in the essay sample. However, these examples are the simplest form of how the tie works in a text. Therefore, I will show some examples from the essay sample in Example (3) where the counting procedure using Halliday & Hasan's (1976) model can yield different counts if applied by different analysts. Let's examine the following example:

(6) **1** There are various reasons why I think that as we **rely** more on computers, our brains will become weaker and more **reliant** on computers to provide us with quick, efficient answers to most of the complex problems we used to use our brains to solve.//

3 The exercises include involvement in solving complex problems, less **reliance** on calculators and computers to solve mathematical problems and engagement in numerical and non-numerical challenges.//

In Example (6), the lexical items *rely*, *reliant* and *reliance* are repeated in T-units (1) and (3). As discussed in Chapter 2, Halliday & Hasan's (1976) approach does not specify how to deal with cases where we have a lexical item in one sentence/T-unit links back to two lexical items in previous sentence/T-unit, or vice versa. It is not clear whether we count only one tie or two ties between such instances of repetition. Hence, studies that apply Halliday & Hasan's (1976) model (e.g., Khalil 1989; Conner 1984) may derive different counts of the same example. In Example (6), a tie could be counted between *reliance* (T-unit 3) and *rely* (T-unit 1), or two cohesive ties between (*reliance* – *rely*) and (*reliance* – *reliant*). There are other examples of

this type in the essay where counting can take different directions. Example (7) is another case:

(7) 7 Instead of trying to **solve** these problems using the power of my own brain, I instead relied on my computer and the access to the internet to identify projects that are relevant or closely related to the ones I was working on.//

8 The accessibility to such information made it easier for me to identify reliable **solutions** to the problems I was trying to **solve**.//

9 If it was not for the computer and for the internet, I would have had to use the power of my own brain to **solve** those problems.//

Halliday and Hasan (1976) would allow for the following options:

a. The cohesive tie between T-units (8) and (7) can be as follows:

solve – solve (1 tie), or:

solutions – solve; solve – solve (2 ties)

b. The cohesive tie between T-units (9) and (8) can be as follows:

solve – solve (1 tie), or:

solve – solutions; solve – solve (2 ties)

Therefore, the total cohesive ties in Example (7) in the three units of the essay can be either 2 or 4. On the other hand, if I apply LRNetM to Example (7), only one option for counting is possible as follows:

solve

solve

solve

solutions

According to the principles of analysis of LRNetM explained in Section 5.6, the frequency count of simple repetition in the lexical repetition network of *solve* is 2 whereas there are two instances of derived repetition. Halliday & Hasan (1976) applied their method of analysis inconsistently to similar examples of the ones provided above. Examples (8) and (9) show different ways of counting lexical cohesive ties when examples such as (6) and (7) are encountered in a text. The sequence of sentences is the same as in the text that Halliday & Hasan (1976) analyse.

- (8) My father's friend went out, he brought two **packs** of seeds back and he gave them to my father (9).

And we keep my sister's **pack** in one half of the box and my **pack** in the other half (10).

(Halliday & Hasan 1976: 354)

In Example (8), Halliday and Hasan (1976: 354) counted only one lexical cohesive tie of same item repetition between *pack* in Sentence (10) and *packs of seeds* in Sentence (9). However, the following example by Halliday and Hasan (1976) shows that they used a different method of counting from the one applied to Example (8) when they analysed a similar example.

- (9) I believed that Salamander of his but an image, and presently I found analogies between **smell** and image (7)

That **smell** must be thought-created, but what certainty had I, that what had taken me by surprise, could be from my own thought, and if a thought could affect the sense of **smell**, why not the sense of touch (8)?

(Halliday & Hasan 1976: 346)

According to Halliday & Hasan's (1976: 347) method of analysis in Example (9), both occurrences of *smell* in Sentence (8) establish a lexical cohesive tie with *smell* in Sentence (7). Therefore, in their coding table of this text, there were two cohesive ties between Sentences (7) and (8): *smell – smell*; *smell – smell*. There are other examples where Halliday and Hasan's (1976) method of counting can be conducted in different ways. As discussed in Chapter 2, Halliday & Hasan's (1976) principles of analysis did not, for example, identify what the lexical item (in a corpus linguistic sense) is. Halliday & Hasan (1976: 343 – 347) sometimes treated a two-word phrase like *carry knife* or *bad thing* as one item, or a three-word phrase such as *group of students*; *fields of rice* as one item. These are examples from the texts they analysed but Halliday & Hasan (1976) did not mention what criteria they used to identify the length of a lexical item – whether this item would be analysed as one unit or two units. This obscurity in determining explicit criteria of analysis can lead to different counts of lexical cohesion by different researchers. So, if I take an example from the essay sample in Example (3) which contains a repetition of phrases like *easy access* or *power of my brain*, there will be varied ways of counting these phrases when Halliday & Hasan's (1976) method of analysis is applied. Example (10) illustrates a repetition of the phrase *easy access* in the essay sample:

(10) **4** However, in this day and age, the computers provide us with **easy access** to vast amounts of data and information that can be used to easily solve some of our most complex problems.//

5 The **easy access** to the internet meant that people are becoming increasingly reliant on their computers to provide them with reliable and efficient answers rather than their own brains.//

In Example (10), there is a possibility of counting the lexical cohesive tie between *easy access* (T-unit 5) and *easy access* (T-unit 4) as one tie if we consider this phrase as one lexical item. Alternatively, there would be two cohesive ties between *easy access* and *easy access* if each item in this phrase is counted separately. This also applies to the phrase *power of my brain*, which is repeated three times in the essay sample in Example (3). If this phrase is treated as one item, there will be two lexical cohesive ties between this phrase in line with Halliday and Hasan's (1976) criteria of analysis. Nevertheless, if other researchers consider this phrase as consisting of two lexical items, there will be 4 cohesive ties between the repeated instances of this phrase, which I illustrated as follows:

power – power

power – power

brain – brain

brain – brain

Another area where inconsistency can arise when applying Halliday and Hasan's (1976) method of analysis is the possibility that some ties are either missed or not recorded during the process of tracking cohesive ties. Example (11) is extracted from Halliday and Hasan's (1976) sample texts and illustrates this case.

(11) [...] The last word ended in a long bleat, so like a **sheep** that Alice quite started (1).

She looked at the Queen, who seemed to have suddenly wrapped herself in **wool** (2). Alice rubbed her eyes, and looked again (3). She couldn't make out what had happened at all (4). Was she in a shop (5)? And was that really – was it really a **sheep** that was sitting on the other side of the counter (6)?

(Halliday & Hasan 1976: 340)

In Example (11), Halliday & Hasan (1976) recorded a lexical cohesive tie between *wool* in Sentence (2) and *sheep* in Sentence (1). However, the lexical cohesive tie between *sheep* in Sentence (6) and *wool* in Sentence (2) was either missed or not counted for a reason. Halliday & Hasan (1976) only identified the lexical cohesive tie between *sheep* in Sentence (6) and *sheep* in Sentence (1). This contradicts their principles of analysis that I explained in Chapter 2 when they suggested that the resolution of a cohesive tie accepts only the nearest cohesive element. Hoey's (1991b) model developed a clear set of analytical principles in identifying what needs to be counted or not as a cohesive link compared to Halliday and Hasan's (1976) model. However, the application of Hoey's (1991b) method to text analysis through the repetition matrix is complex, particularly when analysing a large corpus. His model therefore does not provide a practical improvement for the analysis of lexical cohesion.

Overall, the key issue with the method of analysis when these classic models of lexical cohesion are applied to text analysis is consistency. The examples I cited above showed that the method of analysis by Halliday and Hasan's (1976) model has not been consistently applied, and consistency among analysts cannot be maintained without explicit definitions of the criteria of analysis that eliminate subjective decisions as much as possible. The criteria of

analysis need to allow cohesive elements to be identified and counted systematically by different analysts or by the same analyst at different time. The main reason for such an inconsistency by the classic models is related to the complex notions of directionality or multi-directionality that these models applied in analysing lexical cohesion. As pointed out in Chapter 2, previous studies on lexical cohesion in NNS writing applied these classic models to learner writing but they did not discuss how far the principle of directionality is replicable to large-scale corpora, mainly when comparing two data sets of NS and NNS writing. These studies presented the frequency counts of lexical cohesion without providing working definitions of the categories they analysed or identifying any textual examples where the counting may vary. Actual instances of lexical cohesion in learner writing are complex because learner writing displays a density of lexical repetitions and other cohesive devices that sometimes occur within the same sentence. The tracking of whole chains of cohesive ties in learner writing will produce a variation in the analysis of lexical cohesion among analysts. However, such delicacy of the analysis of lexical cohesion in NNS writing has not been considered. I therefore suggested the LRNetM model as a replicable approach to comparing the frequency counts of lexical cohesion in native and non-native corpora.

As discussed in Section 5.2 and also in Chapter 2, LRNetM does not look at lexical cohesion as directional relationships between lexical items, but it is based on Sinclair's (2004) and Hoey's (2005) observations on how discourse or lexical cohesion is created in a text. Both researchers adopted an alternative approach to analysing discourse; Sinclair (2004), for example, suggests that what has already been mentioned in a text becomes a shared knowledge and as a result we do not need to rely on backward references to interpret the meaning of a text. Instead, the meaning of a lexical item is retrieved from the current unit of

the text which is under analysis. Therefore, the LRNetM model analyses lexical cohesion in line with this assumption and develops a systematic method of analysis which enables researchers to arrive at consistent counts of lexical cohesion. Although LRNetM cannot capture textual cohesive relationships that are crucial to complement the picture of lexical cohesion analysis, its design is well suited for the purpose of this study which is comparing NS and NNS writing. It also provides a starting point by identifying frequency counts for further analysis that will be complemented by text analysis and corpus analysis.

5.8 Counting signalling nouns in ALEC and LOCNESS

Signalling nouns, the third lexical cohesive category in this study, will not be quantified by the lexical repetition network model. They will be analysed using a manual text-analysis method. Benitez-Castro (2014) observes that most previous studies on SNs apply one or more of three types of analysis, i.e., fully automated, semi-automated and manual. Fully automated analyses (e.g., Francis 1993; Biber et al. 1999; Hunston and Francis 2000; Schmid 2000) are those conducted on large corpora (most commonly, the Bank of English (BOE)). This full automation method means the retrieval of examples from the corpus based on pre-defined queries. Such queries typically represent specific patterns such as, the NOUN + CLAUSE and NOUN + BE + CLAUSE patterns. Schmid (2000), for example, compiles a list of 670 shell nouns (i.e., signalling nouns) after searching the BOE for these two patterns. This method though, as Benitez-Castro (2014) points out, might lead to exclusions of potential signalling-noun instances that occur in patterns that are different from the ones that have been identified to make the query. Furthermore, this analysis which is based on the structural patterns cannot discriminate SNs from other nouns. Flowerdew and Forest (2015: 14) explain that these patterns that some researchers depend on in their analysis are not exclusive to the class of

signalling nouns, and so fail to serve as good discriminators of whether an item should or should not be counted as an SN. Therefore, Flowerdew and Forest (2015) suggest that while it is an important aspect of signalling nouns to have a specific structure, this characteristic may only be observed after an SN has already been identified via other means.

Another approach to analysing SNs is the semi-automation method, which represents a moderate degree of the previous method (full automation) in that researchers support predefined queries with a manual analysis. Yamasaki (2008) and Caldwell (2009) apply this approach and use the predefined queries of patterns as a starting point for some sort of experimental manual analysis of contextualised data. The third approach is the manual analysis, which does not depend on predefined patterns and is usually applied when the corpus is small (cf. Hoey 1993; Flowerdew 2003; Mahlberg 2005). The human component of the research process is thus of paramount importance in this type of analysis. I will follow a text-analysis approach that is based on a manual analysis since both ALEC and LOCNESS in the present study are of small size. With assistance of the working criteria suggested in Chapter 2, Section 2.7, I identify SNs in each of the two corpora and provide their frequency counts. The counts I make are based on the number of instances (tokens of SNs) in each text of the two corpora, not on the number of lexical realisations or on particular SN + lexical realisation complexes. Then, the frequency figures of SNs in each corpus are compared against each other to examine differences and similarities between the two groups of writers. Section 5.9.2 will discuss this comparison in more detail.

5.9 Results

This section addresses the first main research question with its three sub-questions.

RQ1: What are the relative frequencies of the tokens of each lexical cohesive category in each variety of writing (NNS vs. NS)?

- How many instances/tokens of simple repetition can be counted in each corpus (NNS vs. NS)?
- How many instances/tokens of derived repetition can be counted in each corpus (NNS vs. NS)?
- How many instances/tokens of signalling nouns can be counted in each corpus (NNS vs. NS)?

In order to answer this question with its three sub-questions, I counted the total number of tokens of simple repetition; derived repetition (see Section 5.9.1) and signalling nouns (see Section 5.9.2) in ALEC and LOCNESS. I then compared frequencies of each cohesive category between the two corpora. This comparison involved computing two types of frequencies: aggregated frequencies and individual text-based frequencies. That is, general comparisons of overall frequencies of simple repetition, derived repetition and signalling nouns in each individual corpus were conducted first. Then, frequencies of simple repetition, derived repetition and signalling nouns were measured for individual subjects.

Flowerdew (2010: 47-48) stresses the importance of considering text-based frequencies along with overall frequencies in corpus studies. He points out that the former provides a more

accurate measure of learners' ability than the latter, particularly when individual variations are expected, which is usually the common case with learner corpora. In contrast, he maintains that aggregated frequencies may disguise internal variations in a corpus. Brezina and Meyerhoff (2014) in their review of corpus-based sociolinguistic studies also criticise the use of aggregated frequencies. On the one hand, these researchers admit that aggregating data is a normal procedure in every corpus design particularly when dealing with large corpora where it is not easy to take into consideration individual variation. They attribute this difficulty to the complex process of information retrieval of the social (or linguistic) characteristics of individual speakers. In addition, they maintain that the samples from some speakers (or writers) can be relatively short and hence they cannot be meaningfully used as individual data points. For such reasons, Brezina and Meyerhoff (2014) observe that the majority of corpus (sociolinguistic) studies to date have relied on general comparisons of two or more sub-corpora.

On the other hand, Brezina and Meyerhoff (2014) claim that the central issue with this procedure of general frequencies is the fact that it considers general inter-group differences and neglects within group variation. They add that this methodology often yields falsely positive results. In this respect, Baker (2010: 56) argues that "when grouping together a large number of speakers we can overlook differences within groups, which may have a skewing effect on our results". Brezina and Meyerhoff (2014) further highlight that the aggregate data procedure creates stereotypes about language and society rather than contributes to our understanding of genuine sociolinguistic (or linguistic) variation. They, therefore, suggest that the procedure that takes into account the variation between individual speakers proved to be more reliable. Accordingly, the present study considered both types of frequencies (i.e.

aggregate and individual frequencies) to ensure that individual variations among learners in the use of lexical cohesion are also taken into account along with the general frequencies.

5.9.1 Frequency (tokens) of simple and derived repetition in ALEC and LOCNESS

As pointed out throughout this chapter, simple and derived repetitions are counted with the LRNetM model. Based on the discussion in Section 5.9, I firstly counted the general frequencies of simple and derived repetition with all of essays in each corpus aggregated. As the essays in the two corpora are of different sizes, the final frequencies of simple and derived repetition were normalised to conduct a reliable comparison. Biber et al. (1998: 265) emphasise that comparing raw frequencies in cases where the different sections of a corpus are not of equal size is misleading. Therefore, when comparing results, it is essential to take into account the different sizes of specific parts of the corpus (Hoffmann et al. 2008: 70-71). ‘Normalisation’, as defined by Biber et al. (1998: 263), is “a way to adjust raw frequency counts from texts of different lengths so that they can be compared accurately”. The basis chosen for normalisation in the present study is per 1,000. This rate of normalisation is also the closest approximation to the average length of essays in ALEC and LOCNESS (i.e. 627 vs. 542 words in length). To count the aggregated frequencies, I adopted the following formula:

Raw (absolute) frequency of simple or derived repetition in all essays together per one corpus/ number of words in the corpus*1,000

Secondly, I calculated the individual text-based frequencies of simple and derived repetition in all the individual essays in ALEC and LOCNESS. I applied this formula to calculate the normalised (relative) individual text-based frequencies:

Raw (absolute) frequency of simple/derived repetition in an individual essay/ number of words in this essay* 1,000

Thus, if we have, for example, an essay from ALEC that contains 101 tokens of simple repetition, and the essay length is 521 words, the relative individual text-based frequency is:

$$\frac{101}{521} \times 1,000 = 193.8$$

This process was applied to each essay in both corpora. From these relative individual text-based frequencies, I calculated the mean by adding up all figures of the relative individual text-based frequencies of each essay in the corpus and dividing them by the total number of essays in each corpus. Table 16 shows both types of frequencies in ALEC and LOCNESS.

Category of lexical cohesion	Aggregated frequencies		Individual text-based mean frequencies	
	ALEC	LOCNESS	ALEC	LOCNESS
Simple repetition	165.16	132.6	162	127.5
Derived repetition	37.8	35.3	36.4	34

Table 16 Frequency of simple and derived repetition per thousand words in ALEC and LOCNESS

In Table 16, the aggregated frequency of simple repetition indicates that Arab NNSs used simple repetition approximately 24% more often than did NSs (165 vs. 133 per thousand). Likewise, the table reveals that the mean frequency of simple repetition in the individual essays by Arab non-native speakers of English ($M = 162$, $SD = 36.5$) is approximately 27% more often than do native speakers of English ($M = 127.5$, $SD = 32$). But with what degree of

confidence can we understand that this is a valid finding about the two groups rather than a result of chance?

In order to be certain that these observed differences in the frequency of simple repetition between the two groups have not arisen by chance, I conducted a significance statistical test. There are a number of different statistical techniques, depending on the type of variables that are being compared (Biber et al. 1998: 275). The chi-square test is the most frequently used significance test in corpus linguistics and also has the advantages that it is more sensitive than the t-test, for instance. It also does not assume that the data are normally distributed. The main disadvantage of chi-square, however, is that it is unreliable with very small frequencies (McEnery and Wilson 2001). Despite the advantages of the chi-square test, I found that the t-test is more appropriate to be performed in the present study. Oakes (1998) points out that the t-test is a reliable technique when we have small samples, which is the case in the present study. Dornyei (2007: 215) suggests that we need t-test statistics if we want to compare two groups but we cannot certainly presume that the observed difference indicates any 'real' difference. He maintains that using a t-test will help check whether we have had a generalisable result or whether the score is likely to be merely an artefact of random variation.

The type of the t-test that was used for the present study was the independent-samples t-test because it is concerned with comparing the results of groups that are independent of each other (e.g., NNSs and NSs), which is the case in this study. Providing that the t-test entails a normally distributed data, I conducted a normality test using SPSS and the Shapiro-Wilk's test ($p > 0.05$) proved that all frequency figures were approximately normally distributed for each repetition category. Back now to the results presented in Table 16, an independent t-test suggested a significant difference between the two groups in terms of their use of simple

repetition ($t = 3.8, p = 0.00017$). To sum up, the results on simple repetition presented above answered the first sub-research question, which compares the relative frequencies of simple repetition in both corpora. The relative overall frequencies of the tokens of simple repetition in ALEC are 165, whereas they are 133 in LOCNESS. Hence it becomes evident that the common view held by many researchers that Arab learners of English tend to employ simple repetition abundantly is supported by the quantitative data of the present study.

For the use of derived repetition in both corpora, Table 16 demonstrates the results of the relative aggregated frequency of derived repetition, which indicate that Arab speakers of English unexpectedly used derived repetition more often than native writers did (38 vs. 35 per thousand). Table 16 also showed that Arab speakers of English scored higher individual frequency of derived repetition 6% more often than did native writers (36 vs. 34 per thousand). However, just by visually inspecting these figures of frequency in Table 16, we can see that the difference is not significant. A t-test confirmed this observation that the difference between the two groups was not significant, $t = 1.05, p = 0.25$. These results answered the second sub-research question that deals with the frequency of derived repetition; the relative overall frequencies of derived repetition in ALEC are 38, while they are 35 in LOCNESS. It can be primarily concluded that both groups used a very close amount of derived repetition in their writing. Nevertheless, more investigation is still needed to understand why Arab speakers of English used more derived repetition compared to native speakers. This investigation will be conducted in the next chapter, where qualitative analysis is used to interpret the quantitative data provided in this chapter.

5.9.2 Frequency (tokens) of signalling nouns in ALEC and LOCNESS

The frequency of signalling nouns as stated in this chapter is analysed using a text analysis method. In order to answer the third sub-research question presented in Section 5.9, I counted relative aggregated frequencies and individual mean frequencies of SNs in ALEC and LOCNESS. I then compared these frequencies between the two groups. I firstly applied the following formula to calculate the relative aggregated frequencies of SNs in each corpus:

$$\frac{\text{Tokens of SNs in all essays together}}{\text{Number of words in the corpus}} \times 1,000$$

Then, I computed the relative text-based individual frequencies of SNs in each essay as follows:

$$\text{SNs per thousand} = \frac{\text{Tokens of SNs in an individual essay}}{\text{Number of words in this essay}} \times 1,000$$

After that, I accumulated all the normalised figures that represent the SNs' relative individual frequencies, and divided them by the total number of essays in each corpus in order to compute the mean. Table 17 shows the relative aggregated frequencies and the individual text-based mean frequencies of SNs in both corpora.

Category of lexical cohesion	Aggregated frequencies		Individual text-based mean frequencies	
	ALEC	LOCNESS	ALEC	LOCNESS
Signalling nouns	8.9	6.8	9	6

Table 17 Frequency of SNs per thousand words in ALEC & LOCNESS

As shown in Table 17, the relative aggregated frequencies of SNs per thousand words revealed that Arab speakers of English used SNs approximately 29% more often than did the native speakers of English (8.9 vs. 6.8 per thousand). Similarly, the relative text-based frequencies of SNs per thousand words approved these results and showed that the mean frequency in individual essays by NNSs ($M = 9$, $SD = 4.5$) is approximately 1.5 times that of NSs ($M = 6$, $SD = 4$). A t-test suggested a significant difference between the two groups, $t = 2.306$, $p = 0.024$. These results answered the last sub-question of the first main research question. The last sub-research question compares the overall relative frequencies of SNs in both corpora, and the results on SNs showed that the overall relative frequencies of SNs in ALEC are 9, whereas they are 7 in LOCNESS. Figure 11 below, which is based on Tables 16 and 17, demonstrates the difference between mean frequencies for each category of lexical cohesion between the Arab NNSs (ALEC) and the NSs (LOCNESS).

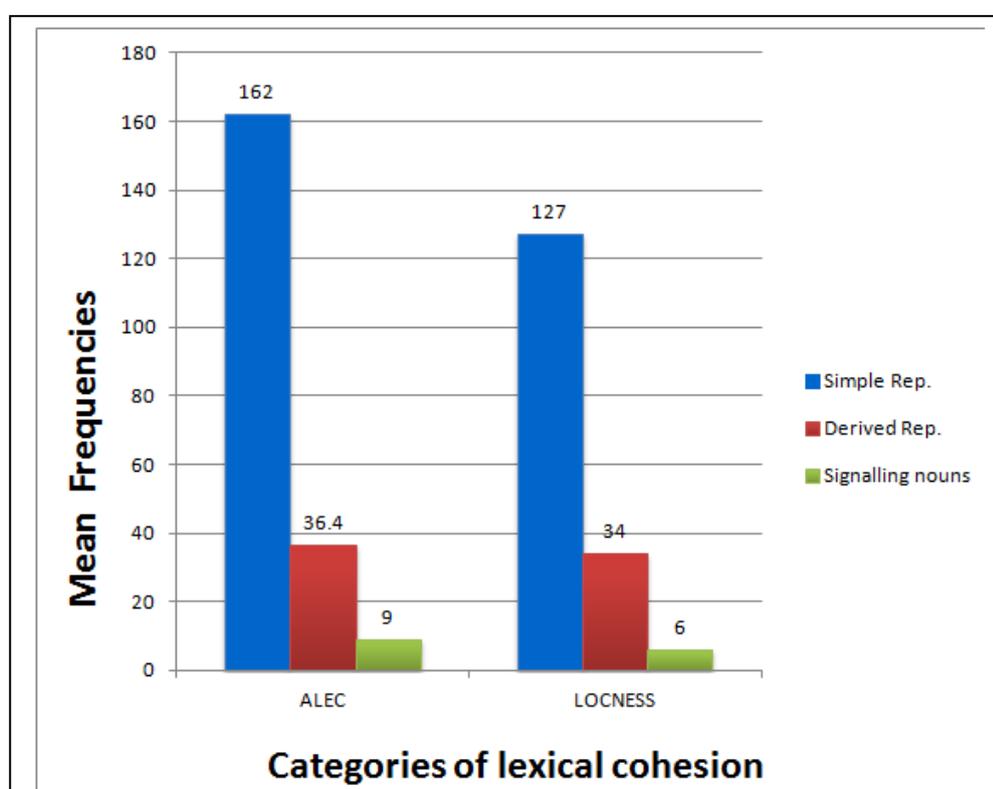


Figure 11 Individual mean frequencies of simple repetition, derived repetition and SNs in ALEC & LOCNESS

Figure 11 illustrates that Arab NNSs achieved higher frequency in all lexical cohesive types: simple repetition, derived repetition and signalling nouns compared to NSs. The figure further reveals that the most used cohesive category by the two groups is simple repetition, while signalling nouns form the least frequently used category among other lexical cohesive categories in both groups.

5.10 Conclusion

This chapter highlights that simple and derived repetitions are better captured with corpus tools than signalling nouns which have textual features. Therefore, the chapter has introduced the lexical repetition network model that I suggested to quantify simple and derived repetitions in native and non-native writing. The LRNetM model is based on grouping wordlists into lexical repetition networks. Within these networks, direction of the textual reference is not a key factor in determining cohesive relationships between lexical items. This model supports the view that lexical cohesion needs to be analysed without considering the traditional concept of directionality, which creates complex patterns of cohesion in which texts have many lines drawn from one part of the text to another to indicate ties, links or chains, etc. The LRNetM model follows Hoey's (2005) suggestion that cohesive properties are inbuilt in the lexical item itself. Therefore, the analysis of lexical cohesion according to LRNetM focuses on the lexical item under analysis and the lexical repetition networks that this item produces. This model also provides a number of quantitative measures that can be used to count tokens of simple and derived repetitions in the lexical repetition networks that make up the text under analysis.

The chapter has stressed that the inconsistency in the application of models of lexical cohesion to text analysis indicates how fuzzy and complex the concept of lexical cohesion is. As evidence of this inconsistency, the chapter showed that frequency of lexical cohesion differs depending on which model of lexical cohesion is applied to analyse the text. This diversity in quantitative methods among models is related to whether cohesion is looked at as a tie, link, or a network, etc. The chapter then highlights that automatic tools fail sometimes to discriminate signalling nouns from other nouns that have similar features to them. As a result, a manual text analysis method is recommended to analyse SNs in their textual contexts particularly when the corpus is small.

Then, the chapter presented the quantitative results that identify frequency counts for simple repetition, derived repetition and signalling nouns. The Arab NNSs have a favourable tendency to use lexical cohesion with its three categories as opposed to the NSs. Many questions arise at this point. For example, does the higher frequency of SNs indicate that the writing of Arab speakers of English is more cohesive than that of NSs? Can an interpretation lie elsewhere other than in the frequency counts? Bearing this in mind, linguistically objective explanations have to be sought for these questions. Therefore, in the two next chapters I will conduct a qualitative analysis of the three cohesive categories in the two corpora. This analysis can help find out what could have possibly given rise to the frequency of cohesive devices in the Arab non-native writing compared to the native speaker writing.

Chapter 6

A qualitative look at the results: A text-linguistic analysis and discussion

6.1 Introduction

The quantitative results in the previous chapter showed that a corpus of argumentative essays by Arab speakers of English are characterised by a higher frequency of simple repetition, derived repetition and signalling nouns, as compared to a corpus of argumentative essays written by native speakers of English. This chapter will, therefore, present a qualitative analysis to investigate what could have caused this preponderance of lexical cohesive devices in the English writing of Arab L2 learners. This chapter principally answers the second research question:

RQ2: What are the advantages and disadvantages of a text linguistic approach in describing the function of the lexical cohesive forms in each corpus (NNS vs. NS)?

Based on this research question, the present chapter will show how a text analysis was conducted to describe the functions of simple repetition, derived repetition and signalling nouns in ALEC and LOCNESS. This analysis examined how these forms of lexical cohesion function in their context at the paradigmatic level (see Chapter 1, Section 1.5 for an explanation of the paradigmatic and syntagmatic levels of language). The analysis of lexical cohesion at the paradigmatic dimension implies how cohesion functions at the formal and semantic level. At the structural level, El-Gazzar (2006) points out that forms of lexical cohesion create the sense that the ideas presented in different sentences are connected, which results in building levels of cohesion within the text. At the semantic level, Reynolds (2001)

emphasises that an appropriate use of lexical cohesion indicates the writer's ability to expand upon his/her ideas and relate new with old information. This function is also underlined by Winter (1977, 1979) who stresses that lexical cohesion (particularly, repetition) has to have an informational value in that it provides a framework for interpreting what is changed. Analysing this function of lexical cohesion using text analysis can further help examine whether there are specific patterns of lexical cohesion that exhibit a distinct style to the writing of Arab L2 learners, thus revealing uses of interlanguage cohesion in the writing of this group.

The present chapter starts with a text analysis of simple and derived repetition. This will be accompanied with a discussion of the potential reasons for the prevalence of these two types of repetition in the English writing of Arab L2 learners, compared to that of native writers (see Sections 6.2 and 6.3). In these two sections, the findings of the present study on simple and derived repetition will be related to the findings of previous studies on these two categories. The discussion of simple and derived repetition will be shorter in contrast to that of signalling nouns. In Section 6.4, I will give an account of the high frequency of SNs in the Arab L2 writing. This section will firstly compare the findings of the present study on SNs with the findings of other studies on SNs in learner writing. In Section 6.5, the procedure that is used to count specific patterns of SNs in ALEC and LOCNESS will be evaluated. Section 6.6 will then highlight other tendencies of using SNs by Arab L2 learners by presenting a number of textual examples from both corpora. Section 6.7 identifies the most commonly used types of SNs in both corpora. This analysis of the types of SNs will reveal how the Problem-Solution pattern (cf. Hoey 2001) is used in each corpus (see Section 6.8). Section 6.9

gives a general discussion on the use of SNs in the Arab L2 writing and NS writing. Finally, Section 6.10 presents a conclusion.

6.2 Simple repetition

In this study, the predominant preference for simple repetition among Arab L2 learners in argumentative essays, compared to English native speakers, might indicate that they favoured this category of repetition because it is a simpler means of maintaining lexical cohesion than other cohesive categories, which need more cognitive effort to be produced such as signalling nouns. This finding is compatible with that of other researchers (e.g., Khalil 1989; El-Gazzar 2006) who found that Arab non-native speakers of English tend to rely mainly on simple repetition for textual cohesion. Khalil's (1989) findings, for example, showed that Arab L2 students overuse reiteration of the same lexical item as a cohesive device compared to other devices. This overwhelming use of simple repetition, as Khalil (1989) notices, prevents students from delivering sufficient information on the assigned topic. Likewise, El-Gazzar (2006) also found that the English writing of Arab students shows poor lexical cohesion because it contains an excessive amount of simple repetition (about 45% of the total lexical types occurring in her study).

In order to illustrate the difference between Arab speakers of English and English native writers in terms of the use of simple repetition, it is useful to provide textual examples from ALEC and LOCNESS. The forthcoming Examples (1-12) are taken from both corpora. In these examples, T-units are numbered according to their sequence in the essay, and square brackets indicate intervening T-units that have not been presented. The T-units are separated by two slashes to indicate the end of each T-unit. Also, instances of simple repetition are

underlined. In a few cases, derived repetition will be marked in bold whenever relevant in order to show the intensive use of lexical repetition networks.

6.2.1 Text analysis of simple repetition in ALEC

This section presents examples that illustrate how simple repetition is used by Arab speakers of English in the ALEC corpus. Example (1) is an extract from an essay about ‘computer and human brain’:

- (1) **1** It is true that computer can run faster, operate pre-defined functions faster and then execute these functions upon the defined parameters. // **2** The computer can not act on its own, // **3** the human brain have to define and determine what function to be executed. // **4** Next to any computer or machine there must be a human to check the operation of that computer. // **5** Computer can not work on their on // **6** they need some kind of lines to be defined, some tasks which is programmed by human then embedded into the machine to do the operation. // **7** computers needs human brain to develop and integrate new functions based on the operation raised. // **8** It is us who think for computers and it is us who embed and program the computer to run some certain operations.

Even though the main idea expressed in Example (1) is both logical and relevant to the subject matter, the paragraph shows that simple repetition is employed redundantly to express a single idea (concept) across the text. The main proposition in the extract above is that computers cannot operate without human supervision. The Arab L2 learner has achieved this through the repetition of lexical items: *operation*, *functions*, *execute*, *define*, *embed* and *programme*. The repetition of the idea starts from T-unit (5), in which the L2 learner begins to rephrase what he has written in the first T-units (1-4).

If we focus on the lexical item *function(s)* (n), for example, it appears 8 times in the whole essay. However, as it is clear from the extract, the Arab L2 writer uses this item to reiterate the same idea over the text without developing the argument adequately. As a result, T-units (5-8) are unnecessary repetitions and can be removed without affecting the message that is being delivered. Let's examine another example to see how simple repetition is utilised by Arab L2 learners. Example (2) is taken from an essay about 'traffic'.

- (2) **1** The rising up will limit the pollution problem and will keep a friendly environment // **2** as people will lead to use less fuel, park their cars and start walking instead which then will help to keep them active and get healthier and be fit. // [...] **6** The use of alternatives like bicycles and walking will affect their health on good way and keep them as active and focused.

As Example (2) demonstrates, T-unit (6) is a paraphrase of T-unit (2). That is, the Arab L2 learner repeats the same proposition that he introduced in T-unit (2), which is the use of alternative ways to reduce the number of cars. Rather than developing this idea by providing a series of specific examples, the learner develops the idea through parallel constructions. This means that, in T-unit (6), the learner repeats the same lexical items that are used in T-unit (2) which are: *walking*, *active* and *health* without adding any new information that can expand the idea further. This use of parallel constructions leads to a repetition of a pattern that renders the text ineffectual. This kind of repetition is prevalent in the English writing of Arab learners in the ALEC corpus. Example (3), which is extracted from an essay about 'dangerous sports', shows another case of repeating entire phrases and T-units through the use of a density of simple repetition networks. Let's examine the example below.

- (3) **1** [...] it is the duty of central governments to ensure the safety of their citizens. // **2** Therefore, it is fair to say that governments should introduce tight regulations in relation to the health and safety of individuals who are involved in this type of sports. // **3** This regulations should focus on the roles of individual clubs in relation to the protection of their players and ensuring the safety of the players and the spectators as well. [...] **5** As I mentioned already, it is the duty of central governments to ensure the health and safety of their citizens // **6** and extremely dangerous sports can ensure neither the safety nor the lives of players and perhaps even spectators.

In Example (3), the Arab L2 writer communicates the main message to the reader in the first T-units (1, 2 and 3). In these T-units, the writer argues that it is the responsibility of the government to ensure health and safety of players and spectators, who are involved in dangerous sports. That is by introducing tight regulations that can protect these individuals. Then, the writer plainly repeats this argument in the subsequent T-units. As Example (3) reveals, there is an abundance of simple repetitions that are realised through these lexical networks: *duty – duty; central – central; governments – governments –governments; ensure – ensuring – ensure; safety – safety – safety; citizens – citizens; regulations – regulations; health – health; players – players; spectators – spectators*. There is a clear over-explication in expressing the main idea whilst the six T-units that compose the text above could be condensed and economically re-written with the minimum of surface constituents. This is a revision of Example (3):

- (3-a) **1** It is the duty of central governments to ensure the [health] and safety of their citizens. // **2** Therefore, it is fair to say that governments should introduce tight regulations for individuals who are involved in this type of sports.// **3** These regulations should focus on the roles of individual clubs in relation to the protection of their players and the spectators as well.

After this revision of the paragraph, Example (3-a) now simply contains two lexical networks of simple repetition which are: *governments – governments*; *regulations – regulations*. Such a tendency of repetition as a strategy of over-assertion is detected in different topics in the ALEC corpus. In Example (4) below, this time about ‘immigration’, the Arab L2 writer repeats the same idea through using simple repetition as a connector device.

- (4) **1** In addition, developed countries should not ignore their responsible (*sic*) towards under developed countries where immigrants come from, // **2** if we returned back for century or nearly it would be **clear** that most of what are happening recently have resulted of **colonialism** period when centuries of injustice had been the situation and when these developed countries were strongly controlling under developing countries. // **3** Regarding to both views, it is **clearly** that countries such as the UK must not come over its responsibilities toward immigrants especially those who come from **ex-colonialist** regions and also for internal benefits for the economy.

Example (4) indicates that T-unit (3) is a reproduction of information that has been introduced in T-units (1) and (2), with the exception of introducing a new aspect of discussion that relates to the economic benefits of immigrants (last line in T-unit 3). The Arab L2 learner introduces this new material abruptly, while he/she has to prepare the ground for it in the earlier text to make the T-units more connected. With the focus on simple repetition, the writer makes little effort to change the wording at all in T-unit (3). The writer uses simple repetition to connect T-unit (3) with T-units (1) and (2). He/she uses these networks of simple repetition: *responsible (sic) – responsibilities*; *immigrants – immigrants*. The writer also employs derived repetition (highlighted in bold) to help him reiterate their argument: (*clear – clearly*; *colonialism – ex-colonialist*). Despite this intensive use of lexical repetitions, T-unit (3) creates, with the previously mentioned T-units, a tedious link that is insufficient for the

progression to be clear to the reader. In addition, in Example (4), the writer should have used a noun *responsibility* instead of the adjective *responsible*. These errors seem to be due a problem of general lack of knowledge of word formation.

All examples of simple repetition discussed above are clear cases of repetition that facilitates production, as Tannen (1989) describes. That is, repetition relieves the writer from the need to think up creative and new ideas to write. Another oddity in the English essays produced by Arab L2 learners is their use of simple repetition in the same T-unit. Although these instances of within-T-unit repetition are not included in the present study, they are still indicative of the Arab L2 learners' tendency to repeat. Examples (5) and (6) indicate this case.

(5) The number of cars on the road is less likely to be reduced in significant number.

In Example (5), the Arab L2 writer does not need to repeat the word *number*. He could simply write:

(5-a) The number of cars on the road is less likely to be reduced significantly.

(6) Although such improvements must be applauded, the policy of higher gas prices does really seem to achieve the aim it is designed to achieve which is to reduce traffic problems.

The Arab L2 writer in Example (6) uses unnecessary post-modifying clause: *it is designed to achieve*, which contains the verb *achieve* that creates simple repetition link with the first

mention of *achieve*. Therefore, the complete post-modifying clause can be deleted and the example could be re-written as follows:

- (6-a) Although such improvements must be applauded, the policy of higher gas prices does really seem to achieve its aim which is reducing traffic problems.

Another interesting finding in the argumentative essays written by Arab speakers of English was that these learners sometimes repeat the same ‘theme’ and ‘rheme’ (Halliday 1985, 1994). This observation will be briefly considered because the topic of thematic progression and its relation to cohesion is not within the scope of this study. However, this allows us to look at the different strategies that Arab speakers of English employ in their use of simple repetition as a cohesive device. Hawes (2015) points out that theme and rheme act as the building bricks of cohesion. The ‘theme’, as defined by Halliday and Matthiensen (2004: 64), is “the point of departure of a message”. This means that theme is typically the start of the clause that provides what the message will be about. The theme is also related to ‘Given Information’ which refers to information that is recoverable from the text (Halliday 1994: 298); that is, information which has been mentioned before. The ‘rheme’, on the other hand, as Halliday (1994: 37) describes, is everything that is not ‘theme’ and it is the part of the clause where the theme is developed. The development of the theme is attained by introducing ‘New Information’ that has not been mentioned before.

Sa’adaddin (1989) observes that English texts written by Arab L2 learners rely mainly on theme and rheme repetition patterns for the purpose of rhetorical persuasion. Sa’adaddin (1989) maintains that this type of repetition is frequently employed by Arab speakers of English in the form of parallel constructions which might create an impression that their

English essays lack progression. With this in mind, Example (7) from a topic on ‘computer and human brain’, illustrates how an Arab L2 learner repeats the theme and rheme in a short paragraph in order to emphasise a single idea. The underlined items show instances of simple repetition, while the theme and rheme are identified in the discussion that follows the example.

(7) **1** This leads to lack on our expression when we intend to express our emotions and daily news/routines. // [...] **5** [...] this have a great impact on maintaining our language referencing // **6** and it will lead to loss of significant treasury meaningful expressions.

Example (7) shows that the Arab L2 writer repeats the theme and rheme in this short stretch of the essay through using parallel T-units. The theme in T-unit (1) is (*this*), which is a non-participant theme that refers back to the excessive use of social media. The complement of the T-unit (1), which is (*leads to lack on our expression when we intend to express our emotions and daily news/routines*), represents the rheme of the first T-unit. Then, in T-unit (5), the Arab L2 learner returns to the same idea and produces a parallel structure that again contains the same theme (*this*) that was used in T-unit (1), with a little paraphrase of the rheme through using *language referencing* instead of *expression*. In T-unit (6), the Arab L2 writer uses (*it*) as a theme instead of (*this*), but both themes signal the same referent which is the use of social media. For the rheme in T-unit (6), the writer reiterates the same elements that were used in T-unit (1). These elements include: *leads to*; *expression*. The writer also uses *loss* instead of *lack* as a synonym.

Although simple repetition can lend an aura of logic to a sequence, it can also degenerate into a plodding succession of links if it is overused. Example (7) above indicates such a case where the Arab L2 writer repeats the theme and rheme constantly without introducing any new information that can develop the argument across the essay. In this respect, Hawes (2015), in his analysis of thematic progression in the writing of international students and professionals, observes that international students concentrate somewhat too much on local cohesion at the expense of whole-text coherence. This means that the use of repetition by overseas students is limited to connecting adjacent clauses linearly and not to linking clauses into extended text. In Chapter 1, I highlighted that although lexical cohesion is manifested through local and surface features, it has to serve a function which is to develop the argument of the text. Hawes (2015), therefore, suggests that students perhaps should be encouraged to systematically remind their readers of the overriding topic to increase cohesion and ensure a sense of logical construction. He adds that students need to be provided with knowledge and skills in the area of thematic progression, i.e. the linking of clauses into extended text by the repetition and transformation of elements in their themes and rhemes.

6.2.2 Text analysis of simple repetition in LOCNESS

It is imperative at this time to look at examples from the English native writing in order to see how native speakers of English use simple repetition compared to Arab speakers of English. The examples below are taken from LOCNESS. The following paragraph is an example taken from a topic about ‘computer and human brain’.

- (8) 1 One area which has drawn particular criticism is the computer games industry. // 2
Since the early 1980’s, when computers such as the Spectrum 48K and the Commodore 64 were introduced into the homes of millions of people, controversy

has followed about their effect on children. // 3 There is a widespread belief among parents that computer games hinder a child's ability to learn. // 4 The development of more sophisticated and technologically advanced computers, such as the Amiga and PC, has served to deepen the problem. // 5 The growing realism of the games appeal strongly to children. // 6 However, parents feel that children should be broadening their minds by reading books and that the computer games industry is encouraging children not to learn.

In Example (8), the main idea that the native writer emphasises is that computer games can have a negative effect on children. In order to deliver this idea, the native writer creates lexical networks of simple repetition by repeating these items: *computer* (adj), *games*, *industry*, *computers*, *children/child's*, *parents* and *learn*. Simple repetition in this passage helped create a clear development pattern for the reader to hold on as s/he navigates the essay. The writer starts his argument by mentioning *computer games* in T-unit (1). Then, in the second T-unit, he gives examples of the type of computers, which contain games that could affect children. In order to define what this impact is, the writer firstly uses simple repetition of *computer*, *games* and *child's* that link back to *computer*, *games* and *children* in T-units (1) and (2).

The native writer employs this simple repetition as a platform for introducing new information that computer games hinder children from learning. In T-units (4) and (5), the writer introduces other types of more advanced computers that might give children access to gaming, and then the writer re-asserts that computer games becomes a fact of existence because of its tempting effect in attracting children's interest. In the final T-unit, the native writer uses simple repetition to supply new information that highlights the importance of reading as a source of learning instead of gaming. The following is another example from the

same topic of ‘computer and human brain’ but by a different writer. The native writer in Example (9) connects the four T-units by means of simple repetition: *computers*, *workplace* and *replace/d*. The example is shown as follows:

- (9) **1** Another moral dilemma that computers have created is their role in the workplace.
// **2** There are many people who fear that computers will eventually replace man in the workplace. // **3** The role of computers is already significant in employment areas, such as accountancy, which require a specialist ability to be able to calculate and manipulate numbers. // **4** These computers have replaced humans.

In Example (9), even though the second mention of *workplace* in T-unit (2) could be discarded, its repetition still has a function in that it links back to the initial occurrence of *workplace* to emphasise that computers will supersede humans in the area of employment. T-unit (3) then expands this idea further by providing an example of which areas of profession that computers might replace humans. We notice then that the last T-unit (T-unit 4) again could be deleted because it echoes T-units (1) and (2) by repeating *computers* and *replace*. However, the repetition of these items in T-unit (4) refers to computers that are used in accountancy sectors, whereas *computers* and *replace* in T-units (1) and (2) are used in a more general argument. Therefore, the repetition in T-unit (4) in this sense is still acceptable because it carries a function. Example (10) is a further illustration of how native writers use simple repetition as a means of continuing the discussion about a given topic. Consider this extract from an essay about ‘traffic’, where the native writer suggests solutions to reduce the number of cars.

(10) **1** So, the only way around the problem is to have less cars. // **2** There are three main ways of doing this: sharing cars, using public transport and walking or cycling. // **3** Sharing cars seems feasible, but is unpopular for various reasons. // **4** It removes the element of independence afforded by a car - // **5** you have to decide exactly when and where you want to travel in advance. // **6** Also people who are paying for a car's road tax, insurance and depreciation prefer to use the car. [...] **12** People walking or using a bike are not protected from the elements // **13** and, until cars and buses are banned, have to breathe everybody else's exhaust fumes. // **14** These methods of transport are slow, and allow you to carry much less.

In Example (10), the native writer uses simple repetition effectively to connect the ideas together smoothly and clearly. The writer, for example, introduces the idea of 'sharing cars' in T-unit (2). Then, in the subsequent T-unit, he/she re-enters the same lexical items *sharing* and *cars* that firstly draw the reader's attention and keeps him/her on track, and also supply new information that helps to develop the argument. The lexical item *cars* in T-unit (2) also links back to *cars* in T-unit (1) and it is further repeated in the succeeding T-units forming a long cohesive network of simple repetition: *cars* – *cars* – *cars* – *car* – *car's* – *car* – *cars*. The missing T-units (7-11) that are indicated by the square brackets explain the second suggestion of using 'public transport' but they are not presented in the passage above in order to capture T-units that are more distant in the essay. In T-unit (12), the writer repeats the lexical item *walking* that he/she firstly introduces in T-unit (2) as a third suggestion to reduce cars. The repetition of *walking* in T-unit (12) serves as a trigger for new material to be added. This new material is provided in T-units (12) and (13). Finally, T-unit (14) encapsulates the three previously mentioned suggestions (*sharing cars*, *public transport* and *walking*) by using the word *transport* which acts as a super-ordinate term. Besides, *transport* forms a simple repetition link with *transport* in T-unit (2).

The following two Examples (11) and (12) are about ‘boxing’. Both examples demonstrate how native speakers of English use simple repetition to develop the main argument of the essay. These examples are outlined as follows:

- (11) **1** So what is the future of boxing? // **2** There must as in most things, be some room for compromise (*sic*). // **3** Suggestions as for the improvement of the sport range from the sensible: reduce the number of rounds, increasing time between rounds, changing the type gloves used and regulating the time span between each fight, to the ludicrous such as only allowing body punches, a measure that would also send the original sport under ground. // **4** Compimise (*sic*) can and must be made if boxings future is to be clear but at the moment their is still sufficient argument for continuing the noble art.

In Example (11), the native writer uses three lexical networks of simple repetition: *future – future*; *boxing – boxing’s*; *compromise – compromise*. As the example shows, the writer commits a spelling mistake (i.e. *compimise*), but this does not affect the cohesive function of the word. The three networks interact together to join the whole 4 units in the passage above. To begin with, the native writer in T-unit (1) questions the future of boxing by entering the two lexical items *future* and *boxing*. In T-unit (2), he/she proposes that boxing should be tackled with *compromise*. *Compromise* in this example could also function as a signalling noun that refers forward to a series of information. The writer introduces this information about compromise in T-unit (3) by negotiating possible suggestions. Then, he/she repeats *compromise*, *future* and *boxing* in T-unit (4) to affirm that compromise is a necessary solution to secure the future of boxing.

In Example (12) below, the main lexical network that spans over the text is the repetition of the noun *behaviour* (*behaviour – behaviour – behaviour*). The noun *behaviour* could also act as a signalling noun as we will see in the passage below. There are also two shorter networks, which consist of the following items: (*boxers – boxers*; *clash* (n) – *clashes* (n)). Let's consider this example.

- (12) **1** This apparent barbarianism is not helped by the behaviour of certain boxers. // **2** The claims of Oliver McCall, prior to his World Championship clash with Frank Bruno, that he would extract revenge for Gerald McClellan upon Bruno, have given the anti-boxing league fresh impetus in their drive for the banning of the sport. // **3** When this behaviour is added to that of Chris Eubank and Nigel Benn before their clashes, and one begins to agree [to ban the sport]. // **4** Whilst such behaviour cannot be condoned, it must be remembered that boxers do realise the risk of their chosen profession, just as other sportsmen do.

In Example (12), there is a constant repetition of the noun *behaviour*. The native writer uses *behaviour* in the first T-unit. Then in T-unit (2), the writer defines the meaning of this *behaviour* that indicates the intentions of boxers to take revenge on their opponents. To establish textual links that readers need, the writer repeats *behaviour* in T-unit (3) that refers back to the *behaviour* of revenge in T-unit (2). Simultaneously, *behaviour* further refers forward to the *behaviour* of famous names of boxers who fight merely to exact revenge. In T-unit (4), the writer repeats *behaviour* to reinforce the idea that this *behaviour* should not be excused. The writer at the same time highlights that boxers are aware and responsible for the risk of this *behaviour*.

Having looked at the examples from both corpora, we notice that both groups of Arab speakers of English and native speakers of English used simple repetition in their argumentative essays. However, the textual examples revealed that this difference is not simply a matter of frequency but it is more to do with how simple repetition functions in each variety of writing. Sections 6.2.3 and 6.3.3 provide a discussion on possible reasons that led Arab speakers of English to use simple repetition more frequently than English native speakers did.

6.2.3 Simple repetition and lexical redundancy

The textual examples in Sections 6.2.1 and 6.2.2 showed that Arab speakers of English used simple repetition redundantly, whereas native speakers of English used simple repetition to develop their argument and connect their ideas cohesively. El-Gazzar (2006) points out that a good essay does not contain this kind of redundant repetition. She claims that good writers not only have a wide range of vocabulary in comparison with poor writers, but also are familiar with the convention of language use for academic writing. Therefore, this linguistic feature of redundancy leads Arab L2 learners to produce poor essays. In contrast, Kai's (2008) study showed that NSs, as compared to NNSs, use more sophisticated forms of lexical cohesion and avoid the clumsy juxtaposition of the same lexical item in their writing. In this sense, Witte and Faigley (1981: 197) compared good and poor writers and pointed out that good writers have a good control of creative writing skills through the use of different lexical cohesive markers and collocation. These skills allow good writers to elaborate and extend the concepts they introduce. In contrast, the poorer writers lack these skills and their essays demonstrate a much higher degree of lexical and conceptual redundancy. Several studies (e.g., Aziz 1988; Ostler 1987) that have been carried out on English-writing problems of L1 Arabic speakers

showed that Arab learners of English tend to have a heavy reliance on redundancy that does not add any new information to the text, and their texts as such lack lexical variety. In this regard, McGee (2008) acknowledges that even though repetition is a prototype form of lexical cohesion, there can be a considerable number of redundant repetitions in L2 students' writing. Ting (2003: 6) comments that although redundant repetition is not a serious hurdle to the meaning of the delivered message, it inhibits the flow of ideas and renders expression dull and boring. Martin (1989) maintains that the main cause of redundant repetition seems likely to be limited L2 vocabulary, which prevents L2 learners from employing a wide range of vocabulary items. Fakaude and Vargs (1992) relate the use of redundant repetition to the fact that L2 learners seem to be unaware that one of the most important features of academic writing is to avoid redundant words and expressions. They add that L2 learners might be accustomed to using redundant words in speech and bring this habit into formal writing.

6.2.4 Simple repetition as a cultural and rhetorical device in Arabic language

Researchers such as Mohamed-Sayidina (2010) relate high frequency of simple repetition in the English writing of Arab L2 learners to cultural influences. This factor has been widely examined by researchers in L2 writing and contrastive rhetoric. For example, Mohamed and Omer (2000) examined differences between Arabic and English rhetoric. They suggested that the high degree of simple repetition could be due to the effect of the Arabic culture. One aspect of this culture is the importance that the Arabs give to oral elements of communication. According to these researchers, Arabic written texts still preserve features of oral communication. They claim that this situation is principally because the influence of the Qurān, the Holy Book of the Arabs and Muslims. The Qurān's style, which Arab writers consider as a linguistic model to be imitated, contains a number of features of oral

communication including repetition. Ong (2003) asserts that repetition is one of the main characteristics of communication in ‘oral’ or ‘oralised’ cultures. The use of repetition was a common practice in the Arab educational history when poetry, the Qurān and literature in general were preserved through memorising and rote learning. As a result, such cultural tendencies by the Arabic community, as Mohamed-Sayidina (2010) claims, have influenced the literacy practices in the English writing classroom in the Arabic-speaking countries.

Other studies (e.g., Allen 1970; Al-Jubouri 1984; Sa’adeddin 1989; Williams 1989) that analyse differences between Arabic and English rhetoric have reported that ESL Arab students transfer Arabic rhetorical modes of propositional development, and connectivity (as realised by cohesive devices) into their English compositions. They also transfer other rhetorical strategies, which include writing organisation, thought pattern, style, language, and writer’s perception of cohesion. Allen (1970: 94), for example, observes that Arabic writers organise their English argument by coming to the same point two or three times from different angles, which give impression to a native English reader that there is no an ongoing argument. Sa’adeddin (1989) describes text development by Arab speakers of English as cumulative and additive with such items as nouns, phrases, or clauses added one after another like beads on a string. In describing the thought pattern of Arabic and English, Kaplan (1966) describes Arabic thought as best illustrated in terms of a zigzag line moving gradually from A to B (see Figure 12), whereas English thought moves directly from A to B by means of a straight line (see Figure 13).

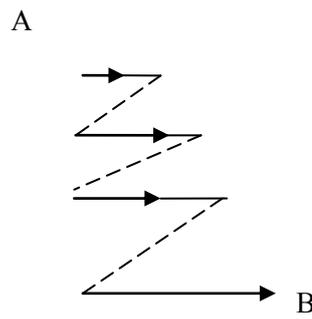


Figure 12 Arabic thought

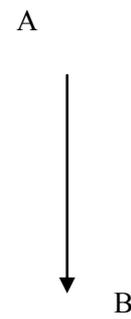


Figure 13 English thought

(Reproduced from Kaplan 1966: 15)

A thought pattern, or what Kaplan (1966) calls ‘rhetoric’, is concerned with how ideas are organised in the mind, which are then reflected in writing. As the diagram above illustrates, Kaplan (1966) identifies thought patterns in different cultural groups by examining how the English paragraph is organised in the English writing of each group. He claims that the English paragraph produced by native speakers of English develops in a straight line, which makes the text argument clear and logical. On the other hand, the English paragraph written by Arab speakers of English, for example, is made up of a sequence of zigzagged lines, because it uses “a complex series of parallel constructions” (Kaplan 1966: 6), which is not acceptable in English discourse. Ostler (1987) argues that the use of parallel constructions is a preferred style by Arab speakers of English, but this does not entail that these learners think in terms of a zigzag line. Several researchers (e.g., Liebman 1992) criticise Kaplan’s (1966) diagram in that it simplifies how rhetoric is described in writing. Liebman (1992: 142) argues that rhetoric is a complex and dynamic notion that is not merely restricted to an arrangement of ideas. However, rhetoric also requires other features such as the inventive thinking, and the interaction between the writer, reader and the world, which helps to deliver a complex piece of writing. Connor (1996) also disagrees with Kaplan’s (1966) assumption that thought

patterns in the writing of ESL students are shaped by culture, and hence each language and culture has its unique thought patterns. Kubota (1999) argues that Kaplan (1966) creates a cultural dichotomy between the East and the West, and his cultural labelling is an oversimplified generalisations of language and culture, portraying ESL students in terms of fixed stereotypical traits, while giving central importance to English writing. Kupota and Lehner (2004) maintain that this separation between cultures prevents discussing different rhetorics in the English writing classroom, and does not take into account the individual variation among ESL writers in terms of the “inventive nature of linguistic and cultural adaptation” (Zamel 1997: 350). Therefore, Kubota (1999) suggests ‘multiculturalism’, as an alternative approach to understanding cultural differences and giving more prominence to the “power struggle within the culture and between cultures” (Kubota 1999: 11). In a later work, Kaplan (1987) admits that it is possible for any language to have multiple cultural thought patterns, and to share many of these patterns with other languages. Nevertheless, each language differs from another in the frequency of specific preferred patterns, which makes each language and culture distinct.

Despite the criticism levelled at Kaplan’s (1966) study on cultural thought patterns, culture is still an important aspect that might have interfered with the English writing of Arab L2 learners, and made their writing different from English speakers in the use of lexical cohesion. However, other kinds of textual analyses are needed, which focus more on the writing process and the relationship between the text and context (Connor 1996), rather than on describing cultural styles between languages. In a different study by Johnstone (1983: 85), she attributes repetition in Arabic to the nature of persuasive discourse, which becomes rhetorically effective through ‘paratactic repetition’. That is, an idea is made believable by being stated,

restated and paraphrased. With a particular focus on argumentative writing, Al-Jubouri (1984) demonstrates that Arabic argumentative texts manifest repetition as an argumentative strategy. While this repetition is acceptable in Arabic, it is not favoured if it is transferred into English, and it will have an impact on text development and its cohesion. As Hatim and Mason (1997: 32) point out, “while recurrence of the lexical item is an option available to users of both Arabic and English, the latter generally see it as a heavily marked form which, to be sustainable, must have some special motivation”. This indicates that English uses simple repetition merely if it performs a linguistic function. Sa’adeddin (1989) concludes that ESL/EFL writing by Arab students displays persuasion devices that are considered inappropriate in English academic discourse and seem to be unlike the persuasive devices in the essays of native speakers of English.

6.3 Derived repetition

The findings of the present study showed that Arab speakers of English used more networks of derived repetition as opposed to native speakers. But, the t-test indicates that this difference is not statistically significant ($t = 1.05, p = 0.25$). Yet, it is still useful to know what has made the Arab non-native learners scored higher frequency of derived repetition than the native group. These findings are not in line with the findings of other studies on lexical cohesion in NNS writing. For example, as discussed in Chapter 2, Kai (2008) found that the percentage of complex lexical repetition (i.e., derived repetition) is much higher in native speaker writing which constituted 46% whereas this percentage decreased to only 4% in non-native speaker writing. Kai (2008) observed that the NS writing shows a relatively higher level of sophistication than that of the NNS writing. Stotsky (1983) explains that the tendency in English is to use complex lexical repetition rather than simple lexical repetition which might

be a salient indication of the text writing development. Also, Reynolds's (1995) findings demonstrated that the overall mean frequency of complex repetition in NS writing was higher compared to that in non-native writing (0.25 vs. 0.12 respectively).

These findings support the common assumption in the area of lexical cohesion that complex lexical repetition is a feature of mature native speaker writing, as Stotsky (1983) points out. Nevertheless, the results of the present study do not confirm this assumption. Therefore, a qualitative text analysis was needed to examine why derived repetition is higher in the Arab NNS writing compared to the NS writing. In order to investigate the difference in the use of derived repetition between these two varieties of writing, I analysed actual examples from both corpora. The examples in Sections 6.3.1 and 6.3.2 are taken from ALEC and LOCNESS. In these examples, instances of derived repetition are in bold. Simple repetition will also be marked (as underlined), whenever relevant, in order to show the density of repetition networks. T-units are numbered based on their sequence in the essay, and they are also separated by two slashes to indicate the boundary of each T-unit. The two square brackets within the examples indicate intervening T-units that are not presented.

6.3.1 Text analysis of derived repetition in ALEC

Example (13) is about 'computer and human brain'.

- (13) 1 One day I was trying to make a simple mathematical **calculation**, // 2 but I struggled and took me awhile to come up with an accurate number, while my younger nephew was quicker in terms of finding the outcome of the multiples I was trying to **calculate**.

In Example (13), the Arab L2 learner attempts to establish lexical cohesion through using a short network of derived repetition that consists of two items: *calculation* – *calculate*. However, the verb *calculate* in T-unit (2) creates merely padding and does not provide a framework for new information to be supplied. Instead, the whole post-modifying clause (*I was trying to calculate*) can be removed because it does not serve any communicative function. Thus, the lexical cohesion in this example is employed unskilfully and creates a clumsiness that unnecessarily labours the obvious. A second example also comes from an excerpt on the differences between computers and human brain. This example illustrates how derived repetition is used by an Arab L2 learner.

- (14) 1 However, in this day and age, the computers provide us with easy **access** to vast amounts of data and information that can be used to **easily solve** some of our most complex problems. [...] 3 The **accessibility** to such information made it **easier** for me to identify reliable **solutions** to the problems I was trying to solve.

At the surface level, Example (14) reveals a density of lexical cohesive networks of derived repetition between *access* and *accessibility*; *solve* and *solutions* and also between *easily* and *easier*. The Arab L2 learner employs derived repetition to connect adjacent T-units but without creating a progression of the main argument because T-unit (3) is merely derived from T-unit (1) rather than actually developing it. The use of the noun *accessibility* in T-unit (3) does not add any new information to the main proposition. In addition, the noun *solutions* also does not have any textual function. It only represents a derived noun of the verb *solve* without suggesting any solution. T-unit 3 is then a rephrase of the first T-unit, which reiterates the same information by repeating a number of lexical items that have already been used in T-unit (1): (*information* – *information*; *solve* – *solve*; *problems* – *problems*). Therefore, repetition in both types (simple and derived) is heavily used in this example carrying an

inadequate textual function. Example (15) is a further illustration of how derived repetition works in text but the topic at this time is about the advantage of rising fuel fees to reduce traffic. The Arab L2 writer in this example discusses the social benefits of rising fuel fees.

- (15) **1** The social factors that raised in such statement can be found in one family for example; who has 3 cars, the father drives on his own, the mother has her own car and the son/daughter uses his/her car, // **2** for this the family will be apart from each other and no interacts between them. // **3** This issue rises and the suggestion of increasing the fuel fees will limit these cars // **4** and all families will be getting one car, // **5** it will keep the family together all the time whenever they travel and keep this kind of perfect relationship we miss these days, // **6** everyone drives his own ways // **7** and there is no more **connection** between family members, // **8** it will increase the chances in camping together and visit the relatives to be **connected**.

As can be seen in Example (15), the Arab subject emphasises the idea that family members will be socially connected if the number of cars has been reduced. In order to convey this idea, the Arab L2 learner uses, for example, multiple networks of simple repetition (*family – family – families – family – family; cars – cars – car; drives – drives*). He/she also uses derived repetition to establish a link between *connection* (T-unit 7) and *connected* (T-unit 8). However, the Arab L2 writer, as the example shows, repeats the main idea many times while he/she could express it in a few lines. For example, T-units (6) and (7) are a rephrase of T-units (1) and (2). T-units (1) and (2) describe how each family member has his own car which, as a result, will loosen the relationship with the family. In T-units (6) and (7), the Arab L2 writer reiterates the same idea in that he/she repeats the phrase *drives his own* (T-unit 6), and also uses the lexical item *connection* (T-unit 7) instead of *interacts* (T-unit 2). Then, the writer in T-unit (8) returns to the main proposition of the social benefits of rising fuel fees and adds two new elements *camping* and *visiting*. However, he/she repeats the idea of connection again

through their use of the lexical item *connected*, which links back to *connection* in T-unit (7). This derived repetition here is clumsily used. The item *connected* (T-unit 8) has no textual value at all. Instead, it is only one constituent of a noisy unsupported statement, as Sa'adeddin (1989) describes.

Examining the same topic that concerns the advantages of rising fuel fees, the subsequent example shows an inappropriate use of derived repetition by an Arab speaker of English. The main proposition in Example (16) is that if fuel fees had been raised, people would think of other inventions and alternative solutions for power. Let's consider the example.

- (16) **1** [...] because of the rising [in fuel fees] there will be more **invention** to replace the use of fuel // **2** and people will start to think for other solutions,[...] **5** It will make people **invent** and think of other solutions for power rather than oil and gas.

In the short excerpt above, there is a textual network of derived repetition between *invention* (T-unit 1) and *invent* (T-unit 5). Also, there is a network of simple repetition that connects two elements *solutions* (T-unit 2) and *solutions* (T-unit 5). With the focus on the network of derived repetition, T-unit (5), which contains *invent* (one member of the derived repetition network), hinges on the same point that has been introduced in T-units (1) and (2). The Arab L2 writer uses abstraction through their use of abstract nouns *invention* and *solutions* without defining these concepts, which as a result creates vagueness in text. Thus, derived repetition in Example (16) creates a case of weak relationship of lexical cohesion because the repetition of the verb *invent* is unnecessarily used. Accordingly, T-unit (5) can be eliminated altogether without affecting the meaning of the passage. The excerpt could be re-written to make it more cohesive than before, as Example (16-a) shows. Grammatical mistakes have not been corrected.

(16-a) [...] because of the rising (in fuel fees) there will be more **invention(s)** to replace the use of fuel // **3** and people will start to think for other solutions for power rather than oil and gas.

In order to provide a range of examples on different topics from ALEC, the following example is about ‘immigration’. Example (17) contains a type of zero derivation (see Chapter 2) in which the lexical item *housing* is repeated in adjacent T-units. One time is used as an adjective or a noun modifier, and the second mention is used as a noun.

(17) **1** More people means more schools, more hospitals and more **housing** capacities. // **2** This high rate of population might result in critical problems in terms of **housing**, hospital and educational institutions. // **3** All of this might create huge pollution and problems to native residents.

Example (17) shows a relationship of derived repetition that connects *housing* in the first T-unit with *housing* in the second T-unit. The use of the second mention of *housing* (T-unit 2) over-asserts what has already been introduced in T-unit (1) that the high rate of population will lead to problems including housing. The Arab L2 writer also uses simple repetition to over-emphasise the main argument that he/she wants to express. He/she achieves this through these networks: *hospital – hospitals*; *problems – problems*. Although the Arab L2 writer introduces one of the problems that will be caused due to the increase of the number of houses (i.e., pollution), he/she uses repetition (simple and derived) redundantly. Example (17) could be simply re-written as:

(17-a) **1** This high rate of population might result in critical problems in terms of housing, hospital(s) and educational institutions. // **3** All of this might create huge pollution and problems to native residents.

As it is clear from the above rephrase of the original example by the Arab L2 learner, the network of derived repetition (*housing – housing*) does not appear at all. This indicates that this relationship is trivial and does not play any communicative role to the meaning of the text, and could hence be discarded. The next example is also about ‘immigration’. The Arab L2 writer employs derived repetition to stress that immigrants will strive to have an entry to the host country by any means. Let’s consider how the writer expresses this idea through the use of repetition.

- (18) **1** As matter of fact the idea of immigration plays big deal for many foreigners // **2** and they will pay all what they have to secure a **clear entrance** to a country for example; the United Kingdom; // **3** many people will try to seek a **legal entry clearance**. // **4** and if they have not been granted a visa then they will seek an alternative in getting in **illegally**.

In the short extract above, there are a number of derived repetition networks: *clear – clearance*; *entrance – entry*; *legal – illegally*. If we look at T-unit (2) in which the two lexical items *clear* and *entrance* occur, we firstly find that the phrase *clear entrance* is not the right construction when the topic of visas and immigration are concerned. Rather, the frequent phrase that is used in this area is *entry clearance*, which the writer himself/herself uses in T-unit (3). However, *entry* and *clearance* in T-unit (3) establish unnecessary relationship of derived repetition with *entrance* and *clear* in T-unit (2). This is because T-unit (3) is plainly a reproduction of T-unit (2) and does not contribute to its development. The only addition in T-unit (3) is the introduction of the adjective *legal* that describes the nature of the entry. The writer could have added this adjective to T-unit (2) to avoid this kind of unnecessary repetition. Thus, T-unit (3) then can be removed and the idea could be re-written more economically as:

(18-a) [...] **2** and they will pay all what they have to secure a **legal entry clearance** to a country for example; the United Kingdom [...]

Then, In T-unit (4) the Arab L2 writer uses the lexical item *illegally* that refers back to *legal* in T-unit (3) and builds up with it a cohesive relationship of derived repetition. This type of derived repetition represents the use of a pair of antonyms that are formed morphologically. The writer's use of this cohesive relationship is to some extent acceptable. Through this relationship, the Arab L2 writer clarifies that not all immigrants enter legally to the targeted country and this is when the government should stop immigration. However, the writer does not elaborate on the illegal ways that immigrants follow to enter the country, but this perhaps because this information is a shared knowledge between readers.

From most of the examples above, the use of derived repetition in the English writing of Arab learners is similar to their use of simple repetition. That is, they use derived repetition to over-emphasise an individual proposition without providing adequate supporting statements that can expand the main argument. These features of over-explication and redundancy have been discussed thoroughly in Section 6.2, and thus there is no need to restate this argument at this point. Let it suffice to mention that the higher rate of derived repetition in English writing by Arab speakers of English, as compared to native writing, could also be attributed to the influence of the mother tongue (i.e. Arabic). Arabic is generally classified as a productive derivational language where many words are derived from the same root. This creates a lexical connection between many items that belong to the same root and thus helps keep the text unified. Al-Jubouri (1984: 100) terms this repetition as 'repetition of root' which is the "multiple use of the same root", as Johnstone (1991: 62) describes. Al-Jubouri (1984: 102) investigates the role of repetition in Arabic argumentative discourse and notes that this root

repetition is frequent in Arabic. It is used to refer to lexical items derived from one root and repeated in one sentence. The most common example of this phenomenon is the ‘cognate accusative’. Johnstone (1991: 63) defines ‘cognate accusative’ as the verbal noun derived from the verb, thus creating a two-faceted repetition (repetition of root and repetition of verb class). This type of root repetition is used in Arabic for emphasis by repeating individual lexical items and ideas. Johnstone (1991), in her analysis of root repetition in Arabic discourse, claims that perhaps the best illustration of the difference between English and Arabic with respect to the repetition of lexical roots is the awkwardness of English glosses like (*differs – differing*) in Example (19):

(19) And so a “Syrian nation” developed there which **differs** from the rest of the nations a fundamental **differing** [differs fundamentally].

(adapted from Johnstone 1991: 64)

In the square brackets in Example (19), Johnstone (1991) shows how this example could be re-written more economically without using repetition. She maintains that English discourse rules encourage writers to avoid repetition of this sort while the situation is the opposite in Arabic. The word *differs* invites the word *differing*. The two words co-occur with each other. In this regard, Dickins et al. (2002: 135) argue that even where English has similar forms, such as (he *drank* a *drink*), there are often more common alternatives such as (he had a *drink*). The following are some examples from the ALEC corpus that show some instances of root repetition. Examples (20) and (21) indicate a type of derived repetition through root repetition (*introduction – introduces; decision – decide*). As the examples below show, these instances of root repetition occur within T-units, while the current study focuses on cohesive relationships between T-units. However, these examples are still worthy of consideration

because they represent a specific linguistic style in Arab L2 writing. Furthermore, some of these within-T-unit lexical connections might have been unconsciously recorded during the process of counting due to the difficulty of identifying the position of each lexical item in the text – whether it is within or between T-units. Accordingly, some of these cohesive relationships might have been counted and affected the overall count of derived repetition.

(20) All in all, it is clear that the **introduction** of a public transportation system also **introduces** a large number of advantages both to the local community and to the more global civilization.

(21) At the end of the day, it is the **decision** of individual persons to **decide** what kind of sports they wish to get involved in.

Examples (20) and (21) illustrate how root repetition occurs in the same T-unit. Dickins et al. (2002: 137) observe that it is also possible to find root repetition in larger stretches of text. This might interpret the density of derived repetition in Examples (13-18) that have been presented in this section. Dickins et al. (2002) point out that this repetition functions in Arabic as a text-building device (i.e. it contributes to the cohesion of the text). However, the textual examples of derived repetition show that when this repetition is used in English writing, it creates unimportant addition to the text.

6.3.2 Text analysis of derived repetition in LOCNESS

This section introduces a number of examples from LOCNESS. These textual examples will help to examine how English native speakers use derived repetition compared to Arab non-native writers. Example (22) is about ‘computer and human brain’.

(22) **1** they [computers] are used to transfer money across the globe, even to create artwork and to **entertain**.// **2** Computer generated pictures, including ‘fractal’ pictures, drawn from equations, seem to be more popular than hand printed images from an artist’s imagination, // **3** and computer games seem to provide more **entertainment** than any game or activity that takes place in the real, physical world.// **4** Virtual reality allows people to have ‘3-D’ entertainment created for them by a computer.

The native writer in Example (22) supports their ideas in the short paragraph above with examples and a proper explanation through using derived repetition. For example, in T-unit (1), the native writer introduces the idea that computers can be used for an entertainment by their use of the verb *entertain*. Then, in T-unit (3), he/she expands this idea and talks about computer games as a kind of this entertainment. In T-unit (4), the native writer repeats *entertainment* to highlight an essential element about computer games, which is ‘virtual reality’ that can create an entertainment through a three-dimensional environment that enables a person to interact with that environment during a game. Thus, we can see how the native writer uses derived repetition as well as simple repetition successfully through this short lexical repetition network: *entertain – entertainment – entertainment*. Example (23) is another example, which illustrates how a native writer discusses the problem of traffic.

(23) **1** We have tried to **solve** this problem by building wider roads and bypasses.// **2** Some councils have tried to show a little more initiative by building bus lanes and trying to entice people to use public transport. // **3** These are mainly short term **solutions** that will eventually make the problem worse. // **4** New roads will just bring new cars. // **5** The problem is not inadequate road systems it’s too much traffic.

Example (23) shows that the text is developed in exemplary fashion in that the native writer suggests a number of solutions to the traffic problem, and he/she then extends their argument by evaluating these solutions. This idea is expressed partially through the use of derived repetition. The native writer builds up a cohesive relationship that connects the verb *solve* in the first T-unit, with the noun *solutions* in T-unit (3). This relationship of derived repetition has a textual function because it has an informational value and it hence contributes to the lexical cohesion of the text. Examining the same topic of ‘traffic’, the following excerpt in Example (24) discusses a proposal on how road traffic could be reduced by suggesting a construction of new roads. The native writer creates a derived repetition network between *constructed* and *construction*. This cohesive network is adequately used because the way it clusters in the short paragraph below helps to connect T-units together and organise the text.

- (24) 1 [...] but the areas where new roads are needed most are in the centre of massive cities where new roads can not be possibly **constructed**, // 2 there is simply not enough room. // 3 Away from cities road **construction** faces the problem of geographical sites.

The following is another example from the topic of ‘traffic’.

- (25) 1 Here the major talking point is **efficiency**. // 2 It is often noted that trains are delayed and are rarely running on time. // 3 Because of this **inefficiency** many of rails customers are turning to other forms of transport, cars mostly // 4 but for long distance journeys flying is an increasingly attractive option.

As Example (25) shows, the main idea of ‘efficiency’ is satisfactorily developed by the native writer. The idea of efficiency is activated by the surface cohesive devices: *efficiency* and its

morphological antonym *inefficiency*, thus creating a short network of derived repetition. These structural forms then operate as channels to convey the meaning of the argument by defining the main factor that has made trains not efficient, and what such *inefficiency* leads people to do. The subsequent example is taken from a different essay that is about ‘boxing’. Let’s consider how the native writer employs derived repetition to create lexical cohesion in the following text.

- (26) **1** One of the main arguments put forward for the abolishment of boxing is the fact that the continual pressure exerted on a boxer whilst being **hit** can result in death, or terminal **injury**.// **2** One **hit** alone is unlikely to **injure** or kill a boxer, // **3** but continual **hitting** can severely **damage** the brain.// **4** The brain is surrounded by **protective fluid**. // **5** This **fluid** is gradually worn down, and eventually leaves the brain tissue with very little **protection**, making it susceptible to serious **damage**.

What is most evident in Example (26) is the linear development of the ideas, which creates a logical means of creating cohesion. The main idea is stated in T-unit (1). Then it is linearly developed by using a density of lexical networks of derived repetition that link the 5 T-units coherently. The cohesive markers that establish the networks of derived repetition are: *hit* (v) – *hit* (n) – *hitting* (v-gerund); *injury* – *injure*; *damage* (v) – *damage* (n); *protective* – *protection*. These lexical forms contribute to extending the meaning of the main proposition in this text, and they also help to build up a unified piece of discourse. A further example that reveals how derived repetition operates in native writing is presented in Example (27). The topic here is also ‘boxing’.

(27) **1** There is always much speculation over the dangers of such a **brutal** sport as boxing. // **2** This is emphasised and exaggerated when a competitor in the sport tragically **dies**. // **3** A recent **death** in the ring has inevitably led to a public uproar on the safety of the sport, // **4** and the controversy over whether the sport should be banned or not is yet again the forefront of discussion. // **5** Let us consider how a professional boxer would feel. // **6** He is clearly aware of the dangers and **brutalism** of the sport, which is possibly why he enjoys it so much.

In Example (27), each T-unit is arranged next to each other, thus forming progressive logical series of ideas. In T-unit (1), the native writer describes boxing as dangerous and brutal. Then in the subsequent T-unit, he introduces the idea of death, which is a result of this violent sport. The idea of death is then highlighted in T-unit (3), which opens a debate on whether this sport should be banned or not. In T-unit (6), the writer repeats the idea of brutality but this time he/she argues that it creates a feeling of enjoyment among boxers. The writer connects their ideas in this short text by the use of lexical networks of derived repetition that links the following items: *brutal – brutalism; dies – death*.

In T-unit (6), the word *brutalism* is used in the right word class in the text above. However, it is not frequently used to describe dangerous sports. Rather, it is mostly used in the field of architecture. The native writer needs to use the noun *brutality* instead to refer back to the adjective that he/she uses in the first T-unit (i.e. *brutal*). Therefore, it could tentatively be claimed that the native writer, on the one hand, achieves lexical cohesion properly at the semantic level. On the other hand, the writer selects an inappropriate word form as regards the right context. Example (28) further demonstrates the above mentioned case in respect to the derivation of the appropriate word-form.

- (28) **1** the only way to stop people boxing is to ban boxing and make it **illegal**.// **2** The **illegalisation** of boxing I believe however could only lead to the introduction of underground boxing matches if out of sight of the law.

In Example (28), the native writer discusses how to stop boxing and make it illegal. In so doing, the native writer uses the wrong word-form in their establishment of derived repetition. On the one hand, the writer successfully creates a derived repetition network between *illegal* and *illegalisation*. The noun *illegalisation* links back to *illegal* and carries new information by clarifying what could happen if boxing has been made illegal. However, the native writer coins the noun form *illegalisation* from the adjective *illegal* while this noun form is not used in English, as the BNC shows. These derivational errors could be considered as form-oriented and are due to wrong derivational process. It might be argued that problems such as these are not really problems with the cohesive function of derived repetition, but are more appropriately considered as problems of word formation. However, more research may be needed to analyse inappropriate word-forms of derived repetition in NS and NNS writing.

The next set of examples (29-31) is drawn from different NS writers on essays about ‘boxing’ and ‘traffic’. These examples further demonstrate the way derived repetition is used in LOCNESS to create lexical cohesion.

- (29) **1** However there is a **tragic** side to the sport, upon which the anti-boxing lobbyists base their opinions.// **2** The **death** of Scottish welterweight Jim Murray in the ring as recent as this October have renewed calls for the immediate banning of boxing. // [...] **12 Tragedies** will always happen in any sport not only in boxing. // Paul Bennett, a young striker at York City **died** during a match a few years ago when he swallowed his tongue.

- (30) **1** During a fight a boxer may receive several hundred **punches** to the head, // **2** and each time that he gets **punched** he loses more and more brain cells.
- (31) **1** All this leads to trains being **unprofitable**, except on a few major commuter routes. [...] // **7** The only way to make **profits** is to get more people on trains, and this requires an improvement in services.

In Example (29), lexical cohesion is established through two networks of derived repetition: *tragic – tragedies; death – died*. Example (30) contains one derived repetition network that consists of two items: *punches – punched*. In Example (31), there is a case of lexical cohesion through using antonyms that are formed through affixes (*unprofitable – profits*). Each lexical network of derived repetition in these examples consists of lexical items that are interconnected in such a way that creates textual continuity. This indicates that each network of derived repetition in each example serves as a carrier of new information which, as a result, helps maintain and develop the communicative function of the text.

To conclude Section 6.3 on derived repetition, we find that the high frequency of derived repetition in the argumentative essays written by Arab speakers of English is not always a guarantee of cohesive writing. This is because Arab L2 learners do not use derived repetition to serve a function in their context. As highlighted in Chapter 2, Stotsky (1983) clarifies that derived repetition in English is used to avoid the clumsy juxtaposition of the same lexical item in a passage. Hoey (1991b) further supports this view and suggests that one of the most important ways for a writer to avoid clumsiness is by means of derived repetition. Nevertheless, derived repetition in the English essays written by Arab L2 learners is clumsily used without serving a textual function or helping to develop the argument. On the other hand, despite the less frequent use of derived repetition in the essays produced by English writers,

their writing looks more cohesive because these writers employ derived repetition meaningfully in order to connect their T-units and add variety to their text. The only problem that some of the examples reveal in the NS writing is that native speakers of English have a problem of word formation (e.g., *illegal* – *illegalisation*). However, as mentioned earlier in this section, this problem does not affect the cohesion of the text.

6.4 Signalling nouns

Signalling nouns are the third lexical cohesive category investigated in the present study. Unexpectedly, the results of the current study suggested that Arab speakers of English used signalling nouns more frequently than did native speakers of English, and this difference proved to be significant ($t = 2.30, p = 0.024$) (see Chapter 5, Section 5.9.2). These results do not conform to previous studies into signalling nouns in NNS writing (e.g., Petch-Tyson 2000; Hasselgård 2012; Flowerdew 2010; Forutan and Nasiri 2011) that found that English native speakers used SNs more frequently than non-native speakers of English. Aktas & Cortes's (2008: 7) study is an exception. As discussed in Chapter 2, their results showed that non-native writing contained a higher frequency of shell nouns (SNs in the present study) compared to native writing which exhibited fewer tokens of these nouns. This difference was emphasised by the result of the chi square test: $X^2 = 75.381$ at $p < 0.01$. Nevertheless, Aktas & Cortes (2008) discovered that this difference in the frequency counts might be due to the distribution of shell nouns in learners' writing which could create an idiosyncrasy that might have increased the frequency counts in the non-native writing. By contrast, these researchers observed that native writers also used shell nouns more than once in their writing but the frequency of types was "better distributed across authors" (Aktas & Cortes 2008: 9).

The discussion above indicates that using frequency analysis alone can give a false picture of the nature of the difference between NS and NNS writing. This is because, in some cases, the difference in frequency could be a result of different factors not related to the good performance of NNS writers. Even studies whose quantitative results showed that SNs were used more frequently in native writing compared to non-native writing, their qualitative analysis suggested that the frequency measure did not always provide accurate results. As demonstrated in Chapter 2, Petch-Tyson (2000), for example, found that the use of anaphoric demonstrative expressions which have a signalling function was only frequent among native writers whereas non-native students preferred reference to other noun phrases. However, her qualitative analysis showed a resemblance in terms of the high frequency of demonstrative noun phrases between native English and French L2 learners compared to other non-native L2 sub-corpora. This high frequency, according to Petch-Tyson (2000), did not indicate a near-native use. Instead, it reflected the way by which non-native and native texts used demonstrative nominal anaphors to encapsulate previous discourse. Examples (32) and (33) demonstrate such a difference. The italicised phrases are the anaphoric demonstrative expressions while the underlined sentences are their referents.

(32) You can *escape* from the reality once in a while and hide yourself in the beautiful world of imagination. [...]. In my opinion, *this kind of escaping* can actually be therapeutic [...]
(Petch-Tyson 2000: 58)

(33) It is important to stress the terrible accidents that occur with other power plants when *arguing for nuclear power.* *This tactic* is very good [...]
(Petch-Tyson 2000: 59)

Example (32) shows how non-native writers use demonstrative nominal anaphors (e.g., *this kind of escaping*) to nominalise verbs (e.g., *escape*) that are previously mentioned in discourse. In contrast, Example (33) illustrates that native writers use demonstrative nominal anaphors metadiscursively to label previous textual segments (Petch-Tyson 2000: 58). This means that the native writer expresses their stance towards the previous discourse by selecting *tactic* that condenses the earlier information into one fresh concept. This concept has been mentioned for the first time and has not been repeated. Section 6.6 will show that these two examples have practical implications for the use of signalling nouns by Arab speakers of English and English native speakers in the present study.

As discussed in Chapter 2, researchers also bring up an important point that relates the high frequency of SNs in non-native writing to factors such as L1 influence. In this respect, Hasselgård (2012: 26) claims that first language (L1) might have a positive transfer of some linguistic features on learners' performance principally if the two languages are very close. Researchers also consider other aspects that might detect the difference between NS and NNS writing in their use of SNs. For example, Flowerdew (2010: 44) did not rely on overall frequencies of SNs to differentiate native writers from non-native writers. Rather, he examined the functions of SNs in NS and NNS writing. His findings, for instance, showed that native writing contained more across-clause anaphoric and in-clause SNs, while learner writing showed a greater frequency of across-clause cataphoric instances. Flowerdew (2010: 48-49) observed that this reliance on cataphora might be a result of the text format of the essays which require eliciting, comparisons and opinions (e.g. discuss the advantages and disadvantages). Consequently, Flowerdew (2010) suggests that frequency information may give an exaggerated impression of the learners' ability to use across-clause SNs. In the same

line of thought, Forutan and Nasiri (2011) stress that quantitative results need to be supported by qualitative analysis. In order to notice differences in the use of SNs between native and non-native writers, Forutan and Nasiri (2011) identified specific types of SNs that can distinguish each variety of native and non-native writing. This procedure will be used in Section 6.7.

Most studies discussed above draw attention to the importance of integrating frequency data with a qualitative analysis that can help recognise the difference between native and non-native English writers in their use of SNs. This indicates that the findings of the present study which showed an increase in the frequency figures of SNs in the Arab NNS writing compared to NS writing should be treated with care before any generalisation of the results is presented. Therefore, in the following sections, I tested some factors that might account for the observed difference in the frequency counts between the two groups in their use of SNs. Firstly; this difference might be due to the counting method that I suggested in Chapter 2. Alternatively, the difference is perhaps a result of one of the factors mentioned above that other researchers also observed in their studies such as L1 influence, for instance.

6.5 Reconsidering the counting method of SNs in this study

During the identification process of SNs in ALEC and LOCNESS, a number of SNs were repeated or appeared in certain linguistic structures particularly in the Arab learner writing. The existence of these structures led me to follow a specific method of counting that takes into account these uses of SNs. This procedure of counting could, therefore, have affected the final frequency of SNs in the two corpora and produced an increase in the number of SNs in the Arab learner writing. One example of these structures, as demonstrated in Chapter 2, is the

use of lexical strings/couplets that are made up of two or more signalling nouns (see Section 6.5.2). In Sections 6.5.1 and 6.5.2, I focus on two linguistic structures that contain SNs. I mainly examined whether or not the rules I set for counting SNs in these structures proliferated the total count of SNs in the Arab learner writing. To do so, I provided a frequency count for each type of these structures in both ALEC and LOCNESS. I then set the frequency of these structures aside from the total count and observed how frequency counts changed.

6.5.1 Structure one (In-t-unit SN + across-t-unit SN)

One of the salient differences between English argumentative essays written by Arab speakers of English and English native speakers is the predominance of this construction in the writing of the Arab NNS group: in-t-unit SN + across-t-unit SN. This structure contains an in-t-unit SN, which is followed by an across-t-unit SN. The latter SN has the same word-form of the former and both SNs share the same referent. Example (34) demonstrates this structure which contains *policy* in T-units (1) and (2). The underlined stretch of discourse is the lexical realisation of *policy*.

- (34) **1** Besides, one needs to predict the outcome of a *policy* that promotes less fuel subsidies. // **2** In my opinion, the impact of such a *policy* depends on the economic outlook of the country where this policy is introduced. (ALEC)

In this structure, Arab speakers of English repeat the same SN in adjacent pairs of T-units. The first occurrence of the SN functions as an in-t-unit label (e.g., *policy* in T-unit 1), and the other mention as an across-t-unit label (e.g., *policy* in T-unit 2). The rule that I set up in Chapter 2 suggests that the second occurrence, which is the across-t-unit SN, needs to be

counted even though it represents a case of simple repetition more than being an SN. This decision has been justified in Chapter 2, and it is perhaps one reason for increasing the frequency count of SNs in the Arab learner writing. Consequently, I firstly quantified the relative overall frequencies of the ‘in-t-unit SN + across-t-unit SN’ structure in ALEC and LOCNESS. This enabled having a general picture of how frequent this structure is in the writing of NSs and NNSs (for more detail on how relative overall (aggregated) frequencies is calculated, see Chapter 5). Table 18 demonstrates overall frequencies of the ‘in-t-unit SN + across-t-unit SN’ structure per thousand words.

Structure type	ALEC	LOCNESS
	Absolute (relative overall frequency)	
‘In-t-unit SN + across-t-unit SN’	11 (0.62)	2 (0.12)

Table 18 Overall frequencies of the ‘in-t-unit SN + across-t-unit SN’ structure per thousand words in ALEC & LOCNESS

Table 18 shows that this kind of structure hardly ever occurs in the writing of the native group, whereas it occurs in the Arab non-native learner group more frequently 5 times than of the NS group (0.62 vs. 0.12). Thus, this juxtaposition of SNs in adjacent T-units in the writing of the Arab non-native group might be one reason for the high frequency of SNs in their writing. Then, I isolated occurrences of the ‘in-t-unit SN + across-t-unit SN’ structure from the individual essays in which this structure takes place. This step helped calculate the individual text-based frequency of SNs again, but without including the SNs that appear in the structure in question. By applying this procedure, I was able to see whether the counts of SNs

between NNSs and NSs changed or remained the same after eliminating the structure under investigation. Table 19 reveals the relative individual text-based mean frequencies of SNs in ALEC and LOCNESS after ruling out the frequency of the structure in question from each corpus.

Category of lexical cohesion	ALEC		LOCNESS	
	Individual text-based mean frequencies	SD	Individual text-based mean frequencies	SD
Signalling nouns	8.47	4.7	6.35	3.8

Table 19 Individual text-based mean frequencies of SNs per thousand words in ALEC & LOCNESS excluding the ‘in-t-unit SN + across-t-unit SN’ structure

As shown in Table 19, the mean frequency of SNs in the individual essays by Arab NNSs ($M = 8.47, SD = 4.7$) is approximately 1 time that of NSs ($M = 6.35, SD = 3.8$). An independent t-test shows that there is no significant difference in the use of SNs between the two groups when we exclude this structure ($t = -1.8, p = 0.06$). This indicates that by excluding instances of the ‘in-t-unit SN + across-t-unit SNs’ structure from the frequency count of SNs, and re-performing the individual text-based frequency, the final frequency output changes and the difference between the two groups becomes no longer significant as it has been proved to be in Chapter 5, Section 5.9.2. Examples (35-39) are further illustrations of the ‘in-t-unit SN + across-t-unit SNs’ structure in the ALEC corpus:

- (35) **1** This is because of the too much use in computer *applications* like spreadsheets and MATLAB. // **2** These *applications* do simplify the task and can do more tasks in time if I planned to work it out using my brain and peace of paper.

- (36) **1** Brain usually control and manage many *functions* in the human body including; information processing, motivation, learning and memory, and motor control. // **2** To be able to understand how the brain can perform all of these *functions*, it might be worthy to review the biology of the brain and how does it work.
- (37) **1** Neuron's *task* is to transmit electrochemical signals to other cells and to respond to similar signals received from other cells. // **2** The brain's function depends on the ability of the neurons to do the previously mentioned *tasks* appropriately.
- (38) **1** Researchers have shown that with new activities, there is a measured *changes* in brain electrical activity which might indicate the learning of new experience. // **2** In my opinion, these new *changes* in the brain can be an evidence to show that the ability of the brain to absorb new information is increasing with every new experience.
- (39) **1** Unlike human brain, computer lack *senses* such as sights, hearing, smell, taste and touch. // **2** These *senses* provide human brain with details that can help in understanding the big picture of a particular situation.

Example (40) is from LOCNESS:

- (40) **1** The green house effect is a major *issue* to most citizens // **2** but very few of them are prepared to give up their car to show it. // **3** the government however, is beginning to address *this issue*.

From the multiple examples presented above, I could tentatively claim that this structure plays a part in increasing the frequency count of SNs in the Arab learner writing. In Section 6.5.2, I examined another common structure, which was another reason for the frequency increase of SNs in the Arab learner writing. This structure is the 'lexical couplets/strings'.

6.5.2 Structure two (Lexical couplets or strings)

As mentioned in Chapter 2, another observed linguistic structure particularly in the English writing of Arab L2 writers is the use of lexical couplets and strings. Lexical couplets or strings, as defined by researchers such as Johnstone (1983) and Rieschild (2006), refer to a structure that contains two or more semantic neighbours which are co-ordinated with *and*, and have a single referent. The ALEC corpus contains some examples of this structure in which Arab L2 learners coordinated two or more SNs, which are synonyms or near- synonyms (e.g., *hatred and animosity*). Both SNs in this structure enclose the same stretch of text where only one SN would do (see more examples below).

According to the counting criterion I identified in Chapter 2, I counted all elements in the couplet or in the string (A and B; or A, B and C). This might have resulted in an increase in the count of SNs in the Arab learner writing giving that NS writing has few cases of this structure, if any. Therefore, in order to verify whether or not this counting method has affected the total count of SNs in both corpora, I counted the overall frequency of lexical couplets in both ALEC and LOCNESS per thousand words. Table 20 illustrates this frequency.

Structure type	ALEC	LOCNESS
	Absolute (relative overall frequency)	
'Lexical couplets/strings'	5 (0.28)	0

Table 20 Overall frequencies of lexical couplets per thousand words in ALEC & LOCNESS

Table 20 shows that the raw frequency of these couplets in the Arab NNS group is 5, while there is no any case of this structure in the NS group. Each couplet incorporates at least two signalling nouns. This means that 5 couplets contain 10 signalling nouns. I excluded these instances from the ALEC corpus. This procedure helped re-calculate the individual text-based frequencies of SNs in ALEC and LOCNESS. To achieve this, I considered the couplet as one lexical unit, and I as such took off 5 tokens only from the individual essays that contain these couplets. Table 21 illustrates the relative individual text-based mean frequency of SNs in ALEC and LOCNESS after I removed the frequency of lexical couplets from the Arab NNS group.

Category of lexical cohesion	ALEC		LOCNESS	
	Individual text-based mean frequencies	SD	Individual text-based mean frequencies	SD
Signalling nouns	8	4.29	6.4	3.8

Table 21 Individual text-based mean frequencies of SNs per thousand words in ALEC & LOCNESS excluding the ‘lexical couplets’ structure

As obvious from Table 21 above, the frequency of SNs is still higher in ALEC compared to LOCNESS even after excluding instances of lexical couplets found in the Arab learner writing. However, an independent t-test suggests that there is no significant difference in the use of signalling nouns between the two groups when we eliminate this structure ($t = -1.63, p = 0.10$). Thus, extracting lexical couplets from the count of SNs affected the degree of difference in the use of SNs that have been found as significant in the previous chapter when this structure was counted (see Chapter 5, Section 5.9.2). Examples (41) and (42) represent

lexical couplets/strings in the ALEC corpus. Lexical couplets are italicised while their referents are underlined:

(41) **1** The wrong employing of computer and social networks per se encourage *hatred and animosity* // **2** some people insult or make fun of others due to their colour, race, religion, or denomination.

(42) **1** With regard to the *positive impacts and advantages* that immigrants can bring to those host countries; there are many.

In Example (42), *positive impacts and advantages* refer cataphorically to the same content, which is introduced in the subsequent discourse. To give another example that shows this phenomenon of juxtaposition of signalling nouns within the same T-unit, Example (43) includes three signalling nouns with an anaphoric function:

(43) **1** Based on the above mentioned *positive points and advantages* of immigration flows, it may not be very wise to stop such *empowering gains*, // **2** this is because their benefits to the host country may outweigh their drawbacks.

The lexical couplet *positive points and advantages* occurs in T-unit (1) and consists of two SNs that encapsulate the same stretch of text, which is mentioned in the preceding discourse. In the same T-unit, there is another SN with a modifier: *empowering gains*. This phrase that contains an SN with its modifier is a synonymy to the previous couplet and it also shares with it the same referent. This type of structure, as Williams (1989) observes, could create with the previous couplet a type of lexical string, which covers cases where a string of two or more words is repeated and refer to the same information. In such a case, the different forms of SNs

(i.e. *points*, *advantages* and *gains*) that cluster together in the same T-unit are all counted according to the counting rules that I defined in Chapter 2.

This phenomenon of lexical couplets has been examined by Johnstone (1983: 51), who attempted to find an answer to this question “why do people keep using two words where one would do?” As part of a larger study of the nature and function of repetition in Arabic, Johnstone (1983) found that Arab L2 learners tend to string together a number of synonyms in the form of ‘lexical couplets’. These Arabic couplets, as Johnstone (1983) observed, are characterised with their creativity because they are a result of a still-productive semantic strategy. Johnstone (1991) described lexical couplets as example of ‘paradigmatic patterning’ which is created due to the repeated juxtaposition in discourse. These repeated patterns indicate a favourable rhetorical style in Arabic language and have a rhetorical function in discourse which is to achieve persuasion in the Arabic texts. Lexical couplets are further considered as a stylistic embellishment (Shalaby et al. 2009). Williams (1989: 164) stresses the fact that these lexical strings are not only used for ornamental purposes in Arabic, but considered essential to the cohesion of the text. However, this feature of lexical strings/couplets as a signalling device displays verbosity in the learners’ English writing because English prose, as Johnstone (1983) highlights, considers lexical couplets as a stylistic flaw. Nevertheless, she maintains that this linguistic structure may also be one aspect of a key process in the evolution of language: “the continual structuring and restructuring of the paradigmatic building blocks” (Johnstone 1983: 60).

I can conclude that the method applied to count SNs that are part of the two structures introduced in Sections 6.5.1 and 6.5.2 is one reason for increasing the frequency counts of

SNs in the writing of Arab speakers of English. The investigation of the counting method showed that when I take these structures out from both ALEC and LOCNESS, the difference between NSs and NNSs in the use of SNs becomes statistically insignificant. This indicates the importance of these structures to the writing of Arab speakers of English. These structures show a tendency by Arab L2 writers to use patterns of repetition in their writing, which reflects a linguistic style that is unique to this variety of writing. Section 6.6, shows other tendencies of repeating SNs by Arab L2 learners. These tendencies might be other factors that have contributed to increasing the frequency of SNs in the Arab L2 writing compared to the English native writing.

6.6 Other patterns of repetition in the use of signalling nouns by Arab speakers of English: Text analysis

Although the frequency of the two structures introduced in Sections 6.5.1 and 6.5.2 is not high, the results showed that these structures are still responsible for increasing the overall frequency of SNs in ALEC and LOCNESS. If the corpus size were bigger, these structures would most likely be more prominent in the ALEC corpus. The textual examples drawn from the ALEC corpus showed that these two structures are typical linguistic features in the Arab NNS's writing and they both share a common function, which is repeating signalling nouns within a short span of text. By contrast, these patterns as reported in the previous sections are rarely to occur in the NS writing. In this section, I had a further qualitative look at the data and considered other uses of SNs that might reflect a specific style of English writing by Arab L2 learners compared to native writers.

i. Delaying the lexical specification of signalling nouns

The English writing by Arab L2 learners displays another feature of using SNs in adjacent T-units, which is the delay of introducing the lexical specification or resolution of SNs (see Chapter 2, Section 2.7.1.1 for more detail on this feature). A case in point is where Arab NNS writers use a cataphoric SN, but tend to hold up its resolution for a while. That is, before introducing the SN content immediately after mentioning the cataphoric SN, Arab L2 learners suspend its resolution by negotiating the meaning and adding a further discussion in which they reutilise the same form of the SN until they reveal its content. Below is an example from the ALEC corpus which demonstrates this feature.

(45) **1** There are so many *factors* that affect the quality of education. // **2** These *factors* that interlinked and can be very complex to analyse.// **3** The educational institutions can't be analysed in isolation, but rather in the environments where they operate. // **4** For example, governments' policy towards higher education is likely to have an influence on the quality of education. (ALEC)

In Example (45), T-unit (1) contains *factors*, which is an SN with a cataphoric function. The Arab L2 writer starts introducing the lexical resolution of this SN in T-unit (3), and then he/she completes the resolution in T-unit (4). In T-unit (4), the writer mentions, for example, the 'government policy' as one factor that affects the quality of education. However, the writer delays this lexical resolution of the noun *factors* by adding T-unit (2) as an intervening unit between the first mention of *factors* and its resolution. More particularly, in T-unit (2), the Arab L2 writer repeats the SN *factors* that he/she firstly mentioned in T-unit (1). The second mention of the noun *factors* acts as a holder till the resolution of the first mention of

factors is resolved. In Example (45), the Arab L2 writer could simply rephrase the statement using the noun *factors* once only by removing T-unit (2):

(45-a) **1** The quality of education has been affected by so many factors that are interlinked and can be very complex to analyse. // **2** The educational institutions can't be analysed in isolation, but rather in the environments where they operate. // **3** For example, governments' policy towards higher education is likely to have an influence on the quality of education.

This feature of delaying the resolution of SNs appears also in Example (46):

(46) **1** The other side of the argument state that immigrants can be the cause of many *undesirable and harmful effects*. // **2** Some people claim that troops of immigrants can be one among many factors that leads to *negative circumstances*. // **3** First of all, there are arguments emphasizing that immigrants compete with local people over facilities and life needs. // **4** Schools, for example, can be full with immigrant's children. // **5** Immigrants will share busses, jobs, health services and every other facility that governments may provide for their local citizens [...]

In Example (46), the Arab L2 writer uses the SN *effects* in T-unit (1). The writer then rephrases the first T-unit by adding T-unit (2) that contains another SN *circumstances*. The SN *circumstances* is a near synonym of the SN *effects*. The second T-unit is adding needless and repetitive information to the T-unit (1). The first T-unit will be comprehensible without the second one. After T-unit (2), the Arab L2 writer introduces the lexical realisation of the SN *effects* in T-units (3), (4) and (5). This use of SNs forms a particular construction in the NNS writing where Arab speakers of English prefer to come to the point in the end compared to NS writers.

It is relevant to pay attention at this point to the text span factor providing that most of the observed SNs in the Arab L2 writing spread over short T-units. Halliday and Hasan (1976: 339) refer to the text span factor as the number of sentences between cohesive items and referred items. Some researchers apply this notion to investigate its correlation with the quality of L2 learner essays. For example, Witte and Faigley (1981: 196) find statistically significant differences between the distance of text span and quality of essays. They consider that immediate cohesive ties are an indication of good writing because students try to remain longer on a topic using strong cohesive bonds that extend and modify the topic. On the other hand, Hoey (1991b) finds that remote cohesive ties point to a more effective writing compared to the immediate ties that could make writing clumsy. Even though the present study does not examine whether cohesive relationships in text are immediate or remote, it seems that Arab English writers tend to use SNs over short distances, which does not help to build up a new argument.

ii. Signalling nouns with vague lexical specification

A further use of SNs repetitively in the Arab L2 writing is that Arab L2 learners use more than one SN, particularly in the introduction part, with a vague and fuzzy lexical specification. More particularly, it is difficult to identify the amount of text that these SNs prospect or encapsulate. This fuzziness of SNs stems from their encapsulation of long stretches of text, which makes the texts written by Arab writers hard to process. Example (47) illustrates this use:

(47) **1** Thinking of using the very available technologies in our daily life has its *advantages* and *disadvantages* and its *effects* on our brain, // **2** to discuss this topic there are a lot of *facts* which can be addressed to make good argumentative case. (ALEC)

Example (47) shows the introduction part of an essay that is written by an Arab L2 learner. In this introduction, the Arab L2 writer uses a cluster of SNs: *advantages*, *disadvantages*, *effects* and *facts*. The lexical content of these SNs in the subsequent discourse is hard to specify. This is because there is an overlap of the lexical specification of these SNs in the text, hence identifying the lexical specification for each SN cannot be recovered. In other words, as Schmid (2000: 363) puts it, a clear semantic match is not formed between the referent and the noun; or sometimes the lexical specification is not mentioned at all by the Arab L2 writer. This makes the analyst in doubt whether they count these cases of nouns as SNs or consider them as general abstract nouns. In Chapter 2, I counted these cases as SNs and therefore the frequency of SNs have been identified as 4 in Example (47). This repetition of SNs that have vague referents is not common in NS writing. Consequently, it could be another factor contributing to the higher frequency of SNs in the ALEC corpus.

iii. Repeating signalling nouns by using near synonymous labels

Another tendency of repeating SNs in ALEC is when Arab L2 learners use two different cataphoric SNs with clear and different lexical realisations. However, these learners add a short summary comment after the full lexicalisation of these SNs in the subsequent discourse. This summary contains another SN, which constitutes with its pre-modifiers a synonymous label to the previously mentioned SNs and adds no new information. An example of this use is illustrated in the following extract:

(48) 1 [...] there are a *downside* to go all these immigrants to the UK, // 2 perhaps they make a damage to local people, traffic congestion and inflation. // 3 In contrast developed countries has many *benefits* of the immigration, // 4 if we take a foreign students for example they pay plenty of money for their education what is more

immigrants give money for taxes, food health etc.// 5 To be more precise immigration has *positive and negative impact* on the West. (ALEC)

In Example (48), the Arab L2 learner uses two different SNs *downside* (T-unit 1) and *benefits* (T-unit 3) with clear and different lexical realisations (underlined). However, the final T-unit in the extract above is merely a summarising T-unit of the previous stretch of text in which the writer rephrases what he/she has already mentioned in the earlier T-units. In the final T-unit, the Arab non-native writer uses another SN *impact* with two pre-modifiers (*positive and negative*). This SN with its pre-modifiers functions as a near synonymous label that encloses the previous SNs. To put it more clearly, *negative impact* is a synonym of *downside*, whereas *positive impact* is a synonym of *benefits*. Thus, the last T-unit is only repetitive and adds no new information to the text and it could be therefore removed without affecting the textual cohesion of the text.

iv. Nominalisation as a signalling mechanism

The quantitative results of the use of SNs in the Arab L2 writing and the English native writing might also have concealed an important difference in the way in which previous stretches of discourse are being labelled. This use of SNs has been highlighted by Petch-Tyson (2000) in Section 6.4. The ALEC corpus shows a number of instances where Arab speakers of English use SNs to nominalise verbs or adjectives previously referred to in the text and which thus add no information to the text. Although the present study counted the feature of nominalisation under the category of derived repetition, it is still useful to demonstrate how Arab L2 learners use SNs as nominalised labels to encapsulate the stretch of discourse. There are only few examples of this type of nominalised labels of SNs in the

LOCNESS corpus. Native speakers of English use SNs as labels that are rich in semantic potential in that they “both change perspective and provide information” (Mauranen 1993: 79). This implies, as Mauranen (1993) describes, that an SN not only encapsulates the stretch of discourse, but also contributes to evaluating it in the text. Below are some examples that illustrate the different uses of labelling in ALEC and LOCNESS.

(49) 1 [...] the current economical climate in the UK is very *hostile*, // prices are sharply increasing // 2 and wages are not keeping up with inflation levels, // 3 and everyone in the country, including immigrants, are feeling the full force of *this economic hostility*, // 4 and basically everyone is getting *frustrated*. // 5 Sadly, as a result of this emotional *frustration* felt by the native population, immigrants are unfairly attributed to the country’s economic misfortunes. (ALEC)

(50) 1 As far as controlling immigration is concerned, I think that is achievable through closely *examining* the reasons that make immigrants leave their native lands and determining which immigrants are most useful to one’s country’s economy, culture, peace, security and most of all the overall health and wellbeing of its native people. // 2 The outcome of this *examination* should be used to draw government policy on immigration. (ALEC)

(51) 1 [...] but if the boxer does not recover, some peace of mind can be found in the fact that the boxer was aware of the dangers posed to himself, was aware of the help at hand and that the best was done to save his life or retain conciousness. // 2 This may seem to be *a cold hearted viewpoint* [...] (LOCNESS)

(52) 1 [...] Cities such as Manchester have become largely pedestrianised and have seen the reintroduction of trams, // 2 this is the *sort of scheme* that can solve our problems. (LOCNESS)

Benitz-Castro and Thompson (2015: 396) summarise this observation about how SNs are used by positing that the non-native use of SNs is cohesive by near-repetition, as in Examples (49) and (50). These two examples contain nouns such as (*hostility, frustration and examination*) which they encapsulate previous mentioned discourse, but they add no fresh information to the text because they had previously been referred to in the text through their adjectives and verbs (i.e. *hostile, frustrated, and examining*). In contrast, Benitz-Castro and Thompson (2015) observe that the native use of SNs is cohesive by a more evaluative shell-like encapsulation, as in Examples (51) and (52). In these examples, the native writer uses the SNs with evaluative pre-modifiers (e.g., *a cold hearted viewpoint, sort of scheme*) to characterise the previous stretch of discourse (underlined).

Overall, the Arab L2 learners in ALEC perhaps resort to use the nominalised form of the verb to avoid any cognitive effort that is required to come up with an appropriate proposition that might be used to encapsulate the foregoing stretch of discourse. On the other hand, the native writers in LOCNESS show more creativity and more cognitive complexity in their choice of a specific signalling noun. If instances of nominalised labels of SNs had been counted in ALEC as SNs, they would have made the frequency count greater than before.

6.7 The most common types of signalling nouns in ALEC and LOCNESS

In order to find other areas of difference between ALEC and LOCNESS in terms of the use of SNs, I examined the specific types of SNs that Arab L2 learners and English native speakers use in their argumentative writing. This comparison was made by rank-ordering the SNs in each corpus in terms of frequency. Table 22 shows the most commonly used SNs in each corpus ordered according to frequency.

No.	ALEC	Freq.	LOCNESS	Freq.
1	impact/s	11	problem/s	20
2	reason/s	10	reason/s	8
3	effect/s	6	solutions	6
4	issue/s	6	view	5
5	views	5	things	4
6	fact/s	4	argument	4
7	advantages	4	ways	4
8	thing/s	4	measure/s	3
9	consequences	4	effect	2
10	activity/s	4	changes	2
11	disadvantages	3	tasks	2
12	factors	3	factors	2
13	opinion/s	3	side	2
14	argument/s	2	viewpoint	2
15	benefits	2	issue	2

Table 22 The most frequent SN types in ALEC and LOCNESS

Table 22 reveals a fair degree of conformity in ALEC and LOCNESS in terms of the selection of the most frequent SNs. Both corpora, for example, share the following seven SNs in their top fifteen: *reason*, *effects*, *issues*, *views*, *things*, *factors*, *arguments*. There are a number of lexical items in the top-15 list, however, that occur in one corpus but not the other; the following eight items only occur in LOCNESS: *problem/s*, *solutions*, *ways*, *measures*,

changes, tasks, side and viewpoint. On the other hand, the following eight items only occur in the top-15 list in ALEC: *impact/s, fact/s, advantages, disadvantages, benefits, consequences, activities and opinions*. If we look carefully at the nouns in Table 22, we notice that Arab non-native learners make use of SNs that are near synonyms such as (*impact/s, effects, consequences*); (*advantages, benefits*). These near synonyms are not so prominent in the top 15-list of the LOCNESS corpus. Besides, from the list of SNs in ALEC, it seems that Arab L2 learners prefer to use a pattern of SNs such as the pair of *advantages and disadvantages*, which does not appear at all in LOCNESS. This finding is in line with Flowerdew's (2010: 49) observation that learners might tend to use a favourite SN such as the pair of *advantages and disadvantages*. He posits that such a preference may be attributed to the teaching or text type effect, and this use of SNs represents a kind of formulaic language, which is not found in the native-speaker corpus.

Of the nouns which are not shared between the two corpora, the possible reason for this might be related to the process of argumentation that each group adopted in their writing. To put it more clearly, researchers such Tirkkonen-Condit (1984) describe an argumentative text as an instance of a problem-solving process. Tirkkonen-Condit (1984) explains that an argumentative text is composed of three basic components: a claim, a justification and an induction of the original claim that offers a solution (see Chapter 2, Section 2.8.1 for further discussion on how an argumentative text is structured). To link the argumentative structure to what Table 22 above shows, we notice that both groups share the feature of justifying the argument in question. This element of an argumentative text is met by both groups through a number of SNs such as *reasons, effects and factors*. However, the mechanism of stating the argument/claim is still different between the two groups. For example, the eight nouns which

appear exclusively in the top-15 list in LOCNESS represent the typical structure of argumentative text that is discussed above. This implies that native speakers of English focus mainly on asserting the problem of the argument and express their viewpoint towards it. They then support their arguments with a justification and offering solutions by suggesting *ways*, *measures* and *solutions* to solve these problems. This argumentative pattern could be elicited, for instance, from their high use of the nouns such as *problems* and *solutions*. On the other hand, it seems that Arab speakers of English follow a descriptive style to present their argument by focusing on building up *facts* and giving more emphasis to the *impacts* of the problem. Such a descriptive style appears also through their use of nouns that predict lists of things such as *advantages*, *disadvantages* and *benefits*. Furthermore, from the specific types of SNs that appear in the list, it seems that Arab non-native learners do not offer solutions to the problem at the end of their argumentative essays, or they might introduce this solution implicitly. This represents a deviation from the typical structure of argumentative text that involves a problem-solving framework. Examples (53) and (54) are taken from both LOCNESS and ALEC and illustrate how native writers use SNs explicitly to offer a solution in their argumentative essays compared to Arab speakers of English. Both examples are about ‘dangerous sports’. SNs are marked in bold.

(53) The only suitable **action** to be taken would be to increase safety regulations, ie. head guards and better gloves in all fights, or to shorten the length of fights to cut down some of the constant pounding. (LOCNESS)

(54) **1** On the other hand, extremely dangerous sports that have high probability of severe injuries or death, I think they should be completely banned. // **2** it is the duty of central governments to ensure the health and safety of their citizens and extremely

dangerous sports can ensure neither the safety nor the lives of players and perhaps even spectators. (ALEC)

In Examples (53) and (54), both English native writers and Arab non-native writers suggest that safety regulations need to be increased. However, in Example (53), the native writer expresses this solution explicitly by using the SN *action* and he/she also elaborates on the solution by providing examples about the kind of safety that is required to protect people who are involved in dangerous sports. In contrast, the Arab non-native writer highlights safety as something that needs to be considered, but he/she does not display this clearly. The Arab non-native writer does not use any SN that can signal the solution part in the essay or explain how safety could be ensured. Examples (55) and (56) are a further illustration of how native writers in LOCNESS introduce the solution part and evaluate it in their argumentative essays overtly. Both examples are taken from an essay about ‘traffic’.

(55) **1** There are many possible **solutions** to the **problems** I have identified many of which would compliment each other. // **2** Firstly the **problem** of traffic congestion of our roads could be mainly tackled in two **ways**. // **3** Either by somehow reducing the amount of traffic or by increasing the number of roads. // **4** To reduce traffic tolls could be introduced on motorways and major roads. // **5** This may deter people from using their cars or it may just push them onto country roads increasing rural traffic problems. // **6** People could be encouraged to use public transport by improving the service of buses and trains or by introducing car-free city centres where trains or buses have to be used. (LOCNESS)

(56) **1** Nothing can be done about the first **problem**. // **2** The second is more interesting.// **3** One possible **solution** is a quite radical one. // **4** Say if each household in Britain was only allowed 1 car (or each registered voter was allowed) the volume would be immediately cut exponentially. // **5** This has been tried on a small island

somewhere and had a positive effect. // **6** However this isn't a viable **solution** as it would make any government very unpopular and also reduce input into its tax coffers. // **7** Another **solution** is to change the supply of petrol. // **8** Maybe a rationing **system** could be introduced, or the tax increased greatly. // **9** However none of these **measures** could be introduced unless something is done about the state of Public Transport first. (LOCNESS)

In Examples (55) and (56), native writers use a number of SNs that mark the solution part in their argumentative essays: *solutions*, *ways*, *solution*, *system*, *measures*. They also provide an evaluation of these solutions. For example, in Example (56) in T-unit (6), the native writer evaluates the solution that he/she suggests by writing: *However this isn't a viable solution as it would make any government very unpopular and also reduce input into its tax coffers*. This use of SNs by native speakers is in contrast to that of Arab L2 learners. Example (57) is drawn from ALEC and shows that the Arab L2 writer assesses the practical implementation of the solution before introducing it:

(57) **1** In conclusion, despite some of its advantages mentioned above, dangerous sports should be banned to reduce catastrophe. // **2** I believe that the government should forbid dangerous sports to minimize its risks of serious accidents. // **3** Those who are lucky enough to survive need an intensive medical assistance. // **4** This will not only cost a lot of money for their family but it also put an unnecessary strain on medical system of the country. // **5** I would strongly recommend that dangerous sports should be banned from our societies. // **6** Because they do not contribute to the development of human life and because of the dangers people could face. (ALEC)

In another example from ALEC on the topic of 'dangerous sports', the Arab L2 writer recommends that the government has to interfere to protect people from dangerous sports. But

he/she then ends their essay without suggesting what the government can do to solve this problem.

(58) **1** Finally, some of the young people are eager to try new things regardless of their health conditions, which possibly makes them at risk after these sports. // **2** Therefore, the governments must **interfere** to protect the people from such a dangerous sports that sometimes looking attractive to participate especially from the young generation. (ALEC)

Example (59) reveals a case where the Arab L2 writer introduces a solution implicitly. But the solution is not expressed clearly and it is also not explained adequately.

(59) **1** To sum up immigration has to be under control, // **2** by this I mean **we need to look at the purpose for living abroad**, // **3** on the other hand many of developed countries in the twenty first century specially the UK depend on overseas people, // **4** it is important to consider what will happen if suddenly these people go back their countries. (ALEC)

The final example shows that the Arab L2 writer concludes the argumentative essay without introducing a solution at all leaving the reader to ask so what we can do about this problem.

(60) **1** In short, although immigration continues to be a debatable issue because of some of its disadvantages such as lower wage, education costs, and population surge, no one can deny the fact that it also offers a number of intangible benefits which far outweigh the drawbacks in many ways. (ALEC)

6.8 Signalling nouns as devices for signalling the Problem-Solution pattern in ALEC and LOCNESS

Table 22 showed that the pattern of the Problem-Solution is realised linguistically differently in the argumentative essays in ALEC and LOCNESS. Hoey (2001: 140) highlights that the Problem-Solution pattern arises as a result of the writer answering a predictable series of questions. Such questions would be of the type: ‘What problem arose for you?’, ‘What did you do about this?’ which are the key questions. Hoey (2001: 123) adds that a Solution does not bring the pattern to an ending and a question such as ‘what was the result?’ might be asked to evaluate if what suggested is really a solution or not. As demonstrated in Section 6.7, SNs in Table 22 indicate that native speakers of English seem to be aware of the Problem-Solution pattern and they shift the attention from Problem to Solution by using explicit SNs such as *problem* and *solution*. These lexical signals trigger the pattern by making it visible to the reader and they are one of the main means whereby a reader decodes the discourse correctly. By contrast, in ALEC, Arab speakers of English do not follow an explicit Problem-Solution pattern; they focus mainly on describing the Problem through the use of SNs such as *issue* without offering a solution, thus leaving the reader to ask, so *what did you do about the problem?* The way they write gives the sense of incompleteness about the problem.

The idea that there are lexical signals that can mark the Problem-Solution pattern in text is also supported by Flowerdew (2008), who used a corpus linguistic method to search for key words that provide linguistic evidence for this pattern. In line with the observation that I made about the types of SNs used in the argumentative text, Flowerdew (2008) also observed the lack of lexical signals that refer to solution in students’ texts compared to professional writers. She remarked that students use lexical signals to assess the practical implementation of the

solution before they introduce it, whereas in professional corpus the solution is introduced. It is important to mention here, however, that Flowerdew's (2008) analysis was conducted from the perspective of systemic-functional linguistics. That is, she examined every individual essay in her corpora and then counted the lexical signals that designate the Problem-Solution pattern. In my case, however, I do not conduct this analysis systematically because the Problem-Solution pattern is not within the scope of the present study. I only brought up Flowerdew's (2008) study as a possible explanation for interpreting the difference between Arab L2 writing and English native writing in their use of SNs. This discussion could also be a suggestion for a further work where I could carry out this analysis of SNs systematically.

This difference between Arab L2 writers and English native writers in the way of organising argumentative essays through the Problem-Solution pattern also brings us to think of what Hoey (2001) calls 'culturally popular patterns of text organisation'. Hoey (2001: 122) argues that the text is usually constructed as a pattern, which is organised rather than structured. He explains that although there are preferred sequences and combinations of elements, there is no impossible sequence or combination. In other words, it is not possible to make firm predictions about what is or is not possible in a particular situation. Hoey (2001) maintains that such a feature of text organisation is probably the reason for differences that exist between writers. These patterns are also characterised by the fact that they are culture bound. That is, they do not have a universal status but occur within particular cultures (Hoey 2001: 122). The crucial point that Hoey's (2001) culturally popular patterns of text organisation emphasises is the role of the reader in textual interaction. Nystrand (1986: 48) assumes that the writer writes a text taking into account the reader's possible reactions and the reader reads the text considering the writer's intention, and names this assumption the 'Reciprocity

Principle'. As Table 22 revealed, writers of the two cultures (English vs. Arabic) manipulated the Problem-Solution pattern differently; the writer's attitude towards the reader in the English culture seems to be different from that in the Arabic culture. The difference lies in how writers construct relationships with readers. English writers give more attention to the reader by their attempt to cover all potential questions that might be asked by the reader. On the other hand, Arab writers of English focus more on the writing process itself than considering the reader. This might reflect a difference in the textual patterning of each group of writers due to their different cultures. Although the analysis of SNs based on the Problem-Solution pattern and the culturally popular items of text organisation belongs to coherence more than cohesion, SNs, as cohesive devices, proved to be an effective tool to unearth the important pattern of problem-solving that could be used to differentiate NS and NNS writers linguistically and culturally.

6.9 General discussion on signalling nouns

Arab L2 learners adopted different strategies of using SNs in their English writing. They either repeated (including near repetition) the same form of an SN with the same referent within a short textual distance; or they used pairs or a cluster of SNs whose meanings are semantically similar, or reused the same SN throughout the whole essay. These tendencies of repetition by Arab speakers of English could be due to the small range of SNs types that these learners have. This might also be explained by Hasselgård's (1994) 'teddy bear' effect with learners relying on those nouns and patterns they are most familiar with. Such a repetition of SNs could also demonstrate an inability of Arab L2 learners to precisely evaluate a specific stretch of discourse or select the proper label to name it. In this context, Flowerdew (2003) points out that SNs are likely to be problematic for non-native speakers of English because of

their cognitive complexity. Flowerdew (2003: 330) summarises the reasons for cognitive complexity as follows:

- a. Signalling nouns refer to abstract entities and are thus removed from the concrete world of reality.
- b. The realisation of signalling nouns must be sought out both within and outside the clause in which they occur, as well as through mutual background knowledge.
- c. Signalling nouns introduce additional propositional density to a text.

The reasons presented by Flowerdew (2003) might have led Arab speakers of English to repeat a number of SNs in their English writing in order to avoid any intellectual effort, which involves producing fresh SNs that evaluate the discourse in a precise way. However, SNs, as Petch-Tyson (2000) points out, need to be an effective device for encouraging the reader to understand the writer's opinion because they are able to add an evaluative flavour to the text. Petch-Tyson (2000) adds that texts which make use of more specific SNs will create a more persuasive effect than those which use little non-specific SNs. This indicates that SNs should be used to condense the content of previous propositions into a nominal phrase which can then be included in new propositions. Thus creating a rhetorical effect of 'brick building', which is particularly important in argumentative writing in which the argument is developed in successive stages, with one idea built upon another (Petch-Tyson 2000). Mauranen (1993: 66) also notes this function of SNs, stating that this type of text references helps to build the argument as a hierarchy with layered elements.

In addition, the tendencies of repetition of SNs by Arab L2 writers can be considered as a stage in their interlanguage, where non-native learners at least have a partial understanding of SNs and how they function. On the positive side, it is from such elementary beginnings that the capacity for establishing extensive cohesive relations will develop, as Murphy (2001) describes. This also has the implication that the use of cohesion might be developmental in nature, as Connor (1984: 310) argues. These patterns of interlanguage that appear in the production of the Arab non-native group could again be attributed to transfer from L1. However, this is only one factor because interlanguage is the complex result of internal cognitive processes that take place during learners' process of second language (SL) acquisition.

6.10 Conclusion

This chapter confirmed that depending on surface cues and frequency counts rather than on deep linguistic analysis is not sufficient to find differences in lexical cohesion between native and non-native writers. This indicates the importance of combining frequency analysis in any study of lexical cohesion with a more fine-grained qualitative analysis such as text analysis. In this chapter, the text-linguistic approach to lexical cohesion helped to uncover differences in the way simple repetition, derived repetition and signalling nouns function in their context in ALEC and LOCNESS. The text analysis showed that these forms of lexical cohesion do not only function as surface devices but can also reveal the coherence of the text and how is organised. This finding answered the second research question that has been introduced in Section 6.1. Overall, the text analysis demonstrated that Arab speakers of English, as compared to native speakers of English, were unable to use lexical cohesive devices to form meaningful connections between T-units over large stretches of text. Rather, they used these

devices to build up facts and emphasise particular ideas within the text several times without being able to define them and make them concrete by adding new information. This tendency of juxtaposing lexical cohesive devices in English writing rather than using them to develop the text argument could be attributed to the influence of Arabic language. The way how Arab speakers of English used lexical cohesion to build up their text argument through patterns of repetition is acceptable in Arabic discourse because Arabic argumentative writing uses repeated juxtaposition as rhetorical devices to achieve persuasion. However, this kind of repetition is not valued in English persuasive discourse. English native writing is expected to be coherent and concise and contains a logical progression of text arguments. Consequently, the negative effect of the L1 transfer of these textual habits into English writing results in interlanguage cohesion. The various ways by which Arab speakers of English employed simple repetition, derived repetition and signalling nouns in their writing might indicate that Arab L2 learners are still in the process of grasping how these devices could be appropriately used.

The qualitative analysis in this chapter that has been based on the traditional text-linguistic analysis is conducive to the description of the function of lexical cohesion. Nevertheless, it only focuses on the analysis of one level of language. That is the paradigmatic level. But corpus linguistics and second language acquisition research have found that language doesn't work in this slot-and-filler fashion and is not stored in the mental lexicon as a giant substitution table. Linear relationships with other words are equally important. Accordingly, syntagmatic relationships could be another distinguishing factor between native and non-native speakers of English in their use of lexical cohesion. The forthcoming chapter will, therefore, analyse lexical cohesion at the syntagmatic level applying corpus-linguistic

concepts. The next chapter will also address the final research question that examines the role of a corpus linguistic approach in describing the function of lexical cohesion.

Chapter 7

A corpus linguistic perspective on lexical cohesion: A qualitative analysis and discussion

7.1 Introduction

In the previous chapter, a text analysis was conducted to examine the paradigmatic relationships of lexical cohesive items in both ALEC and LOCNESS. That is, how lexical items were selected, organised and related to each other in terms of such relations as simple repetition, derived repetition and signalling nouns. The chapter further discussed the frequency data of these lexical cohesive devices and related differences between Arab L2 learners and native speakers of English to a phenomenon such as L1 influence and other linguistic factors. Nevertheless, text analysis, as Sinclair (2004) describes, focuses on ‘point-to-point’ cohesion because each element that enters into a cohesive relationship is less than one sentence long; normally a word or phrase, or at most clause. This kind of pattern is the basis of most accounts of cohesion. Sinclair (2004: 140) further contends that “[t]he tradition of linguistic theory has been massively biased in favour of the paradigmatic rather than the syntagmatic dimension”. The syntagmatic dimension is typically described in terms of lexical co-occurrence (collocation), colligation, semantic preference and semantic prosody. In this regard, Stubbs (2001b: 119) underlines that “syntagmatic lexical patterns both provide a perspective on text cohesion, and also have implications for a theory of communicative competence”. Stubbs (2001b) further explains that lexical cohesion is achieved through the recurrence not just of individual words or their derivatives but of ‘lexico-semantic units’, including collocations and other formulaic lexical combinations, thereby creating “a relatively unexplored mechanism of text cohesion” (Stubbs 2001b: 120). Hunston and Francis (2000)

support this view and suggest that lexical cohesion is much more than a reflex of logical or content relations, and is partly due to the stringing together and overlapping of formulaic lexical combinations. Bublitz (2011: 48), in his support of the role of collocation, also points out that the semantic orientation of an individual collocation along with the negative or positive prosody that it creates is not restricted to a single occurrence but can indicate the overall tenor of the discourse. Furthermore, Baker (2011: 231) underlines that “what actually gives texture to a stretch of language is not the presence of cohesive markers but our ability to recognise underlying semantic relations which establish continuity of sense”.

Such claims by these researchers imply that lexical cohesion could be analysed at a phraseological level. Therefore, in order to look at lexical cohesion from a syntagmatic level in this chapter, I will apply an important concept in corpus linguistics that is suggested by Sinclair (e.g. 1996, 1998) which is the ‘lexical item’. The lexical item characterises extended units of meaning and therefore can be used as a way of analysing text. The lexical item comprises five categories of co-selection: the ‘core’ and the ‘semantic prosody’ are obligatory categories and ‘collocation’, ‘collocation’ and ‘semantic preference’ are optional. These five categories have already been explained in Chapter 3, Section 3.2. The current study focuses on ‘semantic preference’ and ‘semantic prosody’. As discussed in Chapter 3, using an example from Sinclair (2004), Mahlberg (2006) demonstrates how the semantic preference and prosody of the lexical item *true feelings* contribute to cohesion in an individual text. Mahlberg (2009: 114) finds that semantic prosody can uncover how lexical cohesion functions in text. It also plays a role in creating the cohesive harmony of the text. It is crucial to point out that this study draws on Sinclair’s (1996) features of ‘semantic prosody’. That is, semantic prosody is not described simply as ‘good’ or ‘bad’ as Partington (2004a) identifies.

Instead, semantic prosody, expresses the attitudinal discourse function of a unit of meaning that is not always explicitly characterised. Hunston (2007) adds that such meaning is often not reducible to a simple ‘positive or negative’. It is essentially linked to a point of view, so that there is often not one indisputable interpretation of attitude. Another feature of semantic prosody is that it is not a property of a word but it stretches across a span of words, and therefore contributes to the cohesion of a text (Stubbs 2001). In my analysis, communicative purposes such as *uncertainty* were considered. Nevertheless, the good and bad dichotomy was not completely neglected from my analysis. Morley and Partington (2009: 141) explain that the fact that the goodness or badness may come in different forms can be seen in relation to different communicative or pragmatic purposes, such as *danger*, *difficulty*, *uncertainty*, all of which are subcategories of bad, or negative, prosody. Therefore, the analysis in this chapter focused mainly on determining discourse functions of the lexical items under analysis, and at the same time referred to the overall environment around the lexical item whether it is positive or negative, because this helped to observe the differences between the two groups of native and non-native writers more clearly.

This chapter will firstly present the lexical items whose textual patterns were investigated (Section 7.2). The chapter then explains the approach that I used to compare semantic preferences and semantic prosodies between a reference corpus and individual essays from LOCNESS and ALEC (Section 7.3). Next, Section 7.4 introduces the first lexical item for the analysis, which is the verb lemma REDUCE. This section shows concordance evidence from a reference corpus that helped to describe the semantic preference and prosody that are associated with REDUCE. This section is further divided into sub-sections (7.4.1 and 7.4.2) in which the same verb lemma was re-investigated but in individual essays from LOCNESS and

LOCNESS. The subsequent Sections 7.5 and 7.6 present the analysis of the other two verb lemmas (AFFECT and FACE) respectively, where the same steps of comparison and the same format of presenting the results that are explained in Section 7.4 were followed. Then, Section 7.7 compares lexical cohesive networks of the three lexical verb lemmas jointly in an English native essay and an Arab non-native essay. It principally focuses on how the three lexical items interlock in the same essay to establish lexical cohesion. Finally, the chapter ends with general conclusions stressing the importance of semantic prosody in establishing lexical cohesion (7.8).

7.2 Lexical items selected for semantic preference and semantic prosody analyses in this study

As a starting point for analysing lexical cohesion using the ‘lexical item’, I selected three lexical items from a number of individual essays from ALEC and LOCNESS. These items are: REDUCE, AFFECT and FACE. Capital letters are used to refer to a lemma, which stands for different word forms of the verb. This means, AFFECT stands, for example, for *affect*, *affects*, *affected*, *affecting*. The three selected items are lexical verbs because verbs are crucial to building an argument, and Hyland and Milton (1997) further observe that most non-native writers have difficulty in using verbs in argumentative essays. These researchers attribute this difficulty to the fact that verbs involve communicating the writers’ stance towards statements and audience. A more valid reason for selecting these lexical items is that they have been studied in English and I wanted to move from the established patterns of English to investigate the patterns occurring in the English native writing and Arab L2 writing by examining the collocational behaviour and semantic prosody of these words (see Xiao and McEnery 2006). For example, the textual behaviour of REDUCE is similar to the verb

ALLEVIATE, which has been studied by a number of researchers such as Stewart (2010), Whittsit (2005) and Stubbs (2015). Thus, the analysis of ALLEVIATE provided me with an initial conceptual insight into the analysis of REDUCE. For the other two verb lemmas (AFFECT and FACE), they have already been assessed by Stubbs (1995) and Tognini-Bonelli (2001) respectively. Therefore, the analysis of these researchers was the platform on which I built up my analysis.

7.3 Matching individual essays against a reference corpus: A corpus-based approach to textual cohesion

Each verb lemma, REDUCE, AFFECT and FACE, has to occur in the English native essays and the Arab non-native essays that are selected for the comparison. In order to investigate whether the lexical item under analysis shows similar textual behaviour in both native and non-native essays, I studied its semantic profile in a reference corpus. Matching a reference corpus against an individual essay to study the syntagmatic dimension of lexical cohesion is central to this analysis. This is because studying semantic preference and prosody in short texts is theoretically not possible. Hunston (2002: 142), for example, argues that “semantic prosody can be observed only by looking at a large number of instances of a word or phrase, because it relies on the typical use of a word or phrase”. Stubbs (2001a: 123) also claims that the only way to interpret patterns that exist in an individual text is by comparing them in the language as a whole in order to know norms of language use (see Chapter 3, Section 3.2 for how Stubbs (2001a) analyses lexical cohesion matching an individual text against a general corpus). He maintains that relying on one corpus only cannot represent a perfect sample of language in use. The reference corpus, therefore, helped reveal the textual patterns of the

lexical items under analysis that they were reassessed in individual essays from LOCNESS and ALEC.

Accordingly, I initially documented the semantic preference and semantic prosody for every target item in a larger corpus that serves as reference corpus to see the typical use of the lexical item in the language. Once the semantic profile of the lexical item in question was created in the reference corpus, I re-examined its semantic preference and semantic prosody in the individual essays that were selected from LOCNESS and ALEC. In selecting a reference corpus, I took into account, whenever possible, the topic or the domain of the individual essays that contain the target items. This is to follow Hoey's (2005) claim that lexical cohesion is domain specific. For example, the first two lexical items REDUCE and AFFECT both appear in essays about 'traffic' in both corpora. Consequently, the reference corpus also contained essays about 'traffic'.

For the lexical verb FACE, however, I did not draw on the reference corpus of essays about traffic that was used to study the semantic profile of REDUCE and AFFECT. This is because most instances of FACE in the traffic reference corpus indicate the concrete meaning that is related to position and direction (e.g., a traffic jam may result if four vehicles *face* each other side-on). However, this meaning was excluded from my analysis. As a result, I collected a reference corpus that matches other topics in LOCNESS and ALEC in which FACE has an abstract meaning (see Section 7.6 for more detail on FACE and its meanings). FACE is distributed across various topics (other than traffic) in the English native essays and the Arab non-native essays. These topics include: 'immigration', 'computer vs. human brain' and 'dangerous sports'. Nevertheless, FACE is not shared between native and non-native essays of the same domain (with the exception of the traffic topic). As discussed in Chapter 4, topic,

on the one hand, is a key factor that determines the selection of words (Biber 2006), and the type of lexical cohesion constituted as such (Hoey 2005). On the other hand, this variety of topics in both corpora discusses general matters of everyday life. This indicates that a lexical item such as FACE can be discussed in everyday life issues across different topics without having a special meaning. As such, I compared the semantic behaviour of FACE in different topics from both LOCNESS and ALEC providing that this lemma was not found in the same topic in the native and non-native argumentative essays. For the reference corpus, I collected a corpus of essays about ‘immigration’ (see Section 7.6). The reference corpus of essays about ‘traffic’ and ‘immigration’ was collected through Sketch Engine by the WebBootCat tool. This tool enabled me to gather as many texts as possible about the same topic through the seed words option. I then uploaded these texts to the WordSmith software for ease of analysis. The previous steps of comparison can be summarised as follows:

- 1- Study the semantic preference and semantic prosody of the lexical item under analysis in a reference corpus.
- 2- Examine the semantic preference and semantic prosody of the same lexical item in native text/s, and check whether they are in line with the semantic preference and semantic prosody in the reference corpus.
- 3- Examine the semantic preference and semantic prosody of the same lexical item in non-native text/s, and check whether they are in line with the semantic preference and semantic prosody in the reference corpus.
- 4- Compare the semantic preference and semantic prosody of the same lexical item under analysis in the native text and the non-native text.

The principal aim of this comparison was to examine whether there are any differences between English speakers of English and Arab speakers of English when they use a certain lexical item syntagmatically (i.e. how a specific collocation is chosen). As I have pointed out in the introduction, lexical cohesion works on the syntagmatic axis. Therefore, any divergence from the normal use of a certain collocation and the semantic prosody will reflect, as a result, a difference between the texts involved in the analysis at the level of cohesive networks that are established.

7.4 The textual behaviour of REDUCE in a reference corpus

The first lexical item selected for the analysis is REDUCE. This lemma occurs in LOCNESS and ALEC particularly in essays about traffic. As explained in Section 7.3, I created a traffic-specific corpus through Sketch Engine. The corpus size of the traffic topic is 1,470,670 tokens. There are 107 tokens of the lexical item *reduce*. Other word forms of *reduce* with their raw frequencies are: *reduced* (35), *reduces* (25), *reducing* (39). I then documented the semantic preference and semantic prosody of REDUCE in the traffic corpus in order to observe, as frequently as possible, the lexical patterns that co-occur with REDUCE. In Figure 14, I documented 56 instances of the total 206 to show the patterning of REDUCE.

1 to combat smog , these measures also reduce congestion. A weakness of this met
2 cates greater use of road pricing to reduce congestion (a demand-side solution
3 ft of just an hour can significantly reduce the amount of time spent in the ve
4 h been proposed as measures that may reduce congestion through economic incent
5 nvestments throughout the state that reduce congestion and other impediments t
6 ri-Met has implemented procedures to reduce idling and improve vehicle mainter
7 invest \$42 million in 40 projects to reduce costs and improve PV performance,
8 contribution to national efforts to reduce greenhouse gas emissions and promc
9 an undertake concentrated efforts to reduce costs in order to stretch limited
10 time with a lower social cost would reduce total trip costs, the optimum is a
11 ole fashioned walking in cities can reduce pollution by 40 percent by 2050, a
12 presents a number of technologies to reduce congestion by monitoring traffic f
13 r aggravate congestion; most of them reduce the capacity of a road at a given
14 e of the most significant actions to reduce household carbon emissions 5000 40
15 choose to ride public transportation reduce their carbon footprint and conserv
16 s will be ranked by their ability to reduce delay. Performance of the system a
17 patterns, public transportation can reduce harmful CO2 emissions by 37 millic
18 s trend. Experts indicate we need to reduce total CO2 emissions to 60%-80% of
19 e to congestion. Car clubs thus only reduce traffic congestion rather than pre
20 such as motorcycles. Possible way to reduce traffic congestion:It is impossibl
21 pital prompting them to be ready. To reduce traffic congestion, smart road sho
22 tate DOT shall adopt an objective to reduce traffic congestion in the major me
23 tability of travel times can rapidly reduce the cost associated with excessive
24 g between planning and availability. Reduce the demand for road-space Demand f
25 performance-based investment plan to reduce congestion through-out the state.
26 r commute from a private vehicle can reduce CO2 emissions by 20 pounds per day
27 ation Requires Investment to Further Reduce CO2 Emissions and Conserve Energy
28 diate option individuals can take to reduce their energy consumption and greer
29 lic transportation n is estimated to reduce CO2 emissions by 37 million metric
30 ternative for individuals seeking to reduce their energy use and carbon footpr
31 e: The state DOT will be required to reduce the fatality rate, as measured by
32 o improve traffic-law compliance and reduce road fatalities. South African Mar
33 fic police. This policy is aimed at reducing the number of vehicles on the roa
34 raffic calming : Schemes that aim to reduce the speed of road traffic. For exa
35 ublic health problems. By helping to reduce the number of cars on the road, pu
36 easures that are feasible to help us reduce global warming pollution," Replogl
37 al streets. Physical capacity can be reduced by the addition of "intentional"
38 revent rat running Congestion can be reduced by either increasing road capacit
39 d-space Demand for road-space can be reduced in two ways; by decreasing car ov
40 cing ways. Each of these congestion reducing strategies has a role in major ci
41 le way to commute that saves energy, reduces traffic congestion and helps the
42 closing time may play a key role in reducing traffic congestion. Shifting time
43 transportation as being a means of reducing traffic congestion, providing an
44 ycle, etc.) is another great way of reducing congestion. Active transportatio
45 e effective incident management that reduces annual incident conges-tion delay
46 viable and sustainable approach to reducing traffic congestion. Although driv
47 cient choices. Public transportation reduces overall greenhouse gas emissions
48 used by traffic congestion which can reduce the productivity of warehousing ar
49 sed and the Report discusses ways to reduce them. But the day-to-day variatio
50 licies Some cities adopt policies to reduce rush-hour traffic and pollution ar
51 h been proposed as measures that may reduce congestion through economic incent
52 being viewed as a solution that can reduce public opposition to HOV lanes. St
53 e: The state DOT will be required to reduce the injury rate, as measured by ir
54 ously my actions did more than just reduce the size of the jam. In order to c
55 er of vehicles on the roads and thus reduce rush-hour traffic intensity. Meter
56 educing road speeds in cities, could reduce the frequency and severity of roac

Figure 14 Sample of concordance lines of REDUCE in the 'traffic' reference corpus

In Figure 14, the lexical item *reduce* itself seems to imply a positive meaning. This may be the effect of its basic meaning, which denotes an improvement of something. The lexical verb *reduce* is similar in its textual behaviour to the verb *alleviate*. Stewart (2010: 73) points out that *alleviate* would be considered to have a favourable basic meaning, but it habitually co-occurs with words, which indicate conventionally undesirable things or states of affairs (see also Whitsitt (2005: 297) and Stubbs (2015) for further detail on the analysis of *alleviate*). However, in order to describe the semantic preference and semantic prosody of a lexical item, we do not need to look at it alone. Instead, we need to observe what surrounds this lexical item in the co-text and in the context as a whole. This observation has been supported by Partington (1998: 68) who describes semantic prosody as “the spreading of connotational coloring beyond single word boundaries”. Accordingly, looking at the right co-text of REDUCE in Figure 14, I found a noticeable presence of items that indicate undesirable things/or large quantities of bad elements. Among the collocates that occur repeatedly with REDUCE, for instance, in the traffic reference corpus are:

congestion, the number of, emissions, rush hour, road, Co2, pollution, intensity, roads, vehicles, costs, carbon, cost, fatalities, delay, demand, jam, injury, idling.

It is clearly observable that the vast majority of collocations are abstract nouns and refer to undesirable elements. There are very few cases of REDUCE that collocate with positive elements such as *capacity* in (line 13) and *productivity* in (line 48). It is not straightforward to find a single label for the various collocates of REDUCE in order to identify its semantic preference. That is because REDUCE collocates to the right with an inconsistent group of collocates. For example, it co-occurs with items that share the semantic meaning of ‘harmful substances’ (e.g., *emissions, carbon, Co2, pollution, gas*). Consider, for example, lines (8, 11,

14, 15, 17, 18, 26, 27, 29, 36 and 47). Other collocates of REDUCE on the right side include: *congestion* (lines 1, 2, 4, 5, 12, 19, 20, 21, 22, 25, etc), *the amount of* (line 3), *the number of* (line 35), *rate* (line 53), *the size of* (line 54), *intensity* (line 55), *frequency and severity* (line 56). All these collocates share a semantic preference of ‘something that is large in amount, number, size, degree or extent’. The third group of collocates has the semantic preference of ‘expenditure’ (*time, cost/s, energy, consumption*). The final group of collocates characterise the meaning of ‘fatality’ (*fatality, fatalities, injury*).

If we want to simplify this analysis of REDUCE, we could suggest general labels of semantic preference such as ‘environmental, economical, financial and physical problems’. These labels encompass most of the semantic features identified earlier. This simplification is useful here because my aim is not to provide a depth analysis of semantic preference and semantic prosody. Rather, I principally needed to know the communicative function as well as the general aura that REDUCE creates in the reference corpus in order to be able to examine the same lemma in ALEC and LOCNESS to search for any possible lexical cohesion. As a further evidence to prove that REDUCE has a mixture of semantic preferences, I looked at the left side of REDUCE. It collocates with items and phrases that express ‘solutions or a plan of action’ (e.g., *measures, procedures, policies, actions, strategy, idea, solution, schemes* and *option*). There are also a number of adjectives that describe the importance of these solutions (e.g., *significant, effective, immediate*), and other lexical items and phrases that indicate that these solutions are urgent and have to be provided immediately (e.g., *will be required to, efforts, mandated to, prompting*). Besides, these solutions may be only proposals that might be not successful. This meaning is implied in the following items: *helping* (line 35), *help* (line 36), *can* (lines 3, 11, 17, 23, 26, 28, 37, 38, 39, 48 and 52), *seeking* (line 30), *could* (line

56), *estimated* (line 29), *aimed at* (line 33), *aim* (line 34), *may* (lines 4, 42 and 51), *only* (line 19), *possible* (line 20), *planning* (line 24), *plan* (line 25), *would* (line 10), *efforts* (lines 8 and 9), *proposed* (line 4).

This analysis of REDUCE in the reference corpus showed that it has diverse semantic preferences, on the left and right. The following examples each contain more than one of the features discussed above. Any lexical item that participates to the meaning of semantic preference and prosody is in bold, while the core item is underlined. This procedure of illustration is applied to all examples in this chapter.

- (1) [...] these **measures** also reduce **congestion**.
- (2) [...] implemented **procedures** to reduce **idling**
- (3) [...] number of **technologies** to reduce **congestion**
- (4) [...] **significant actions** to reduce household **carbon emissions**
- (5) [...] **seeking to** reduce their **energy use** and **carbon emissions**
- (6) [...] will be **required** to reduce the **fatality rate**
- (7) [...] **could** reduce the **frequency and severity**
- (8) [...] national **efforts** to reduce greenhouse **gas emissions**
- (9) An **effective strategy** to reduce greenhouse **gas emissions** [...]
- (10) [...] adopt **policies** to reduce **rush-hour traffic**
- (11) [...] **methodologies** could reduce **congestion** in real traffic
- (12) By **helping to** reduce **the number of** cars on the road [...]
- (13) **Schemes** that **aim** to reduce the speed of road **traffic** [...]
- (14) [...] **plan** to reduce **congestion** throughout the [...]

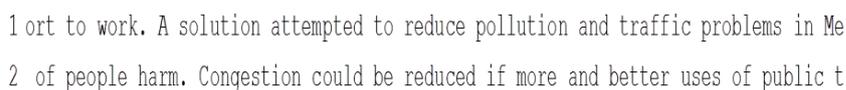
The various preferences, on the left and right of REDUCE, determine the semantic prosody, which can be described in this case as: ‘an effort or an attempt to find an urgent solution to help alleviate unsatisfactory situation/ or decrease something unpleasant or harmful that is usually large in amount, number, size, degree or extent’. This characterisation of REDUCE represents its communicative function in the reference corpus. As it is seen here, both semantic preference and semantic prosody are almost fused. They are both closely related and carry a similar meaning. Nevertheless, I do not plan at this point to distinguish between them because this will distract me from my main goal, which is examining lexical cohesion through these two concepts. Therefore, it is suffice to stress that the semantic preference and semantic prosody of REDUCE in the traffic corpus work together to create a general aura of ‘uncertainty or scepticism’ giving that problems are not fully solved because the solutions that are suggested are usually not feasible.

The analysis of the semantic profile of REDUCE revealed that this verb lemma prefers to occur in a textual sequence between references to a problem and its partial solution. This supports what Stubbs (2015: 112) observes in his analysis of *alleviate* that a single lemma can contribute to text management. Likewise, Hoey (2005) observes that lexical items may be primed to occur in specific types of semantic relation, or in a specific textual pattern. This has an implication for the analysis of lexical cohesion because when this lemma is repeated in the text, its textual pattern with its individual elements will also be repeated, which will then reinforce the strength of lexical cohesive networks. What is central at this time is to check whether some of the collocates that have created the semantic profile of REDUCE in the reference corpus also take place in the native and Arab non-native texts in LOCNESS and

ALEC. Based on this examination, I was able to see whether or not REDUCE displays the same textual pattern in both varieties of writing.

7.4.1 The textual behaviour of REDUCE in LOCNESS

In this section, I examined the lemma REDUCE in an individual native essay that is about traffic. The essay was extracted from LOCNESS. I mainly checked what kind of semantic preference and semantic prosody this lemma creates in the native essay compared to the reference corpus. To begin with, the word-form *reduce* occurs one time in the native text. The text also contains another inflected form of *reduce*, which is *reduced*. Thus, at the surface level, the lexical item *reduce* already establishes a lexical cohesive network of simple repetition (*reduce* – *reduced*). At the syntagmatic level, I looked at the concordance lines of *reduce* and *reduced* as derived from the WordSmith. This gave an idea about the collocates around them.



1 ort to work. A solution attempted to reduce pollution and traffic problems in Me
2 of people harm. Congestion could be reduced if more and better uses of public t

Figure 15 Concordance lines of REDUCE in a native text

As the concordance lines in Figure 15 demonstrate, *reduce* and *reduced* co-occur with unpleasant collocates (*pollution, congestion*). The native writer also associates the two forms of REDUCE with other lexical items that appear in the immediate context around the core item (e.g., *problems, harm*). The native text, in line with the reference corpus, shows that REDUCE collocates with abstract nouns that refer to something bad. More particularly, these collocates denote an environmental problem in this context.

I also considered the left side of REDUCE in Figure 15, and I found that the native text contains the lexical item *solution*, which belongs to the semantic field of ‘solution’ that has been identified in the reference corpus. This use of explicit abstract nouns yields the problem-solution pattern overtly. This pattern, as explained in Section 7.4, plays a role in reinforcing lexical cohesive networks in the text because once the lemma REDUCE is repeated across the native text, the textual and semantic patterns around it are also repeated (see Examples (15), (16), (17) and (18) for an illustration). Furthermore, the concordance lines in Figure 15 reveal that the native writer intends to express the meaning that this solution is incomplete. This could be realised through the lexical verb *attempted*. I then looked beyond the concordance lines of the native text in question in order to find more evidence of the cohesive use of REDUCE. Consider the following text fragment.

- (15) **1 Solutions** to any road **problems**, are always going to **cause** groups of people **harm**. // **2 Congestion could** be reduced **if** more and better uses of public transport were used. // **3** making people use buses and “underground systems” **seems** logical when they are operating anyway. // **4** Persuading people to leave their cars at home must become a major issue. // **5** Increasingly people cycle to work and walk, but because of the unreliability of U.K. weather, a rainy day means back into the car and comfort to work. // **6** A **solution attempted** to reduce **pollution** and **traffic problems** in Mexico City was to **only** allow certain coloured cars onto the roads on specific days. // **7** Surely the U.K. **can** come up with something **more** socially acceptable than this.

The semantic preference of REDUCE in Example (15) is apparent. The first mention of the word-form *reduced* in T-unit (2) is preceded by negative collocates such as *problems*, *cause*, *harm* and *congestion*. Also, to the left of the node item, the use of lexical items such as (*could* and the second conditional *if* statement) reinforces the meaning of semantic preference where

a solution is suggested but it is not effectual. Additionally, the use of the hedging verb *seems* in T-unit (3) adds to this meaning. The other word-form *reduce* that appears in T-unit (6) also attracts negative collocates around it (*pollution* and *traffic problems*). Besides, the native writer uses the item *solution*, and shows that this solution is merely a trial. This meaning is fulfilled by the lexical items (*attempted, only, can, more*). There are similar lexical sets that spread over the text and share the same semantic meaning (*help, suggested, cannot be, possibly, not enough*). All these semantic features help express the attitude of the native writer of the uncertainty and scepticism about finding a proper solution to a problem. This latter meaning manifests the semantic prosody of REDUCE. The following are more examples that illustrate how native writers use REDUCE in their argumentative writing.

(16) [...] **1** One **possible solution** is a quite radical one. // **2** Say if each household in Britain was only allowed 1 car (or each registered voter was allowed). The volume would be **immediately** cut exponentially. // **3** This has been tried on a small island somewhere and had a positive effect. // **4** However this isn't a viable **solution** as it **would** make any government very unpopular and also reduce input into its tax coffers. // **5** Another **solution** is to change the supply of petrol. // **6** **Maybe** a rationing **system could** be introduced, or the tax increased greatly. // **7** However none of these **measures** could be introduced **unless** something is done about the state of Public Transport first.

(17) **1** The simple **way** to solve this **problem** would be to either build more roads or make the present ones larger. // **2** However this, as seen in recent events for example Newbury, has been met with strong opposition from environmental groups and local residents. // **3** another **way** to reduce **congestion**, especially in the large cities, **could** be to persuade people not to use their vehicles at all. // **4** Introducing **schemes** such as the 'park-and-ride' scheme and encouraging people to use bicycles **would** certainly reduce the **number of** cars on the road.

(18) **1** There are many **possible solutions** to the **problems** I have identified // **2** many of which would compliment each other. // **3** Firstly the **problem** of traffic **congestion** of our roads **could** be mainly tackled in two **ways**. // **4** Either by somehow reducing the **amount of** traffic or by increasing the number of roads. // **5** To reduce traffic **tolls** could be introduced on motorways and major roads. // **6** This **may** deter people from using their cars // **7** or it may just push them onto country roads increasing rural traffic **problems**.

Examples (16), (17) and (18) are extracted from LOCNESS. They display most of the semantic features of REDUCE that have been established in the reference corpus in Section 7.4. Besides, in these examples, the semantic preference of ‘solution’ is clearly articulated using lexical items such as: *solution, system, measures, way, ways* and *schemes*. The general attitude expressed in these examples indicates uncertainty and scepticism about the solution that has been suggested. Native writers establish this meaning of semantic prosody through using lexical items such as: *possible, would, could* and *may*. Overall, native writers exhibit a semantic preference and semantic prosody that are present in the reference corpus.

7.4.2 The textual behaviour of REDUCE in ALEC

To start with, I selected an individual essay from the ALEC corpus. This essay is written by an Arab learner writer about ‘traffic’ and contains the lemma REDUCE. In this essay, the word-form *reduce* is repeated 4 times while *reduced* is 2 times. Therefore, this represents in itself a typical type of lexical cohesion through ‘simple repetition’. The non-native essay also shows a type of derived repetition that could be seen in this lexical network (*reduce – reduce – reductions – reduced – reduce – reduce – reduced*). These networks of simple and derived repetition work at the paradigmatic level by showing what kind of lexical items that are selected by the writer. This analysis has already been examined in Chapter 6; therefore, I

needed to check how REDUCE was used by Arab L2 writers at the syntagmatic level. In order to examine whether the semantic preference and semantic prosody of REDUCE are shared between the Arab non-native essay and the ‘traffic’ reference corpus, I ran concordance lines of REDUCE in the non-native essay. This step helped partially investigate the kind of collocates of REDUCE in the non-native writer’s essay, and revealed his/her attitude of the use of this lexical item. Consider concordance lines of the word-forms *reduce/reduced* in Figure 16 below.

```
1 l consumption rates, but also helped reduce the toxic emissions of cars. Althoug  
2 t is designed to achieve which is to reduce traffic problems (i.e. congestions)  
3 ars on the road is less likely to be reduced in significant number. Instead, car  
4 s increase as the fuel subsidies are reduced. Employees can claim higher expense  
5 he poor. The rich are less likely to reduce their consumption of gas even if the  
6 ernments should raise gas prices and reduce their subsidies, but I think any sur
```

Figure 16 Concordance lines of REDUCE in an Arab non-native text

Despite the few number of concordance lines of REDUCE in Figure 16, they are still revealing. For example, we could observe some collocates around REDUCE that carry unpleasant or unfavourable meaning (e.g., *toxic emissions*, *traffic problems*, *consumption*), or collocates that indicate a large number (*significant number*). All these collocates were also present in the reference corpus. As in the concordance sample discussed in Section 7.4, these preferences, in the Arab non-native essay, which occur on the right side, can be blended in one broad label to indicate ‘environmental/financial problems’. However, from Figure 16, it is noticeable that the semantic preference of REDUCE that has been characterised through lexical items such as *solution* on the left side of the node in the reference corpus is not present.

With reference to semantic prosody, Figure 16 shows that the Arab non-native speaker of English, in line with the reference corpus and the native texts, uses REDUCE when he/she talks about problems and the urge to find solutions to them. This urge is met by the use of *should* in line 6. The NNS also uses REDUCE to express uncertainty about the success of the solution that is suggested to lessen the problem. From the concordance lines, we can see that there are lexical items such as *helped* and *less likely* that indicate that the solution to the problem is merely a trial towards an improvement. In order to find more detailed information about the use of REDUCE in the Arab learner essay, I looked beyond the concordance lines of REDUCE for more extended units of meaning. Consider Example (19) which shows how the semantic preference and prosody of REDUCE interact to create the textual function of this lemma in the Arab learner essay.

- (19) **1** For example, in the UK you can see hundreds of cars everyday that use a smaller engine or use diesel instead of petrol. // **2** This positive **impact** did not only affect fuel consumption rates, but also **helped** reduce the **toxic emissions** of cars. // **3** **Although** the **policy** of higher gas prices does really **seem** to achieve the **aim** it is designed to achieve which is to reduce **traffic problems** (i.e. **congestions**), it often leads to changes in the consumption patterns of the society instead. // **4** **The number of** cars on the road is **less likely** to be reduced in **significant number**. // **5** Instead, cars will **only** get smaller with less powerful engines.

The fragment text in Example (19) contains a number of semantic features that were also recognised in the reference corpus. For example, I have identified in the reference corpus that REDUCE links solutions to problems, and produces the problem-solution pattern. In Example (19), the Arab non-native writer uses *reduce* in T-unit (2). On the left side of *reduce* in the same T-unit, the writer attempts to express the solution component in the problem-solution

pattern around REDUCE by selecting the SN (*positive*) *impact*. The writer uses this SN to label T-unit (1), which contains this content: *use a smaller engine or use diesel instead of petrol*. This content indicates that people need to start considering other alternatives to petrol that could be used for their cars. However, the Arab non-native writer selects an inappropriate noun from the lexical set that the semantic preference of REDUCE belongs to. The learner perhaps should have used nouns such as *approach*, *strategy* or *solution* to condense the first T-unit.

However, in Example (19), the performance of the Arab non-native learner varies within the same text. The learner in T-unit (3), for example, uses the lexical item *policy* in the surrounding environment of the word-form *reduce*. *Policy* is from the same lexical set of (*solution*, *strategy*, etc) that exhibits the meaning of semantic preference that has been established in the reference corpus (see Section 7.4). The learner this time successfully uses *policy* to label *higher gas prices* (T-unit 3) as a solution to the traffic problem. Thus, I could tentatively posit that, on the one hand, the Arab L2 learner is aware that an abstract noun has to be used on the left side of REDUCE. However, the learner sometimes lacks the knowledge to select the appropriate signalling noun in the context of REDUCE. Flowerdew (2006: 352) argues that “errors such as these [inappropriate use of SNs] might be considered developmental in terms of second language acquisition”.

With respect to semantic prosody in Example (19), the Arab non-native learner maintains the overall evaluative meaning of the semantic prosody of REDUCE, which expresses an attempt to find a solution to help alleviate unsatisfactory situation. This meaning of semantic prosody is carried out by lexical items such as the verb *helped* (T-unit 2) to the left of *reduce* and other similar collocates that imply that the reduction or improvement of a problem may be only

partial and not completely successful: *although, seem, designed, aim* (T-unit 3); *less likely* (T-unit 4); *only* (T-unit 5).

I then continued the analysis and examined the wider co-text of REDUCE in the text in Example (19); I found other co-textual items that fit in with the semantic preference and semantic prosody of REDUCE, as established in the reference corpus. For example, the Arab non-native text shares a number of collocates with the reference corpus such as *gas, prices, traffic, price, problems, congestions, number, rises, emissions, hundreds, surpluses* and *toxic*. The accumulation of all these semantic features in the co-text and the context of REDUCE in the Arab learner's text in Example (19) displays the sceptical attitude of the writer about problems that are not fully solved. Examples (20) and (21) demonstrate further examples from the ALEC corpus that show the textual function of REDUCE in the argumentative writing of Arab speakers of English.

(20) 1 Even though the lives of the people playing this kind of sports can be saved, they still possibly have to **suffer** from permanent **injuries** which can make them **disabled**.// 2 Hence, **preventing people from this kind of sports can reduce** the chances of permanent **injuries** from sport related **accidents**.

(21) 1 [...] **dangerous sports should be banned to reduce catastrophe**. // 2 I believe that the government should forbid dangerous sports to minimize its risks of serious **accidents**.// 3 [...] I would strongly recommend that dangerous sports should be banned from our societies [...]

It is important to note here that even though Examples (20) and (21) are taken from essays about 'dangerous sports' rather than 'traffic', the use of REDUCE still shows similar textual

behaviour to REDUCE in the traffic topic. The surrounding environment of REDUCE in these examples indicate undesirable situations. This negative aura is expressed by the Arab L2 writers through these items: *injuries*, *accidents* and *catastrophe*. These collocates occur in the right side of REDUCE and they suggest a semantic preference of a ‘physical harm or problem’. This semantic preference characterises part of the semantic preference that has been identified in the traffic corpus, which refers to ‘environmental, economical, financial and physical problems’ (see Section 7.4).

However, on the left side of REDUCE, Examples (20) and (21) illustrate how the semantic preference of ‘solution’ is not fully satisfied. The Arab non-native writers communicate the semantic field of ‘solution’ without using any explicit markers such as *solution* or *measures*. This implicit description of the problem-solution pattern by Arab L2 learners is in contrast with that by native writers who express this pattern explicitly, as demonstrated in Section 7.4.1. This observation has already been made in Chapter 6, where the text analysis of SNs in the Arab L2 writing showed that Arab L2 learners do not provide an overt problem-solution pattern, and they often do not offer a solution to the problem in their essays.

Overall, Arab non-native writers, as compared to native writers, deviate partially from the typical semantic preference of REDUCE due to the inadequate knowledge of using the proper abstract noun on the left side of REDUCE, as shown in Example (19), or the absence of any SN that indicates the solution, as in Examples (20) and (21). On the other hand, these learners, like native writers, maintain the semantic prosody in their essays, which indicates an unfavourable state of affairs.

7.5 The textual behaviour of AFFECT in a reference corpus

The second lemma analysed in this chapter is AFFECT. According to Stubbs (1995: 43), AFFECT has a clearly negative prosody. Stubbs groups AFFECT with the semantically related lemmas CAUSE, CONSEQUENCE, CREATE, EFFECT, HAPPEN, REASON. He points out that things are usually badly or adversely affected. He lists a number of top collocates that co-occur with the word-form *affect* that include:

adversely, badly, directly, disease, seriously, severely, worst, drought, floods, f(o)etus, negatively

Stubbs (1995) also notes that other word-forms of AFFECT such as *affected* and *affecting* carry a sense of negativity particularly when these forms occur with medical collocates. He illustrates such meaning with medical examples: *a stroke affected the brain; haemophilia, one younger brother being affected; malaise adversely affecting his physical health*. Stubbs (1995) observes that AFFECT has also neutral collocates (e.g., *areas, changes, countries, people, factors, lives*). However, he argues that the clearest fact is the absence of positive collocates. He maintains that even with no explicitly negative word in the collocational span of AFFECT, the negative prosody on AFFECT can make it difficult to interpret utterances positively. He, for example, demonstrates that if something affects the accuracy of the solution, or if interest rates affect the cost of land, there is almost unavoidably negativity embedded in such examples. According to Stubbs's (1995) analysis, AFFECT has a negative prosody. Taking Stubbs's (1995) analysis into account, I studied AFFECT in the reference corpus of 'traffic' and identified the type of collocates that surround this lemma. This helped profile its semantic preference and semantic prosody. In the reference corpus of traffic, the word-form *affect* occurs (13 times), whereas the raw frequencies of other word-forms of AFFECT are: *affected*

(9), *affecting* (4), *affects* (11). Figure 17 demonstrates 33 concordance lines of the lemma AFFECT as extracted from the traffic corpus.

1 ad, lateral movement of the vehicles across the road affected the noise measurements. In particular, the two measur
2 obstacles such as buildings, walls, trees, etc. would affect the measurement. The residual noise generated by other
3 eved under controlled conditions. Another factor that affects the accuracy of the prediction could be the measureme
4 hich are quite close to the traffic flow line can be affected easily due to this movement. This can be avoided by t
5 roadways of Sri Lanka were selected. The factors that affect the modelling of traffic noise can be categorized as t
6 asuring site may have different conditions that could affect the noise measurement. The condition of the road, the
7 n a vehicle moving at steady speed. Sound levels are affected by meteorological conditions such as wind, temperatur
8 Varma , I need to always remember how my driving is affecting the other person on the road. Is it impeading him or
9 to do but aggressive driving acts because of how they affect the safety and convenience of others." "Mostly, it's t
10 puter models. Traffic situation The traffic situation affects the parameters of interest considerable. The queue le
11 Capacities $S=S_0 * \beta_1 * g/C$ The saturation flow (S) is affected by the multiplication of geometric and traffic condit
12 turbances to the traffic on the motorway and can thus affect the capacity and level-of service of the motorway. The
13 traffic behaves in a less disciplined manner and that affects the average speed of the vehicles. Eddie was the firs
14 ic congestions not only waste time of people but also affect the financial situation of a country. With the growth
15 d sensors are- a) the operation of the system may be affected due to environment condition like fog; b) installatio
16 crime is perceived as pandemic Personal security affects many peoples' [sic] decisions to use public transport
17 ignals or other events at junctions that periodically affect the smooth flow of traffic. Alternative mathematical t
18 . [4] However, unlike a fluid, traffic flow is often affected by signals or other events at junctions that periodic
19 ive routes are attempted (' rat running '), which may affect neighborhood amenity and real estate prices. Higher ch
20 e United States. Saves energy: Our energy consumption affects everything from the price at the pump to national sec
21 tly related to the transport operation (ii) factors affecting other aspects of the logistics activity. Information
22 e carriers Figure 3: Factors, other than congestion, affecting delivery reliability Adaptation of Logistics Systems
23 elected officials, and the public. Since we are all affected by congestion, it is important that we all work toget
24 ravel time is a more direct measure of how congestion affects users. Travel time is understood by a wide variety of
25 ngestion extends to more time of the day, more roads, affects more of the travel, and creates more extra travel tim
26 t to/from work in the "rush hour." But congestion now affects more trips, more hours of the day and more of the tra
27 e extra travel time and amount of the day and system affected by travel delays is not the same every day. It affect
28 gestion. As an example of how travel time reliability affects highway users, consider the following (Figure ES.3).
29 fected by travel delays is not the same every day. It affects not only commute trips, but any trip during the peak
30 rapidly before the traffic volume would begin to be affected by overcrowding. FC: What part of your research could
31 congestion and other impediments to traffic flow that affect safety and the environment. Such actions will include
32 stuck in traffic jams, I discovered that I cannot to affect them by "peeling out" after I'd made my way through th
33 g from numerous on-ramps. The "rolling barrier" can't affect these extra inputs, and if the major portion of the tr

Figure 17 Sample of concordance lines of AFFECT in the ‘traffic’ reference corpus

From the concordance lines displayed in Figure 17, I was able to built up a picture of the semantic environment in which AFFECT occurs. Such an environment includes a number of negative collocates that are distributed to the left and right of AFFECT. The following are some of these collocates (both single lexical items and phrases).

obstacles, conditions, driving acts, traffic situation, less disciplined manner, events, rat-running, congestion, rush hour, delays, traffic volume, overcrowding, impediments, stuck, peeling out, barrier

It is also noticeable from the concordance lines that even when the lexical item that precedes or follows AFFECT is one that is often used positively or neutrally, it can be shown with more co-text that it is used negatively in the environment of AFFECT. To illustrate this case, these are a number of examples (22, 23 and 24) drawn from lines 12, 16 and 21 in Figure 17. The items that contribute to the negativity of AFFECT are marked in bold while the core word is underlined.

- (22) Ramps are sections of roadway that provide connections from one motorway to another motorway or normal road. Entering and exiting traffic **causes disturbances** to the traffic on the motorway and can thus affect the capacity and level-of-service of the motorway.

In Example (22), *affect* is followed by positive items *capacity* and *service*. However, the surrounding environment of *affect* indicates a negative evaluation. This could be understood from lexical phrases such as *causes disturbances*. Example (23) is another example from line 16:

- (23) In the United States expansion of public transportation systems is often opposed by critics who see them as vehicles for **violent criminals** and **homeless** persons to expand into new areas (to which they would otherwise have to walk). According to the Transportation Research Board, “[v]**iolent crime** is perceived as **pandemic** Personal security affects many peoples’ [sic] decisions to use public transportation.” Despite the occasional highly publicized **incident**, the vast majority of modern public transport systems are well designed and patrolled and generally have low

crime rates. Many systems are monitored by CCTV, mirrors, or patrol. Nevertheless, some systems attract **vagrants** who use the stations or trains as sleeping shelters, though most operators have practices that discourage this.

As it is evident from the excerpt in Example (23), the collocates that take place in the immediate span of AFFECT (i.e. *security, people* and *decisions*) are neutral. Nevertheless, their expanded context reveals that they are used negatively. For example, the word-sequence *personal security* when it is in isolation means a protection of people from physical violence, personal harm or domestic abuse. However, this phrase gives a negative interpretation in the extract above. The text implies that people are having negative attitudes towards their personal security and tending to be sceptical about it. This feeling of insecurity, therefore, affects people's decisions to use public transport. Example (23) contains other collocates that are spread beyond the boundaries of AFFECT. These collocates are clearly negative such as (e.g., *violent criminals, homeless, violent crime, pandemic, incident, vagrants*). All these collocates help construct the negative meaning of AFFECT.

Line 21 in Figure 17 shows a further example where AFFECT (in the form *affecting*) is associated with positive collocates *logistics activity* but conveys a negative meaning. This negativity is realised by the lexical verb *disturb*, as illustrated in Example (24) below:

- (24) Previous research has established that traffic congestion is only one of many factors that **disturb** logistical schedules. These other factors can be broadly divided into two categories: (i) factors directly related to the transport operation (ii) factors affecting other aspects of the logistics activity.

From the concordance lines in Figure 17 along with the textual examples provided, it is clear that AFFECT in the traffic corpus shares the negativity that Stubbs (1995) assigns to AFFECT in his corpus. The only difference is that AFFECT in the traffic corpus does not collocate with medical collocates which are apparent in Stubbs's (1995) analysis of AFFECT. In contrast, AFFECT in the traffic corpus prefers to co-occur with lexical items that indicate a 'barrier' or 'hindrance', which makes it difficult for something to happen or be achieved smoothly or efficiently. This barrier could be a concrete thing such as 'a wall or building that influences the performance of something' (see line 2). Alternatively, this barrier could be something that is abstract such as 'a situation, condition, an event or an act'. For example, in line 8, the barrier is the 'driving act', which hinders the other person from moving easily (see Example (25)).

(25) [...] remember how my **driving** is affecting the other person on the road. Is it **impeding** [*sic*] him or [...]

Therefore, this meaning of a barrier or hindrance can portray the semantic preference of AFFECT in the traffic corpus. This semantic preference could be attained through lexical items such as: *across* (line 1), *obstacles* (line 2), *impeding* [*sic*] (line 8), *disturbances* (line 12), *congestion/s* (lines 14, 23, 24 and 26), *delays* (lines 27, 29), *overcrowding* (line 30), *impediments* (line 31), *stuck* (line 32), *barrier* (line 33). However, this semantic preference is not always straightforward to observe. Some collocates that surround AFFECT, for instance, do not clearly belong to a lexical set such as the one that has just been presented. They instead refer to the meaning of a barrier/hindrance implicitly. A case in point is the phrase *personal security* (line 16), which has already been introduced in Example (23). In that text, people have concerns about their personal security which can be a real barrier to their use of public

transport. Other phrases such as *metrological/environment conditions* can also serve as a barrier that prevents an accurate measurement of sound levels to be taken (cf. lines 7 and 15). The lexical item *conditions* itself, in lines 3 and 6, means ‘something that limits or restricts something else’ (Online Collins English Dictionary). Thus, *conditions* can also be interpreted as an obstacle. However, there are a number of examples that do not fit well with this meaning of semantic preference either overtly or covertly. Lines 14 and 20 are among these examples. Examples (26) and (27) show this case:

(26) [...] traffic congestions not only waste time of people but also affect the financial situation of a country.

(27) Our energy consumption affects everything from the price at the pump to national security.

The meaning of hindrance is not obvious in the two examples above. Even after I looked at the broader co-text of AFFECT, I could not find lexical items that refer to a barrier or hindrance. Rather, the word-form *affect* in Example (26), for instance, denotes that traffic congestions cause an undesirable change. Similarly, *affects* in Example (27) is used to indicate that the overuse of energy leads to bad effects or a negative change. Thus, both examples share the meaning of ‘a change’. This meaning is, however, abstract and based on an evaluation more than on lexical items that are visible to the eye. This, consequently, helped me to define the semantic prosody of AFFECT.

The semantic prosody of AFFECT, in the traffic corpus, can be characterised as ‘causing difficulty/ or an undesirable change that limits the ability of someone or something’, and this limitation is lexicalised through the semantic preference. The sense of change is apparent in

most of the examples and concordance lines introduced in this section. Thus, I could posit that cases such as Examples (26) and (27) have only semantic prosody, while their semantic preference is not clear compared to other instances of AFFECT in the concordance lines.

7.5.1 The textual behaviour of AFFECT in LOCNESS

After I identified the semantic preference and prosody of AFFECT in Section 7.5, I then examined the lemma AFFECT in individual essays from the LOCNESS corpus. Particularly, I checked whether AFFECT shares the same semantic preference and semantic prosody with those that were documented in the traffic corpus. The native essay is taken from the LOCNESS corpus and it is about traffic. In this essay, there is only one occurrence of the word-form *affects*. Example (28) is an extract from this essay that contains the core item in question.

(28) **1** A solution attempted to reduce pollution and traffic problems in Mexico City was to only allow certain coloured cars onto the roads on specific days. // **2** Surely the U.K. can come up with something more socially acceptable than this. // **3** Some say increasing prices on petrol and taxing of the use of roads will help. // **4** This is **disastorous**, // **5** as this only affects the **poor**. // **6** Why should those with more disposable incomes be allowed to pay these taxes and use the roads. Where the poor who have as much right as anybody **struggle** to pay extra road tax. // **7** This system simply will **lead to** the richer earners to use our roads.

In Example (28), the semantic preference of ‘hindrance’ is not clearly present in this little extract. However, the lexical verb *struggle* (T-unit 6) could express this meaning because the extra road tax will be a hindrance to the poor. Furthermore, I checked the complete essay that is presented above and I found other lexical verbs that illustrate the meaning of ‘hindrance’:

blockading, hinder and *stopped*. These three verbs establish, in the first place, cohesive links through paraphrase, and they get a further cohesive dimension through the link with *affects*, which is the core of a lexical item. Example (28-a) is an extended text of the extract in Example (28) in which these verbs take place.

(28-a) [...] **1** The second major problem is that not many new roads can be built. // **2** Ever increasing numbers of protestors and environmentalists are **blockading** the paths of construction of ‘by-passes’, e.g. the Dwyford Downs. // **3** This is not a significant problem, // **4** but the areas where new roads are needed most are in the centre of massive cities where new roads can not be possibly constructed, // **5** there is simply not enough room. // **6** Away from cities road construction faces the problem of geographical sites. // **7** The shape of the land and its landforms extremely **hinder** the paths of new ventures. // **8** Rivers need bridges to cross them costing more money which is not already there. Which means construction is **stopped** until further time when resources and money again become readily available.

With reference to the semantic prosody of AFFECT in Example (28), the surrounding co-text of *affects* in the native text implies the meaning of ‘change’. We could elicit this meaning from T-unit (2) which suggests that the *U.K. can come up with something more socially acceptable than this*. T-unit (2) refers to the need of making a change by trying another solution to reduce the traffic problem instead of the one that is used in Mexico. Also, in T-unit (3) people suggest that increasing prices of petrol and road tax could produce a change in the traffic problem. This change is described by the native writer as unfavourable because it will principally affect the poor most in that they will face difficulty in paying the extra rates for the road tax. This is perceived by the adjective *disastrous* in T-unit (4) and the lexical verb *struggle* in T-unit (6). The writer also uses the phrasal verb *lead to* (T-unit 7), which indicates a transformation or a change from one situation to another, usually an unpleasant one.

Example (29) is another example from LOCNESS, which is taken from an essay about ‘boxing’. This example includes two word-forms of AFFECT: *affect...affected*.

(29) 1 [...] so why should it **banned** it will cause **displeasure** to so many people. // 2 Boxing, because it is a big money business, should not be **banned** // 3 because it would affect not only the fighters themselves but also the managers, the professional referees and judges, security guards, organisers and betting offices. // 4 Gambling is a major part of the boxing game, and huge amounts of money would be **lost** by betting offices if the sport was **banned**. // 5 The people who organise boxing matches would also be affected by **losing** business and maybe **becoming unemployed** if the sport was **banned**. // 6 The players themselves would be left **unemployed** also, because many of them being professional boxers have no other jobs to fall back on. // 7 Only the multimillionaire boxer who win every match, and get lots of money from sponsors etc. could survive **unemployed**. // 8 Many boxers have participated in the sport all their lives and it would be a great **loss** to them if the sport was **banned**.

In Example (29), the native writer expresses the semantic preference of AFFECT, which is ‘hindrance’, through the simple repetition of the lexical item *banned* overall the text. According to the Online Collins English Dictionary, ‘to ban something means to prevent something from happening’. Thus, this meaning denotes a restriction and hindrance (to boxing), which characterises the semantic preference of AFFECT. For the semantic prosody of AFFECT, the ‘undesirable change’ could also be elicited from the extract above. For example, in T-unit (5), the immediate co-text of the word-form *affected* contains these lexical items: *losing*, *becoming*, and *unemployed*. The lexical item *becoming* refers to a ‘change’, whereas *losing* and *unemployed* convey the negativeness of this change.

Example (30) is a further illustration of how the lemma AFFECT is used in LOCNESS. This example, however, does not contain the word-form *affect* but its derived noun *effect*. As mentioned in Section 7.5, Stubbs (1995) groups AFFECT with other semantically related lemmas that has a clear negative prosody. The abstract noun *effect* is among these lemmas. Example (30) is taken from a topic about ‘computer vs. human brain’.

(30) 1 [...] One area which has drawn particular criticism is the computer games industry.
// 2 Since the early 1980’s, when computers such as the Spectrum 48K and the Commodore 64 were introduced into the homes of millions of people, controversy has followed about their effect on children. // 3 There is a widespread belief among parents that computer games **hinder** a child’s ability to learn. // 4 The **development** of more sophisticated and technologically advanced computers, such as the Amiga and PC, has served to **deepen** the "**problem**". // 5 The **growing** realism of the games appeal strongly to children.

In Example (30), the native writer uses the lexical item *effect* in T-unit (2). The writer then describes this effect in T-unit (3) where he/she mentions that computer games hinder children to learn. The use of the lexical item *hinder* reflects the semantic preference of AFFECT, which is ‘hindrance’. The native writer further expresses the semantic prosody of AFFECT by using lexical items such as *development*, which implies a change or improvement. The native writer describes this change as unfavourable because the development in computer games will cause more problems to children by limiting their ability to learn. The lexical item *growing*, in T-unit (5), also refers to this change. Example (31) also contains the lemma EFFECT. The text is also about ‘computer and human brain’, but it is written by a different native speaker of English.

(31) **1** However, although the human brain may not be **redundant** it could **become** programmed like a machine influenced by the computer's programming. // **2** Therefore, computers can have an **adverse effect** on the human brain's greatest facility, its imagination. // **3** Whilst the user **becomes submerged** in the computer and its programming he may well **lose** his force of imagination.

In Example (31), the semantic preference of 'hindrance' is not lexicalised clearly. The lexical item *submerged* might carry the meaning of 'hindrance' implicitly. *Submerged* in T-unit (3) means to make the computer user fully involved in a computer's programming. This submerging can stop or hinder the development of the user's imagination. The semantic prosody of 'change' is reflected by the repetition of the verb *become/s* in T-units (1) and (3). The negativeness of change is described by the use of the adjective *adverse*, which co-occurs with *effect* in T-unit (2). Other lexical items that convey the meaning of the undesirable change are: *redundant* and *lose*. Consequently, the analysis of AFFECT in the native essay shows that the native writers share the semantic preference and semantic prosody of AFFECT that have been documented in the reference corpus. However, not all native speakers of English lexicalise semantic preference and semantic prosody overtly. But the clear observation in these examples by native writers is that these writers maintain the negative aura that AFFECT creates.

7.5.2 The textual behaviour of AFFECT in ALEC

In this section, I test whether AFFECT in the Arab non-native essays agrees with the reference corpus in having the same semantic preference and semantic prosody. I selected for this analysis an Arab L2 essay about 'traffic'. In this essay, the word-form *affect* occurs once. There is also one mention of the noun *effects* that constitutes with *affect* a cohesive link of derived repetition. Below is an extract from this essay, which illustrates how the lexical item

affect behaves semantically (i.e. at the level of semantic preference and semantic prosody), compared to the reference corpus.

(32) **1** The effects on increasing the fuel fees from the economical impacts will **help** on **maintaining** fuel consumption to some limitation // **2** because of the rising there will be more invention to **replace** the use of fuel // **3** and people will start to think for **other** solutions, // **4** economically this is very good to increase the industrial and the manufacturing process across any country.// **5** The rising up will limit the pollution problem and will keep a friendly environment // **6** as people will lead to use less fuel, park their cars and start walking **instead** which then will help to keep them active and get healthier and be fit. // **7** The use of **alternatives** like bicycles and walking will affect their health on **good way** and keep them as **active** and **focused**.

In contrast to the reference corpus, the lexical item *affect* is used positively in Example (32) above. For example, in T-unit (7), *affect* associates with the lexical item *health* and other positive adjectives such as *good*, *active* and *focused*. Besides, the whole paragraph conveys a positive meaning. It explains that there are various benefits that could be gained by rising fuel fees. This positive atmosphere can be understood from lexical items and expressions such as *help*, *maintaining*, *limit the pollution problem* and *keep a friendly environment*. The semantic preference of AFFECT, which has been assigned a label of a ‘hindrance’ in the reference corpus, does not seem to be apparent in this extract even when I looked at the complete essay. This supports Sinclair’s (2004) assumption that semantic preference is optional.

On the other hand, the semantic prosody of ‘change’ is noticeable in the Arab non-native text above, in which the Arab L2 writer describes how increasing fuel fees could lead to a change. The writer employs a number of lexical items that help reveal this meaning of ‘change’ (e.g., *replace*, *other*, *instead* and *alternatives*). However, the writer characterises this change as

positive, which is not in line with the semantic prosody of AFFECT that is marked as negative in the reference corpus and in the native texts. As such, we could see how the Arab non-native writer uses the lemma AFFECT in a positive environment and this creates a shift in its semantic prosody. This may lead to misunderstandings with respect to the tone or content of the original message. I also observed that the word-form *affect* was used positively in another Arab non-native essay from the ALEC corpus. This essay is also about ‘traffic’. Example (33) shows a short paragraph from this essay.

- (33) **1** [...] raising fuel prices let to **positive changes** in individuals’ behaviours in terms of gas consumption. // **2** For example, in the UK you can see hundreds of cars everyday that use a smaller engine or use diesel **instead** of petrol. // **3** This **positive** impact did not only affect fuel consumption rates, but also **helped** reduce the toxic emissions of cars.

As in Example (32), the Arab L2 text in Example (33) does not exhibit a semantic preference of ‘hindrance’. There are no lexical items, either in the small quote above or throughout the complete essay, which indicate this meaning. However, the meaning of ‘change’, which represents the semantic prosody of AFFECT, is present. The items in bold in Example (33) *changes* and *instead* characterise this meaning. The text indicates that increasing fuel fees leads people to find alternative fuels other than gas. This meaning is perceived by the item *instead*. The text also shows that increasing fuel fees will produce a positive change in the consumption rates and will help decrease environmental problems. This change is again favourable and can be understood from the context that surrounds *affect* in which the Arab L2 writer uses items that give positive reading (e.g., *positive*, *helped*).

To conclude, the semantic preference of AFFECT is absent in both Arab non-native essays, while the semantic prosody of this lemma is used positively rather than negatively. As a result, this use of AFFECT in the Arab non-native texts does not comply with the typical semantic preference and prosody of AFFECT in the reference corpus (see Section 7.5). This alteration of the typical textual function of AFFECT produces a positive environment that contains positive lexical items that do not fit in with the node item. To put it the other way round, the prosody of AFFECT clashes with prosody's consistent series of collocates that surround it. Such an atypical use of this lexical item may consequently create an inconsistency in the lexical cohesive networks throughout the text.

7.6 The textual behaviour of FACE in a reference corpus

The third verb lemma which was selected for the analysis is FACE. According to the Online Collins English Dictionary, *face* has two meanings. The first meaning refers to a concrete area of meaning where “something or someone are positioned opposite to another thing or person or are looking at that direction”. The second meaning is more abstract which indicates that “if you face or are faced with something difficult or unpleasant, or if it faces you, it is going to affect you and you have to deal with it”. The analysis in the present study, as I have explained in Section 7.3, considered only the second meaning of *face* and identified its semantic profile in a reference corpus. In Section 7.3, I have already mentioned that FACE was not studied in the reference corpus of traffic. Rather, the reference corpus, which was used to study the semantic profile of FACE is about ‘immigration’.

The lexical verb *face* has been studied by Tognini-Bonelli (2001: 20). She observes that FACE has a semantic prosody of negativeness and “whatever you are likely to *face* is usually

a very undesirable thing or event". She identifies a set of words that co-occur with *face* such as *grim, dilemma, obstacles, challenges, difficulties, problem, task, threat*. As such, I based my analysis on this ground and studied this textual behaviour of FACE in the reference corpus of 'immigration'. The corpus size of the 'immigration' reference corpus is 588,601 tokens. The lemma FACE occurs very frequently in this corpus. The word-form *face*, for example, occurs (143) times, while other word-forms have the following frequencies: *faces* (10), *faced* (28), *facing* (59). Below is a partial concordance for FACE from the immigration corpus

1 threats of political persecution they faced were greater. Zucker (1983) argues a
2 efugees to the place where they would face persecution. Each year, the President
3 unities. Additionally, refugees might face more problems with reintegration beca
4 e refugees, more problems. And, let's face it, there will be no end to the Syria
5 is pensioners. But young workers have faced a 'double whammy': greater falls in
6 nancial crisis. Youngsters have since faced a ''double whammy'' of scarcity of w
7 rates differential selection-refugees face persecution or well-founded fear of p
8 anticipate the challenges they might face along the way. The increase in human
9 leading up to their flight, and often face crowded and unsanitary conditions in
10 ncourage migrants who do not actually face persecution to use the asylum route t
11 s a nightmarish future for us all. We face a bleak transportation future if toda
12 s power of subpoena. While aliens may face suit under tort or commercial laws, t
13 be 200 000. 29 People without status face serious barriers to health care, 34 b
14 r to it . Why should a family have to face war because they're not Christian? it
15 osed, have their licences removed, or face prosecution if they continue to flout
16 aurants and other takeaway firms will face closure if they are found to be emplo
17 and persons in private accommodation face problems sending their children to sc
18 f their claim for asylum because they face persecution in their home country on
19 o a country where that individual may face persecution on the basis of race, rel
20 s in need of international protection face difficulties accessing the procedures
21 ination offices. Domestic non-profits face several issues that challenge their a
22 and fight. Conflicts many communities face in making sure immigrants understand
23 of the key policy challenges that we face as a state grow out of these demograp
24 o cooperate with the INS or otherwise face loss of specified state funds; and bu
25 are not found by 2020, residents will face a water shortfall nearly as great as
26 o live in the UK - especially if they face a jail sentence for getting it wrong.
27 ion concentrations in low-lying areas face greater flood risk as sea levels rise
28 tland and the UK. Scottish debt would face a larger risk premium because of bein
29 o promote international stability-but face constraints which rein in the actual
30 onal firm and lives in a big city may face considerably less problems of adjust
31 ing illegal immigrants to 'go home or face arrest'. The pilot project, which res
32 l asylum to some LGBT individuals who face potential criminal penalties. [33] [3
33 des protection to certain persons who face persecution, harm or torture upon ret
34 to their claim takes place. They also face strict limitations on work eligibilit
35 petitiveness? would be harmed if they face tough restrictions on who they are al
36 erations. Refugees and asylum-seekers face a variety of protection challenges: E
37 and the victors of the election will face a series of major challenges as the e
38 ome people forced into sexual slavery face challenges of charges of illegal immi
39 raised here reflects the same problem faced by the U.S. government and refugee p
40 ducing event (e.g., the elite in Cuba faced more persecution under the Castro re
41 tates. [220] It allows a person being faced with the threat of removal to obtain
42 he States in the Western Balkans were faced with a set of multiple challenges. T
43 may be different, and the challenges faced by today's societies different from
44 man and Austro-Hungarian Empires, was faced with multi-faceted challenges includ
45 reflect the realities and challenges faced by most or all countries in the regi
46 iolence, discrimination and prejudice faced by persons or groups of persons on g
47 different issues and needs than those faced by prisoners or even other aliens, a
48 the refugee-producing event that they faced-the persecution or threat of persecu
49 protection and assistance challenges faced by various categories of persons on
50 light on the multi-faceted challenges faced by Governments in the region, the pa
51 of the biggest challenges the country faces. Confirmed speakers will include Pro
52 he biggest problems the Welsh economy faces, yet it has not become a major issue
53 the desert border with Egypt. Israel faces substantial illegal immigration of A
54 in situations when an EU Member State faces a sudden increase of irregular migra
55 . If California continues to grow, it faces environmental degradation. Many of t
56 icies to promote protection of those facing life threatening situations. Althoug
57 t refugees asylum seekers and others facing life threatening situations if retur
58 ations to address current challenges facing the US asylum system. The first set
59 ration was the most pressing problem facing immigration authorities, a perceptio
60 Recognizing that one of the problems facing California, if not the United States
61 on - with the world's poorest people facing the greatest danger from possible ec

Figure 18 Sample of concordance lines of FACE in the 'immigration' reference corpus

From the concordance lines in Figure 18, we can see very clearly that FACE always has a negative association and it co-occurs with abstract nouns. These are among the collocates which appear in the immediate environment of FACE, to the left and right, and carry a negative meaning:

persecution, challenges, problems, double-whammy, difficulties, limitations, restrictions, discrimination, prejudice, constraints, risk, arrest, jail, shortfall, loss, conflicts, barriers, threat, closure, harm, torture

Most of these collocates agree with the list that Tognini-Bonelli (2001) introduces for the verb *face*. The item *persecution*, however, does not appear in her list. This indicates that this collocate might be specific to the context of immigration. To sum up, FACE prefers a set of items denoting ‘problems and challenges’ and also ‘suffering’. Thus, these meanings could construct the semantic preference of FACE. Besides, the semantic prosody is also fused with the semantic preference and they both, therefore, express the same meaning.

7.6.1 The textual behaviour of FACE in LOCNESS

As pointed out in Section 7.3, the topic of ‘immigration’ is not included as one essay title in the LOCNESS corpus. As a result, comparing the same topic in the reference corpus and native texts was not possible in this section. This means that I examined the semantic behaviour of FACE in native texts, which are about different topics rather than immigration. The first text that I selected is about ‘computer and human brain’. Example (34) below illustrates the use of FACE in this topic.

(34) **1** There is however a more dangerous **threat** from computers, // **2** it is that they can do the work for man. // **3** This could lead to high **unemployment**. // **4** Those people who work with computers for long periods of time every day face **problems**. // **5** The repetition of tapping keys all day and staring at the screen can be **harmful** and not only that it is highly **boring** to do the same thing over and over again.

In Example (34), the verb *face* takes place in T-unit (4) and collocates with the abstract noun *problems*. Furthermore, the whole paragraph carries a negative evaluation. This could be realised through lexical items that are in bold such as *threat*, *unemployment*, *harmful* and *boring*. This sample of native writing shows that FACE maintains the same semantic preference and prosody of ‘problematic areas’ that are identified in the reference corpus of immigration in Section 7.6. Another example, which is also about ‘computer vs. human brain’, illustrates how FACE is used by another native writer (the square brackets indicate T-units that are not presented).

(35) **1** Another moral **dilemma** that computers have created is their role in the workplace. // [...] **3** These computers have replaced humans and such changes have sparked **fears** that the level of unemployment could vastly increase as a result. // **4** As the computers are: more efficient than humans, do not require payment for their work, are less temperamental than humans and will never have a day off because it is ill, // **5** it seems mankind is faced with a great **problem**.

In Example (35), the verb lemma FACE occurs in T-unit (5) and it is in a passive-voice form *faced*. T-unit (5) serves as a summary unit to the previous paragraph and means that the problem of replacing humans with computers would have an effect on mankind. This meaning is expressed through *problem*, which occurs in the immediate span of *faced*. *Problem* characterises the typical semantic preference of FACE. There are also other lexical items in the text fragment above which are semantically related to *problem* (e.g., *dilemma*, *fears*). The

text fragment conveys ‘problems’ as a semantic prosody distributed across the text. The following two examples are produced by different native speakers about ‘traffic’. Both examples contain the lemma verb FACE.

- (36) **1** The basic **dilemma** facing the UK’s rail and road transport system is the general rise in population. // **2** This leads to an increase in the number of commuters and transport users every year, consequently putting **pressure** on the UKs transports network.
- (37) **1** Hundreds of people are **killed** on the roads every year, and as cars become faster and traffic denser, the number of accidents will surely increase. // **4** [...] Now that rail privatisation has gone ahead, many people are likely to lose faith in trains, due to the perceived inefficiency of the operators. // **5** Fares are likely to increase, and many rural lines that used to be subsidised by the government face **closure**. // **6** Very little freight is transposed by rail these days, // **7** so the railways have lost income due to this. // **8** Many tracks and trains are in need of renovation or replacement, but funds are not available for this.

In both Examples (36) and (37), the verb lemma FACE prefers to co-occur with lexical items such as *dilemma*, *pressure*, *killed* and *closure*. Most of these lexical items have been identified in the reference corpus in Section 7.6. In general, these items denote problems, difficulties or suffering. The selection of these items by native speakers of English indicates their attitude to the topic, which is negative. To sum up, native speakers of English seem to be aware of the semantic preference and prosody of the verb lemma FACE. In the following section, I re-examined the semantic behaviour of FACE in the Arab non-native essays from ALEC.

7.6.2 The textual behaviour of FACE in ALEC

I started the analysis with an essay about ‘immigration’. In this essay, the Arab non-native writer uses the word-form *face* two times. Example (38) below illustrates an extract from this essay.

(38) **1** Economically, the opponents to immigration would argue that immigrants would compete with the original people on housing and job opportunities which could bring a significant economic burden by increasing prices of houses and decrease the salaries offered by the employers. // **2** While this is true in reality, the governments could minimize this effect by controlling the job market minimum wage paying and enforce big companies and universities to build appropriate accommodations for their employees on long term or permanent contracts. // **3** Indeed, immigrants would **refresh** and **boost** the job market as they **encourage** the investment in those countries especially when the investors are seeking the availability of workers accepting a reasonable levels of paying. // **4** Consequently, the economy will face a significant **growth** and **development**.

In T-unit (4) in Example (38), the Arab non-native learner successfully selects abstract nouns to co-occur with *face*. However, the nouns *growth* and *development* in T-unit (4) are not found as collocates of the verb *face* in the reference corpus or in the native texts. In addition, these lexical items do not reflect the semantic preference or the negative semantic prosody that have been both established in Section 7.6. This means that FACE goes against the norm of usage in the Arab non-native learner’s text. Instead of *face*, the writer needs to use verbs such as *benefit from* or *bring*, for instance, which carry a positive connotation that can fit in with the rest of the paragraph. I further looked at the immediate environment that surrounds *face* in the extract above, it contains other positive items such as *refresh*, *boost* and *encourage* (T-unit 3).

Interestingly enough, the Arab non-native writer repeats the verb *face* in the same text in a different paragraph. However, he/she uses *face* with a different communicative purpose from the one that has been mentioned above in Example (38). More particularly, the Arab non-native writer this time expresses the meaning of semantic preference of ‘problems’ by selecting the lexical item *difficulties* to co-occur with *face*. The item *difficulties* is an abstract noun and carries negativeness. Example (38-a) is an extended text of Example (38).

(38-a) On the other hand, it would be argued that immigrants would face **difficulties** in homogenizing themselves in the new culture whatever the duration of their settlement. // There is no doubt that the roots of the immigrants remain influencing their behavior.

Example (39) below is also taken from the ALEC corpus but it is about ‘traffic’. This example is a further illustration of how FACE is used by Arab speakers of English.

(39) **1** Developed countries face **no** overpopulation **problems**, and in some cases under population is actually the main concern. // **2** By investing in such human capital and incorporating them into the diminishing population as citizens and providing them with employment or entrepreneurial opportunities, this economic burden will turn into an economic **benefit**.

In Example (39), the communicative function or the semantic prosody of *face* does not indicate negativeness, or more specifically this use of *face* is not in line with the established semantic preference and prosody that express ‘problems’ in Section 7.6. Although *face* in Example (39) is followed by *overpopulation problems*, which refers to a problem, the use of *no* turns the idea of this fragment into a positive one. The paragraph above carries the meaning that immigrants will help developed countries compensate for their low population and they will bring economic benefits. Thus, the overall evaluative meaning of the paragraph

is positive rather than negative. The negative item *no* does not collocate at all with the verb *face* in the immigration corpus (see Section 7.6). I also checked the written part of the BNC and it showed that *no* collocates with *face* only 10 times. The relative frequency of its occurrence relative to the size of the BNC corpus, which is 87,903,571 tokens of running text, is 0.113. This means that *no* is not a strong collocate with *face*.

To conclude this section, the semantic behaviour of FACE in the Arab non-native texts fluctuates between good and bad from a paragraph to another within the same text (Examples 38 and 38-a). FACE, as pointed out by Tognini-Bonelli (2001), has a negative semantic prosody and not neutral as it is expressed by Arab speakers of English. The Arab L2 learners also use FACE in a positive atmosphere despite their use of lexical items that denote problems (cf. Example 39). This implies that the Arab non-native learners are perhaps unaware of the semantic prosody of the verb *face*. Thus, they use the same verb one time to express a positive meaning, and another time to indicate something bad.

7.7 Lexical cohesion through the interlocking of textual functions of lexical items

The previous sections described the three verb lemmas (REDUCE, AFFECT and FACE) separately in different examples from native writing and Arab learner writing. In this section, I gathered these text fragments together in order to provide a complete picture of how networks of lexical cohesion are established when the three verb lemmas interact with each other in the same text. Most parts of the texts that were used in this section were analysed in detail in the earlier sections of this chapter. However, in this section, these parts were examined collectively. In the two essays presented below, the T-units were numbered and displayed in order so that they indicate the cohesive development of the text. The node items

of the cohesive network were underlined and circled in order to highlight that they represent the prime items or the start items that attract particular collocates to them. The lexical items belonging to the semantic preference and prosody are in bold. The first essay is written by a native writer, and contains the three lemmas simultaneously. Figure 19 illustrates the cohesive networks that REDUCE, AFFECT and FACE create in the native text.

1 Away from cities road construction faces the **problem** of geographical sites. // 2 The shape of the land and its landforms extremely **hinder** the paths of new ventures. // 3 Rivers need bridges to cross them costing more money which is not already there. Which means construction is **stopped** until further time when resources and money again become readily available. // 4 **Solutions** to any road **problems**, are always going to cause groups of people **harm**. // 5 **Congestion** could be reduced if more and better uses of public transport were used. // 6 making people use buses and "underground systems" seems logical when they are operating anyway. // 7 Persuading people to leave their cars at home must become a major issue. // 8 Increasingly people cycle to work and walk, but because of the unreliability of U.K. weather, a rainy day means back into the car and comfort to work. // 9 A **solution** attempted to reduce **pollution** and traffic **problems** in Mexico City was to only allow certain coloured cars onto the roads on specific days. // 10 Surely the U.K. can come up with something more socially acceptable than this. // 11 Some say increasing prices on petrol and taxing of the use of roads will help. // 12 This is **disastorous**, as this only affects the **poor**. // 13 Why should those with more disposable incomes be allowed to pay these taxes and use the roads. Where the **poor** who have as much right as anybody **struggle** to pay extra road tax. // 14 This system simply will lead to the richer earners to use our roads.

Figure 19 REDUCE, AFFECT and FACE cohesive networks in a native essay from LOCNESS

Figure 19 shows that although the three lemmas REDUCE, AFFECT and FACE enact different semantic preferences and prosodies, they are still related in some respects. By examining these three items in the native essay in Figure 19, it is noticeable that they share a

number of general semantic features. For example, they all co-occur with abstract nouns that usually denote problems or a negative state of affairs. To start with, *face* in T-unit (1) co-occurs with a *problem*. Then, in T-unit (4), the lexical item *problem* is repeated but in a plural form. This repetition of *problems* forms with the previous mention of *problem* a cohesive link of simple repetition and it also takes place in the surrounding environment of the word-form *reduced*. Thus, we can see how the textual patterns of lexical items interweave through their shared collocates. In the same T-unit (T-unit 4) in which *problems* occurs, there is also an occurrence of the noun *harm*, which, on the one hand, implies ‘suffering’, which has been identified as one of the semantic preferences of *face*. On the other hand, *harm* refers also to an unfavourable thing that could still be semantically and collocationally relevant to the lexical item *reduced* in T-unit (5).

In T-unit (5), the item *congestion* also appears as one of the collocates of *reduced*. *Congestion* further acts as a hyponymy of the general word *problem* in the previous T-units. Then, in T-unit (9), the node item *reduce* is repeated and its textual pattern is repeated as well. More specifically, in T-unit (9) the item *problems* appears again as one of the collocates of *reduce* along with the item *pollution*, which both refer back to *congestion*. I could then posit that many of the collocates of *face* and *reduce* intersect. With reference to the item *solution* in T-unit (9), it belongs to the semantic preference of *reduce*, and it links back to *solutions* that appears in T-unit (4) establishing a cohesive link of simple repetition. If we move on to T-unit (12), there is a mention of the adjective *disastrous* that occurs in the co-text of the node item *affect*. The adjective *disastrous* links back to the previous occurrence of the related item *harm*, and this adjective also implies a problem. So, *disastrous* can be related back to the earlier mentions of *problem/s*. The semantic preference of ‘hindrance’, however, that has been

identified to describe *affect* in the reference corpus is not present in the co-text of *affect*. Instead, this meaning is distributed in other places of the text, particularly, in the immediate environment of *face*, in T-units 2 and 3 (*hinder, stopped*). This is an indication that semantic preference and prosody of a lexical item are subtle and they do not necessarily occur in the immediate span of the lexical item as linear links, but they might be spread in different places in the text forming non-linear links with their node items. Mahlberg (2009: 117) points out that “non-linear links illustrate how different lexical items merge beyond the occurrence in a sequence”.

These semantic preferences of the three verbs helped uncover the overall semantic prosody in the native essay above. The semantic prosody of every node item (e.g., *face*) shades into prosodies of other lexical items (e.g., *reduce* and *affect*) and creates a mechanism of lexical cohesion. For example, the node item *face* indicates the meaning of a ‘problem’, as a semantic prosody. It further describes the causes of this problem (cf. T-units 2 and 3). Then, through the node item *reduce*, a potential solution to this problem is provided. However, *reduce* expresses ‘an uncertainty’ of finding a proper solution to the problem and this can be realised through lexical items such as *could be* (T-unit 5), *attempted* (T-unit 9). This uncertainty is created due to the fact that the suggested solution is usually ineffective, which could be understood from these expressions (*cause... harm; affect the poor*). Such a negative effect is communicated through the node item *affect*, which demonstrates that any attempted solution will have bad effects or consequences that usually cause a negative change that is undesirable.

Based on this analysis of semantic prosodies of each node item in the native essay, we saw how the three lexical items interlock and convey a discourse prosody (Stubbs: 2001) that is generally negative. Even though the node item *reduce* has inherently a favourable prosody (cf.

Partington 2004a), it is used in the ‘traffic topic’ to express ‘uncertainty and skepticism’. Moreover, the analysis of semantic preferences and prosodies of the three lexical items reveal an important aspect of the text, which is that lexical items can organise the text. The three node items connect two different textual patterns, which are very common in argumentative writing: a Problem-Solution pattern through *face* and *reduce* and Cause and Effect pattern through *face* and *affect*.

Next, I focused on the same three verb lemmas REDUCE, AFFECT and FACE in an Arab non-native essay that is also about ‘traffic’. I examined particularly how Arab L2 writers establish the networks of lexical cohesion through the semantic preference and prosody of the three items as compared to the native writers. Figure 20 illustrates how the three lemmas interact with each other in an individual non-native essay.

1 I agree in rising the fuel fees, // 2 people will start to think of the money spending carefully // 3 and the transportation system of the country will **face** **growth** and **improvement** in more bus lanes with fixed time scheduled to get more people from-to their work. // 4 The effects on increasing the fuel fees from the economical impacts will help on maintaining fuel consumption to some limitation // 5 because of the rising there will be more invention to replace the use of fuel// 6 and people will start to think for other solutions, // 7 economically this is very good to increase the industrial and the manufacturing process across any country. // 8 The rising up will limit the **pollution** problem and will keep a friendly environment // 9 as people will lead to use less fuel, park their cars and start walking instead which then will help to keep them **active** and get **healthier** and be **fit**. // [...] 10 The use of alternatives like bicycles and walking will **affect** their health on **good** way and keep them as **active** and **focused**. // 11 The social factors that raised in such statement can be found in one family for example; who has 3 cars, the father drives on his own, the mother has her own car and the son/daughter uses his/her car, // 12 for this the family will be apart from each other and no interacts between them. // 13 This issue rises and the suggestion of increasing the fuel fees will limit these cars and all families will be getting one car, // 14 it will keep the family together all the time whenever they travel and keep this kind of perfect relationship we miss these days [...] 15 The increase of fuel fees must follow an increase on the salaries for people to cope with their daily expenses, // 16 the idea from increasing the fuel prices in the first place is to limit the number of cars that is been used by most working people/not working, even student at some stages // 17 they all intend to use cars and drive instead of taking shuttle bus or even walk short distances. // 18 Such cases can be found in some countries whom are very rich in oil and gas, // 19 these counties for example Libya the fuel cost as little as 0.150 LYD a Litre which is about 7p:10p GBP. // 20 That figure is very cheap // 21 and this is result of low rate in salaries, // 22 people do **face** the **trouble** of car traffic // 23 and too many people driving because there is no alternative for buses or any other transportation system.

24 To sum up, increasing the fuel fees has **affected** us **positively**. // 25 It has **advantages** in **reducing** the cars on the streets and limit these cars to 1 one car per family. // 26 It **helps** in keeping the family members traveling together. // 27 It will **help** in limiting the **pollution** // 28 and people will breath fresher air rather than the **smoke** of car exhausters. // 29 And to do all these the country itself needs to make changes and start by increasing the salaries and work on alternatives, buses, trams and trains inside each city and across the whole country.

Figure 20 REDUCE, AFFECT and FACE cohesive networks in an Arab non-native essay from ALEC

In Figure 20 above, the Arab non-native writer supports the argument that is in favour of rising fuel fees. The writer believes that this solution has economical, environmental and social benefits. In order to construct a positive argument, the Arab non-native writer makes use of lexical verbs such as *face* and *affect*. Figure 20 shows that these verbs provoke different cohesive networks from those in the native text in Figure 19. To exemplify this difference, the first mention of the verb *face*, in the Arab non-native essay, takes place in T-unit (3). *Face* occurs with a pair of abstract nouns *growth* and *improvement*. These nouns carry a positive meaning in contrast to the noun *problem* that occurs with *face* in the native essay, as illustrated in Figure 19. As a result, the semantic preference and prosody of ‘problems’ are not present at all in the Arab non-native essay.

Then in the subsequent T-units, the Arab non-native writer keeps the argument positive across the essay. For example, in T-unit (9), the writer uses adjectives such as *good*, *active*, *healthier* and *fit* that appear in the surrounding environment of *affect*. Then, in T-unit 10, the non-native writer uses the node item *affect*, which associates with positive adjectives *good*, *active* and *focused* that link back to previous adjectives *good* and *active* in T-unit (9). Therefore, the semantic prosody of AFFECT, which indicates ‘an unfavourable change’, is absent. After a long distance of 12 intervening T-units, the node item *face* is repeated in T-unit (22), and occurs with *trouble*, which does not build on the previous use of *face* in T-unit (3). This, consequently, creates an inconsistency in the lexical cohesive network. In T-unit (24), the node item *affect* is repeated (the word-form *affected*) and occurs with the adverb *positively* that refers back to the adjective *good* in T-unit (10). With reference to the node item (*reducing*) that occurs in T-unit (25), it somewhat maintains its semantic preference and prosody in the Arab non-native essay above. It co-occurs, for example, with similar semantic

fields to those identified in the reference corpus and the native essay in Figure 19. The item *pollution*, for instance, takes place in the extended co-text of *reducing* in T-unit (27), and relates back to *pollution* in T-unit (8). In addition, there is the item *smoke* in T-unit (28), which links back to the previous mention of *pollution*. Both *pollution* and *smoke* represent a semantic preference of ‘environmental problems’. However, the semantic preference of ‘solution’ is not present, compared to the native essay in Figure 19. For the semantic prosody of *reduce*, which is described as having the meaning of ‘uncertainty’, it is relatively noticeable with the use of the verb *help* that occurs in T-units (26) and (27). Although the node item *reduce* (in the word-form of *reducing*) shares a number of its textual patterning with *reduce* in the native essay in Figure 19, it does not interact cohesively with *face* and *affect* as it does in the native essay where collocates of some of these verbs intersect and establish a consistent lexical cohesive networks. This is perhaps due to the clash between the textual patterning of *reduce* and the other two node items of *face* and *affect* in the non-native text.

I can conclude that the Arab non-native writer attempts to employ the three node items cohesively by keeping the general attitude of the argument positive and selecting positive items throughout the essay. However, the textual patterns of the node items that the Arab non-native writer uses in the building of lexical cohesive networks do not match those identified in the reference corpus. Furthermore, the non-native writer changes the semantic prosody of a lexical item depending on the argument he is developing whether positive or negative. This could be seen in the essay through the verb *face*. Overall, the cumulative meaning of the textual patterns of the three lexical items in the Arab non-native essay differs too much from that in the native essay. In this analysis, I only looked at cohesive links in the essay from the

point of view of three lexical items. Mahlberg (2009: 117) suggests that “[t]he picture will only be complete if we do the same for every item in the text. Then we can truly describe how the ‘interlocking’ and ‘overlapping’ of prosodies works in this text”.

7.8 Conclusion

The analysis of lexical cohesion by means of a corpus linguistic approach complemented the analysis of lexical cohesion, which was conducted in the previous chapter using a text-linguistic approach. In this chapter, the corpus linguistic analysis was useful in adding further detail to the description of the function of lexical cohesion, and showed that lexical cohesion is not simply a one-to-one relationship between individual lexical items. Instead, it is the accumulation of the relationships between the textual patterns of a lexical item across text. This finding answered the third research question in this thesis, which examines the role of corpus linguistics in describing the function of lexical cohesion.

The analysis of the three verb lemmas in individual native and non-native essays showed that lexical cohesion works differently at the level of semantic preference and prosody in both varieties of argumentative writing. These differences were of various types. The first case has been where Arab non-native learners maintained the semantic prosody of an item but deviated partially from its semantic preference. This partial deviation has been made, for example, by selecting the wrong lexical item from the lexical set that characterises the typical meaning of the semantic preference. This local divergence from the typical semantic field of a lexical item will stimulate different collocational patterning around the node word whenever it is repeated throughout the text. This, consequently, will disrupt the cohesive structure of a text. A further difference between both groups of learners has been when Arab non-native writers

used the wrong semantic prosody of a lexical item, or they sometimes change the semantic prosody of the same lexical item within the same text as they move on from one argument to another in their text. In both cases, the progression of the text argument as well as the cohesive harmony of the text as a whole will be affected, because using a wrong semantic prosody will produce, in turn, cohesive links that are also different.

These differences between native and non-native learner writing imply that Arab non-native learners are perhaps not aware of the fact that semantic prosody has to be respected. Tognini-Bonelli (2001) points out that this is not a matter of a mistake proper, but it is a lack of the quality of naturalness in language. She maintains that the potential danger for the unaware language student is the clash with a norm. She contends that a native speaker can expect that a certain lexical item has to co-occur with something negative and he will, hence, successfully create an aura of negativeness. In contrast, the non-native student lacks this intuition and may select a positive word that follows this lexical item.

The findings reported in this chapter, nevertheless, reflected only personal and idiosyncratic uses of a number of individual writers. Other writers from the ALEC corpus might have used the three lemmas correctly. However, my aim in this chapter was to find the odd uses of the lexical items in question because such uses might trigger the attention to an interlanguage phenomenon that requires a deep analysis. Furthermore, my analysis did not use statistics to measure the strength of the collocations that are responsible to determine both semantic preference and prosody of the lexical items under analysis. Instead, the analysis was based on observations for the repeated patterns in the reference corpus. The findings, therefore, could not be generalised and should be taken as indicative of the corpus size and the methodological steps proposed. In order to avoid the effect of any idiosyncratic use, the findings would still

need to be explored further and validated with a larger corpus of (Arab) non-native English texts. Only in this case, we might treat the texts as one discourse community which would then represent their typical linguistic use of specific lexical items. In addition, it would be more valid if a larger number of lexical items have been examined for their semantic and pragmatic behaviour, because this can depict more types of lexical cohesive networks. Overall, these findings are still valuable because they provide preliminary statements regarding the use of semantic preference and prosody. They also draw the attention to the role that semantic preference and prosody play in establishing lexical cohesion, an area that has not received sufficient consideration in the previous literature about cohesion.

Chapter 8

Conclusions and implications for pedagogy

8.1 Introduction

As highlighted in Chapter 1, L2 writers encounter difficulties in the use of lexical cohesion. Therefore, this study compared the frequency and function of lexical cohesion in English argumentative writing by English native speakers and Arab speakers of English to better understand the specific problems that Arab learners face. This combination between frequency and function in the analysis of lexical cohesion helped to reveal differences between the two corpora and identified specific strategies that Arab speakers of English tend to use when they are trying to achieve lexical cohesion. This study also aimed to overcome some of the limitations of the classic models of lexical cohesion for the analysis of texts, and suggested a systematic approach to analysing lexical cohesion that is replicable across corpora. The current chapter brings the results of the study together. In Section 8.2, I will summarise the key results that have answered the research questions I presented in Chapter 1. These results were obtained by combining a text analysis with a corpus analysis. Based on these results, in Section 8.3, I will highlight five key contributions of this thesis. Then, in Section 8.4, I will look at the pedagogical implications for teaching lexical cohesion in the L2 writing classroom. Next, I will suggest directions for future research in the field of lexical cohesion analysis (Section 8.5). Finally, an overall conclusion will be presented.

8.2 Summary of the results

The results will be divided into two sections. Section 8.2.1 will outline the results in respect to the frequency of lexical cohesion in ALEC and LOCNESS. Section 8.2.2 will then focus on

the results that are related to the function of lexical cohesion in both corpora. For practical reasons as I demonstrated in Chapter 2, my analysis focused on three categories of lexical cohesion: simple repetition, derived repetition and signalling nouns.

8.2.1 The frequency of lexical cohesion

The first main research question in this thesis sought to quantify simple repetition, derived repetition and signalling nouns in 58 argumentative essays produced by Arab L2 writers and English native writers. As stated in Chapter 1, the first main research question contains three parts, as it is indicated below:

RQ1: What are the relative frequencies of the tokens of each lexical cohesive category in each variety (NNS vs. NS)?

- How many instances/tokens of simple repetition can be counted in each corpus (NNS vs. NS)?
- How many instances/tokens of derived repetition can be counted in each corpus (NNS vs. NS)?
- How many instances/tokens of signalling nouns can be counted in each corpus (NNS vs. NS)?

To answer the first two sub-questions that examine the tokens of simple and derived repetition in both corpora of argumentative writing, I suggested my model of the Lexical Repetition Network (LRNetM). LRNetM is based on corpus tools, i.e. the wordlist as a strategy to help

find instances of simple and derived repetition. LRNetM adopted Thornbury's (2010) view that wordlists provide the base for building semantic networks of the text. As explained in Chapter 5, LRNetM depends on grouping an alphabetically sorted word-list into networks of simple and derived repetition. The wordlist method enabled to view a text as lexical networks of repeated lexical items that interconnect with each other. Each network in the wordlist moves out from a node word or a prime item and represents a relatively self-contained centre of unity. In each network in the essay under analysis, I applied my counting procedure to quantify instances of simple and derived repetition. This quantitative analysis, as illustrated in Chapter 5, considered individual-based frequencies and overall frequencies. The results, based on these two types of frequencies, demonstrated that the frequency of simple repetition was significantly higher in the writing of the Arab L2 group compared to that in the native group. Likewise, derived repetition was also higher in the English writing of the Arab L2 group than the native group. However, the difference between the two groups was insignificant. The third sub-research question addressed the frequency of SNs, which I calculated using a text analysis approach. This approach is more suitable to capture textual relations than LRNetM, which is mainly designed to identify and tally simple and derived repetitions. The results, either that were attained from the individual-based frequency or from the overall frequency, showed that SNs were significantly more frequent in the English writing of the Arab L2 group than that of the native group.

Overall, investigating the frequency of simple repetition, derived repetition and signalling nouns provided a basic understanding of the structure of the two corpora in terms of the number of lexical repetition each corpus contains. However, differences in the frequencies of the three lexical cohesive forms between the two corpora did not contribute to understanding

the reasons that made frequency higher in the Arab L2 writing compared to the native writing. As emphasised in Chapter 2, researchers such as Reynolds (1995) observe that quantitative measures of repetition do not pinpoint any significant differences between NNSs and NSs. Reynolds (1995) suggests that differences can be marked by examining how NNSs and NSs use repetition to communicate. Hartnett (1986) maintains that researchers should not link overall writing quality to the quantity of cohesive devices. They should instead pay attention to the function of these devices. Therefore, the frequency analysis of lexical cohesion was complemented with a functional analysis. This analysis was important to know whether high frequency of lexical cohesion in the Arab L2 writing indicates a good writing.

8.2.2 The function of lexical cohesion

The function of lexical cohesion in both corpora was firstly assessed by means of a text analysis (cf. Section 8.2.2.1). Then, the text analysis was reinforced with a corpus analysis to comprehensively understand the complexity of the function of lexical cohesion (cf. Section 8.2.2.2).

8.2.2.1 Text analysis

The second research question in this thesis aimed to examine the paradigmatic choice of simple repetition, derived repetition and signalling nouns. This means how these lexical cohesive forms associate with each other at the non-linear level in the argumentative writing of NSs and NNSs. This question was phrased in Chapter 1 as follows:

RQ2: What are the advantages and disadvantages of a text linguistic approach in describing the function of the lexical cohesive forms in each corpus (NNS vs. NS)?

The examination of this question was conducted from the viewpoint of text-analysis. In Chapter 6, the qualitative analysis revealed that the Arab L2 group used simple repetition, derived repetition and signalling nouns redundantly without adding new information that can contribute to the development of the argument across the essays. In contrast, the native group employed lexical cohesion in a way that helped them develop their argument by adding new materials. These results conform to Ouaouicha's (1986) findings that English arguments of Arabic speakers, as compared to the native writing, contain more data but fewer claims, warrants, backings, and rebuttals. These results also confirm what Connor (1984) observes that ESL writing has lexical and conceptual redundancy that does not allow them to extend concepts they introduce. The results further showed that the Arab L2 group transferred a number of linguistic features from their L1 (i.e. Arabic) to their English writing. Such an L1-transfer was particularly obvious when the Arab L2 group used lexical strings or couplets which contain a pair or more of SNs. This use of lexical couplets was not found in the writing of the native group.

Generally, the text-based analysis showed how NS and NNS writers construct their argument through simple, derived repetition and SNs at the paradigmatic level. This analysis constitutes, in my view much-needed, alternative to the limited quantitative approach to lexical cohesion currently available on this topic, as indicated Chapter 2. However, this analysis of lexical cohesion that is based on the traditional text analysis stops at fixed phrases or describes cohesive devices by clear-cut categories. This point of view towards cohesion simplifies how language is described. As Mahlberg (2006) points out, language is action and meaning is use, and as a result lexical cohesion should be described in a more rigorous

fashion by analysing the way in which the flexible boundaries of lexical items link in with other lexical items. Thus, a further analysis was needed in the present study.

8.2.2.2 Corpus linguistics

This section addressed the third research question which investigates the following:

RQ3: What are the advantages and disadvantages of a corpus linguistic approach in adding further detail to the description of the function of the lexical cohesive forms in each corpus (NNS vs. NS)?

The use of a corpus linguistic analysis was important to complement the qualitative results that were obtained from the text analysis. Therefore, as explained in Chapter 7, these results were further checked through applying corpus-linguistic concepts, namely semantic preference and prosody. Three lexical items *reduce*, *affect* and *face* were selected from both corpora as case studies. These lexical items already showed a lexical cohesive behaviour by means of simple and derived repetition in both sets of data. These lexical items were examined this time through analysing their typical semantic preference and semantic prosody in a reference corpus (collected through Sketch Engine). Then the cohesive behaviour of these lexical items was compared against ALEC and LOCNESS. The findings revealed that a number of Arab NNS essays deviated from the typical semantic preferences and prosodies of the selected lexical items. This as a result broke up the lexical cohesive harmony in their writing. Native speakers of English, on the other hand, showed more respect to the semantic preference and prosody of the items under analysis and their writings were more cohesive as such. This analysis of semantic preferences and semantic prosodies suggested that a corpus theoretical approach to cohesion has been able to garner a deeper set of insights into the

dynamics of lexical cohesion in native and non-native writing. It revealed how lexical cohesion functions in a text by describing how lexical items interconnect and hang together creating a coherent text.

8.3 Contribution of my thesis

The innovative contribution of this thesis is that I attempted to bring together text analysis and corpus analysis to provide a detailed analysis of frequency and function (i.e. meaning) of lexical cohesive relations in English argumentative writing by native and non-native speakers of English. Such a mutual relationship between frequency and function involved combining both quantitative and qualitative features of lexical cohesion to have a comprehensive understanding of the problems that Arab L2 learners have in their use of lexical cohesion. Based on these key aspects of my research, my thesis has made five main contributions.

The first contribution of this thesis is the creation of the ALEC corpus, which contains argumentative writing produced by Arab speakers of English. I pointed out in Chapter 4 that this type of corpus is absent from the existing learner corpora. In this thesis, the use of the ALEC corpus enabled to understand the challenges that Arab L2 learners confront in their use of lexical cohesion. This corpus can be used as a starting point for linguistic descriptions or as a means of verifying hypotheses about the English writing produced by Arab L2 learners. ALEC could also be either added to ICLE or published in future as a new corpus in the field of learner corpora.

The second contribution of this thesis is that I have developed a set of criteria for determining signalling nouns that take into account a specific variety of learner writing. This analysis has

been conducted by means of a text analysis. As pointed out in Chapter 2, most studies on SNs in the previous literature focused on SNs that are commonly used among L2 learners, but they fail to examine L1-specific tendencies of using SNs. Such tendencies, however, deserve more explicit attention, as they may bring to light the various rhetorical purposes for which L2 learners from various L1 linguistic background employ these nouns. With the criteria of analysis in the present work, this thesis, as demonstrated in Chapter 6, has been able to identify specific patterns of SNs that are distinctive to the writing of the Arab non-native group. This, in turn, highlights some features of rhetorical uses and interlanguage cohesion. My thesis further provides a unique contribution to SNs by advancing our understanding of the way SNs signal the structure of a text, and how they can reveal differences in culture between native and non-native writers.

The third main contribution of my thesis is my proposal of the Lexical Repetition Network Model (LRNetM). As demonstrated in Chapters 2 and 5, LRNetM identified areas of inconsistency in the application of the classic models of lexical cohesion to text analysis, and it therefore developed a systematic method of analysis that can be replicated by other analysts. What LRNetM achieves is that it gives us a way of analysing lexical cohesion that has no problem with directionality. Directionality, as demonstrated in Chapter 2, is a complex notion that was employed by the classic models of lexical cohesion mainly Halliday and Hasan's (1976) model and Hoey's (1991b) model. These models analyse cohesion by tracing backward or forward cohesive referents that involve having a specific direction in a text – whether this direction is of a unidirectional or multidirectional type. This kind of pattern is the basis of most accounts of cohesion. By suggesting the LRNetM model, the focus in the description of lexical cohesion is on the communicative function of each T-unit and not on

what the textual antecedents might have been. The T-unit is the current unit of a text that is being processed and not dependent on anything else (Sinclair 2004). Furthermore, one of the main principles of LRNetM is that it takes into account Hoey's (2005) conception of cohesion that cohesive properties are inbuilt in the lexical item itself and hence each lexical item could be primed for cohesion. LRNetM put these concepts into practice by adopting a corpus linguistic wordlist method, as explained in Section 8.2. This quantitative procedure can handle large amounts of data in contrast to the classic models of lexical cohesion that are based on the analysis of single texts and invented sentences. As emphasised in Chapter 5, the way by which I designed LRNetM makes it possible to compare lexical cohesion in native and non-native speaker writing from a corpus-linguistic perspective. This means that the methodology that I adopted in this thesis to count lexical cohesive forms is that each corpus of native and non-native speaker writing was not analysed as a single entity. This methodology is not a typical method of corpus-based studies. Rather, the whole texts in each corpus were analysed individually. On the one hand, these individual texts make up the native corpus and the non-native corpus of argumentative texts. On the other hand, the integrity of each text was preserved during analysis. Each text in the corpus was treated as a self sufficient set of data. From the separate but related analyses of the texts in each corpus, I was then able to make general statements about the frequency use of lexical cohesion in the argumentative texts produced by each corpus.

The fourth contribution of my thesis has been the analysis of the function of lexical cohesion at the two levels of language patterning: the paradigmatic and the syntagmatic levels. By interlinking these two dimensions, the description of the function of lexical cohesion depends on combining forms and function. This indicates that the cohesive function of a lexical item is

more complex than it has been described before where the function was limited to the paradigmatic choice only. My thesis showed that the function of lexical cohesion has to take into account lexical, grammatical, semantic, and pragmatic patterning. It is this relation that links lexical and textual properties which is important to provide the full picture of how cohesion functions in a text. Mahlberg (2006) also stresses that cohesion created by lexical items works on these two dimensions. However, her analysis was restricted to the Newspaper Domains 'UK broadsheets', and included only one lexical item, *true feelings*. The paradigmatic and the syntagmatic levels have not been combined before in the study of lexical cohesion in learner writing, particularly in Arab L2 writing. Building on Sinclair's (2004) descriptive model of a lexical item, my research, as demonstrated in Chapter 7, was able to examine the function of lexical cohesion applying corpus concepts of semantic preference and prosody to individual essays. This approach makes it possible to study cohesion as a feature that spans across the whole of a text, incrementally contributing to the construction of textual meanings and the overall tone of the text.

The fifth contribution is the comparative analysis of lexical cohesion between Arab speakers of English and English native speakers. This comparison provided specific findings on the use of lexical cohesion by Arab speakers of English. The results of this comparison accumulated evidence that Arab speakers of English have problems in using lexical cohesion at both the paradigmatic and the syntagmatic levels. Overall, Arab speakers of English overused lexical cohesive forms compared to native speakers. However, this high frequency of cohesive forms provides a false picture of the cohesiveness of the Arab L2 writing. Based on these results, I can suggest direct pedagogical implications for teaching cohesion in the L2 writing classroom. One main suggestion is to shift the teaching of lexical cohesion to an approach that

focuses on function and text instead of only form. This approach needs to show how overt surface links contribute to the interpretation and the communicative purpose of a text.

8.4 Pedagogical implications for teaching lexical cohesion in the L2 writing classroom

In Chapter 1, I stressed that the findings of the present study will be of most immediate relevance to ESL/EFL teachers and learners. The results showed that L2 learners have difficulties in using lexical cohesion at the functional level. A number of factors have led to these difficulties. I indicated in Chapter 1 that ESL teachers conceptualise lexical cohesion as a list of formal and semantic devices that can be used to connect the text, and they apply this belief to the classroom. As Mahlberg (2006) points out, issues that we confront in pedagogic approaches to cohesion are not merely an outcome of the requirements of the classroom, but they also reflect general linguistic beliefs. Witte and Faigley (1981) suggest that cohesion can be better taught if it is better understood. A different factor which might be another reason for the problems that L2 learners have in using lexical cohesion was referred to by Forutan and Nasiri (2011) in Chapter 2. This factor is the absence of explicit teaching of lexical cohesion principles in second language writing classes. Therefore, Sections 8.4.1 and 8.4.2 will suggest a number of pedagogical interventions of teaching lexical cohesion to L2 learners. These interventions will mainly focus on raising L2 learners' awareness that lexical cohesion is not just lists of categories, but also has a complex function which works at different levels: lexical, syntactic, semantic and pragmatic. There is a case here for a conscious-raising approach, rather than a rote-learning one (Flowredew 2015). Section 8.4.3 will then suggest a direct application of a corpus linguistic approach to teaching lexical cohesion in the L2 writing classroom.

8.4.1 L2 learners need to recognise that lexical cohesion should provide a framework for new information and not to be used redundantly

As highlighted in Chapter 6, what distinguished the writing of native and non-native writers was that English native speakers used lexical cohesive forms (i.e. simple repetition, derived repetition and signalling nouns) to connect their T-units to develop their argument. In contrast, Arab speakers of English employed forms of lexical cohesion repetitively within and across T-units without performing any communicative function. This communicative function that Chapter 6 presented is the addition of new information by which the text argument can be developed. Hoey (1991b: 243) suggests that if a learner is to avoid clumsiness, he or she must be taught how to avoid it. McGee (2008: 219) also recommends increasing student awareness of redundant repetition in their writing. She points out that highlighting overuse of forms of lexical cohesion may help students to avoid this redundancy.

One of the most important ways to avoid unnecessary repetition of lexical cohesive forms is to advise L2 learners not to go around in circles (Hoey 1991b: 243) when they produce their English writing. As demonstrated in Chapters 1 and 2, Winter (1979) emphasises that using lexical cohesive forms, mainly lexical repetition, entails changes or additions to the repeated element which give it new meaning as a stretch of text. Hoey (1991b: 244) proposes that learners need to be reminded that new information always accompanies repetition in mature writing. For example, derived repetition could be a very effective technique that allows learners to avoid clumsiness and develop coherent texts, as Hoey (1991b) observes. However, as illustrated in Chapter 6, the Arab L2 learners used this category clumsily and did not use it as a means to add new information or to make the text more varied. This might indicate that

L2 learners are aware of derived repetition but they do not know how to employ it as a cohesive device.

One activity that the teacher can do to raise learners' awareness of derived repetition is to collect a batch of uses from students' writing where derived repetition is used excessively. The teacher can then present these uses in class and highlight examples of unnecessary use of derived repetition. Learners could then be asked to think of how they can employ derived repetition in a way that helps them develop their argument. Materials of native speaker writing could be used as a guideline to show how derived repetition is used to add informative value to the text. The teacher can also remove the unnecessary repetition of derived repetition from a number of students' writing and then rephrase some of these examples. Learners can then be asked to notice the difference of how writing becomes more concise than it was before. These activities can be also applied to simple repetition or signalling nouns. Arab L2 learners, for example, overused SNs by repeating the same SN over the text, or they used SNs to nominalise verbs or adjectives previously referred to in the text. This use of SNs added no information to the text which implies that L2 learners are not familiar with SN usage. Flowerdew and Forest (2015) recommend that L2 learners need to learn the subtle semantics of individual SNs and the various syntactic patterns, as they are used in discourse. L2 learners also need to understand that SNs have to serve a function, which is to provide a new characterisation of the discourse segment that it encapsulates or prospects. This means that the SNs that L2 learners use need to produce some fresh insight into the writer's understanding of the segment. Understating this function of SNs will help L2 learners develop their argument and push the discourse forward.

The superficial arguments and the lack of specific and new details that may convince the reader of the writer's point of view could be due to the insubstantial exposure to reading materials and general information. This has led to a poor level of concise and cohesive writing. The teacher could, therefore, spend some time on the topic that learners will write about and provide them with materials to read before they start writing. This reading activity will help learners to gain new information that they can use to support their argument in their writing. Malgwi (2016) suggests that reading for writing activities need to be increased and advises that reading activities in the ESL/EFL class should include a discussion on the lexical connections of the words within the text. This will expose learners to a rich variety of lexical cohesive forms that characterise texts with a high level of sophistication. Stotsky (1983) further stresses the need for a more reading/writing integrated approach in developing students' writing as this will raise students' awareness of the academic nature of good writing.

Students could also be taught to improve the informative function of lexical cohesion by teaching them the elements of thematic progression. As explained in Chapter 6, thematic progression is a strategy that writers can use to link the themes and rhemes in a clause to those of surrounding clauses. Alonso and McCabe (2003) point out that while ELT writing materials provided some focus on cohesive devices, little attention was paid to the progression of information in texts. Hawes (2015) suggests that thematic progression is a key factor in the organisation of information because it functions as a bridge between sentence level and discourse level, integrating cohesion with coherence. Therefore, as Wang (2007) proposes, students need to be taught how to arrange old and new information, and need to be reminded that this pattern is an important dimension for improving textual cohesion in English writing. Students, for example, can be trained to evaluate, in their writing, the theme-rheme structure

for a clause/t-unit to follow given clauses/t-units. Hawes (2015) further advises that students should be made more aware of the needs of their readers and of the importance of extended text coherence to the reader as opposed to merely local cohesion. He therefore recommends adding a module on information structure to language courses where appropriate.

The excessive use of lexical cohesive forms, as Chapter 6 showed, might be due to an L1 transfer. For example, Arab L2 writers used lexical couplets (e.g., *positive impacts* and *advantages*) that consist of a pair of SNs which are synonyms and refer to the same stretch of text. A lexical couplet/string is a common structure in Arabic language. However, L2 learners have to be aware that using this structure creates a monotonous effect on the English text. The teacher needs to instruct students to avoid such a redundant use of SNs and explain that only one SN is sufficient to condense the previous or the subsequent text. A different L1 writing strategy that was adopted by Arab L2 writers was the way how they structured their argumentative texts. L2 learners, for instance, structured their argumentative essays by using SNs that emphasise facts and problems while their writing lacks SNs that indicate the solution part. This focus on building up facts rather than developing arguments manifests the nature of persuasive discourse in Arabic, as explained in Chapter 6.

However, transferring such an L1-strategy into an English writing affects the cohesion of a text because focusing on facts involves the use of a great deal of repetition to emphasise a specific idea and achieve persuasion. The written product is as a result an argumentative essay which lacks information that is important to the reader to understand the text development. In contrast, English native writers did not concentrate only on facts and problems but they covered all parts of an argumentative essay, which include the situation, the problem, justification and the solution. Covering all these parts made their writing more informative

and hence more cohesive than the Arab L2 writing. In this case, L2 learners should be informed about L1-L2 differences in the text genre. The teacher could explain that in writing an English argumentative essay, the writer does not need to focus merely on presenting facts and problems as this creates unwanted repetition, and does not contribute to the development of the text. Rather, the writer has to learn that writing is an interactive process between the writer and the reader and both parties are important. Arab L2 Learners need to consider the reader when they write by answering all the potential questions that might be asked by the reader about: the situation, the problem, justification and the solution. Silva (1990) proposes that student and teacher awareness of the typical characteristics of the written modes may enhance effective teaching and learning. Learning genres, structures and lexis related to tasks relevant to the academic community is crucial in the successful development of students' writing.

To sum up, this section showed that the main pedagogical implications for teaching lexical cohesion is that L2 learners have to learn how to use repetition skilfully by balancing the given information with the need for variety and a change of focus. As Reynolds (1995) recommended, the use of repetition in a particular context has to be viewed with reference to its function and how it presents a smooth flow of information from one paragraph to another. If repetition is used reasonably without redundancy, repetition of lexical devices helps in engaging the reader in a text, facilitates comprehension, and most importantly, provides the basis for introducing new information, ultimately assisting the reader to establish the appropriate schemata.

8.4.2 L2 learners need to be aware of the function of semantic prosody in creating textual cohesion

The pedagogical suggestions in the previous section have addressed how lexical cohesion functions at the formal and semantic level. This function represents the paradigmatic use of lexical cohesion. The current section presents the pedagogical implications for teaching another function of lexical cohesion, which is concerned with how lexical cohesive forms integrate with each other in a text through their semantic prosodies to create textual cohesion. As presented in Chapter 7, the analysis of lexical cohesion showed that Arab L2 writers employed cohesive devices improperly in the wider context (i.e. at the syntagmatic level). They were not aware of the typical semantic prosodies associated with the usage of certain words which might suggest over-extension in usage. This, for example, has been illustrated by *affect* that a number of Arab L2 writers used in a positive environment which does not concur with the semantic prosody observations in the literature. This may possibly be due to the fact that a big challenge in learning a word or a unit of meaning lies in mastering its pragmatic function (Zhang 2008), which is related to its semantic prosody (Partington 1998; Zhang 2009; Hunston 2002).

Partington (1998: 8) further suggests that information about semantic prosody is “vital for non-native speakers to understand not only what is grammatically possible in their language production but [...] also what is appropriate and what actually happens”. For this purpose, a preliminary step may be to help ESL/EFL learners to understand the notion of semantic prosody, and then consider how to integrate semantic prosody in vocabulary instruction. Consciousness-raising activities of teaching semantic prosody are important for L2 learners. As Xiao and McEnery (2006) observe, ESL/EFL learners’ intuition of the second language is

typically less reliable than that of native speakers and thus cannot help such learners detect the usage of a lexical item in terms of its semantic prosody. Inter-language studies of semantic prosody also indicate that ESL/EFL learners, when learning a lexical item, seldom notice its semantic prosody and often make semantic prosodic errors in communication (Wang & Wang 2005; Wei 2006).

The teacher can, therefore, teach L2 learners that they need to master not only a lexical item's structural and semantic features, but also its pragmatic and evaluative meaning. Additionally, L2 learners are required to understand the function of semantic prosody in creating textual cohesion. For example, students can be instructed that when they repeat a lexical item using one of the forms of lexical cohesion such as simple and derived repetition, they need to bring with the repeated item its lexical, grammatical, semantic, and pragmatic patterning. The teacher can also compare the collocational behaviour and semantic prosody/preference of a number of lexical items in L1 and their close translation equivalents in L2, and make learners aware of L1–L2 differences. This activity will help reduce the number of errors from L1–L2 semantic prosody differences, as Xiao and McEney (2006) recommend. Teaching materials of lexical cohesion, therefore, need to take account of semantic prosody. One approach to teaching lexical cohesion by means of semantic prosody is corpus linguistics.

8.4.3 The application of a corpus linguistic approach to teaching lexical cohesion in the L2 writing classroom

The traditional method of vocabulary and more specifically lexical cohesion instruction must be improved and the traditional teaching concept should be changed, as Mahlberg (2006) suggests. As highlighted in Chapter 1, teaching materials that aim at developing writing skills

provide a simplified picture of lexical cohesion by, for instance, introducing a list of clear-cut cohesive devices such as synonyms, antonyms and connectors (e.g., *start – break out; on the one hand, on the other hand*). These lists are presented for learners to help them connect their arguments and produce cohesive texts. However, the emphasis of these materials, as Cheng (2009) observes, tends to be on surface devices and no explanation about how the different devices contribute to coherence is provided. More specifically, these materials overlook features that span larger contexts and that create complex networks of meaning relationships.

In Chapter 1, I argued briefly that a corpus linguistic approach to cohesion can help to cope with the limitations of the traditional approach to cohesion. With the observation of recurrent patterns of words, corpus linguistics highlights the importance of lexical patterns. This observation by corpus linguistics has implications for language teaching. Early ideas on lexically oriented and corpus informed approaches in language teaching were introduced by Sinclair and Renouf (1988), who outlined a ‘lexical syllabus’, which was further developed by Willis (1990). The central argument about the lexical syllabus, as Sinclair and Renouf (1988: 148) point out is that “for any learner of English, the main focus of study should be on (a) the commonest word forms in the language, (b) their central patterns of usage, and (c) the combinations which they typically form”. However, the lexical syllabus seems to move on too fast from lexis to larger units of discourse without paying enough attention to the role of lexis in creating lexical cohesion. This may be due to the fact that the lexical syllabus is informed by a corpus-linguistic theoretical approach to teaching whereas cohesion is principally not a corpus linguistic concept. Therefore, a lexical syllabus needs to be taken further to consider how we can move on from lexical patterns to connected texts. A possible suggestion is by using the concept of semantic prosody as a pedagogical tool. Analysing semantic prosodies

associated with lexical items in a single text is one way to reveal the link between lexical and textual properties of a text, which is important to the teaching of lexical cohesion.

One activity that the teacher can do in the classroom is to instruct students to use any available free software such as Antconc. Firstly, students, with the help of the teacher, are asked to run a wordlist of their written essays. Using students' own writing is advocated by Seidlhofer (2002) and Mukherjee and Rohrbach (2006). These researchers suggest that a learner corpus that contains students' own writings can be used directly for learning by coping with students' questions about their own or classmates' writings, or analysing and correcting errors in such familiar writings. Then, the teacher can prepare a stop wordlist and assist students to apply it to their written essays. Subsequently, by applying the lexical repetition network model (LRNetM), every student is instructed to group the wordlist of his essay alphabetically. This is to group the wordlist into networks of lexical cohesive relations. Such grouping will help students visualise the different types of lexical repetitions they used in their essays. For example, students can check how many times they repeated a certain lexical item, and how many derivatives they produced for this item.

LRNetM can further help students detect other categories of lexical cohesion such as antonyms, synonyms and superordinates with their hyponyms. Figure 21 illustrates how LRNetM is applied to an Arab L2 learner's essay to analyse its lexical cohesion. It is clear from the network view of the essay that it is rich of lexical cohesive devices. The teacher can ask students to highlight the longest lexical networks in the essay. For example, the view of the essay below shows that the longest network is (no. 39) which contains 28 instances of repetition. That is because the main theme of the essay is about sports. The prime word in this network is the lexical item *sport*. In the same network, the student can see that he uses *sporty*

which constitutes with other items in the same network derived repetition. *Sports* also could function as a superordinate term that includes within it the lexical items: *boxing* in (network no. 7) and *football* in (network no. 18). The network view can also reveal synonyms for *sports*. Lexical networks (no. 2 and 20) include, for example, *activity*, *activities*, *game*, and *games*. The second longest network is number 10 which consists of *dangerous* and *dangers*. These two lexical items establish both simple and derived repetitions. Besides, they can link in with other lexical networks. For example, they are semantically related with networks 22, 35 and 36. These networks contain synonyms of *dangerous* (e.g., *hazard*, *hazardous*; *risks*; *serious*).

Network N	Word	Freq.	total number of tokens in the lexical network	Tokens of Simple Rep.	Tokens of derived Rep.
1	ACCIDENTS	3	3	2	0
2	ACTIVITIES	2			
	ACTIVITY	1	3	2	0
3	ADVERTISEMENT	2	2	1	0
4	AWARE	2	2	1	0
5	BANNED	3	3	2	0
7	BOXERS	1			
	BOXING	1	2	0	2
8	CHALLENGING	3	3	2	0
9	CONTRIBUTE	2	2	1	0
10	DANGEROUS	13			
	DANGERS	2	15	13	2
	DISABILITIES	1			
11	DISABLED	1	2	0	2
12	EFFECTS	2	2	1	0
13	ENCOURAGED	1			
	DISCOURAGE	1	2	0	2
14	EQUIPMENT	2	2	1	0
15	FACE (v)	2	2	1	0
16	FAMILY	2	2	1	0
	FINANCIAL	1			
17	FINANCIALLY	1	2	0	2
18	FOOTBALL	3	3	2	0
19	FORBID	2	2	1	0
20	GAME	1			
	GAMES	1	2	1	0
	GOVERNMENT	2			
21	GOVERNMENTS	2	4	3	0
	HAZARD	1			
22	HAZARDOUS	1	2	0	2
23	HEALTH	3	3	2	0
24	HUMAN	3	3	2	0
25	INJURIES	5	5	4	0
26	LIFE	4			
	LIVES (n)	1	5	4	0
27	MEDICAL	3	3	2	0
28	MONEY	4	4	3	0
29	PEOPLE	13	13	12	0
30	PERMANENT	2	2	1	0
31	PERSON	2	2	1	0
	PLAY (V)	1			
32	PLAYING	4	6	4	2
	PLAYERS	1			
	PROTECT	2			
33	PROTECTIVE	3	5	3	2
34	REDUCE	2	2	1	0
35	RISK	2			
	RISKS	2	4	3	0
36	SERIOUS	2	2	1	0
37	SITUATION	1			
	SITUATIONS	1	2	1	0
	SOCIETIES	1			
38	SOCIETY	1	2	1	0
	SPORT	5			
39	SPORTS	22	28	26	2
	SPORTY	1			
40	SYSTEM	2	2	1	0
41	TALENTED	2	2	1	0
	THREAT	1			
42	THREATENING	1	2	0	2
	UNSAFE	1			
43	SAFER	1	3	0	3
	SAVED	1			
	WEAR	1	3	2	0
44	WEARING	2			
45	YOUNG	2	2	1	0
			169	112	23

Figure 21 A Lexical Repetition network of an Arab L2 essay

After this preliminary analysis of lexical cohesion at the surface level, students are instructed to pick up any lexical items from any lexical network in the essay, and then consider these words as cores. They can then run concordance lines of the selected item and examine how it fits in with other items in building up the text. These concordance lines also allow the learner to observe repeated patterns and meanings, and thus help them to be aware of collocation and semantic prosody. In this step, the teacher needs to pay the students' attention to semantic prosody. For example, students are asked to look at the surrounding environment of the lexical item under analysis and evaluate how it is used in their essays – whether it is positive or negative.

However, the students' essays are short and cannot reveal a large number of textual patterns that could show the typical semantic prosody of the item in question. Thus, the teacher can in this case use a tool such as *WebCorp* (2006) that easily provides concordance samples which represent authentic examples of cohesion. The students therefore can check the lexical item in *WebCorp* first, and try to describe its semantic prosody. Then, they can move from *WebCorp* to their individual essays and assess if they used the same semantic prosody of the lexical item under analysis. The students can repeat this exercise with other lexical items in their essays. They can further see similarities between lexical items in their essays and identify semantic prosodies that are shared by different lexical items in the same essay. This activity will help learners recognise how lexical cohesion links lexical and textual features of a text.

8.5 Directions for future research

The limitations of the present study have been already addressed in the previous chapters. Therefore, in this section, I will focus on how to take up these limitations as directions for

future research. First, I indicated in Chapter 4 that the corpus used in this study was small. This was because no corpus of argumentative essays by Arab L2 learners was found, and I therefore collected a corpus of Arab L2 argumentative essays from scratch. This collection was challenging in terms of time and availability of participants who were willing to take part in my study. Despite the small size of the Arab L2 corpus by corpus linguistic standards, it was suitable for the stated aims of the present investigation. Nevertheless, the obtained results should be interpreted with care and not be generalised. A small corpus can be non-representative for the patterns that have been identified. It might also provide data which is insufficient to show linguistic regularities. This has been proved to be true when I analysed semantic prosody of selected lexical items in Chapter 7. For example, I only found few cases that indicate how semantic prosody of the lexical item in question is violated in a number of essays of the Arab L2 writers. This might create the problem of ‘local densities’, as Moon (1998: 68) describes. This case of linguistic irregularity could be avoided by using a larger corpus. Sinclair (2004: 189) explains that “the main virtue of being large in corpus is that the underlying regularities have a better chance of showing through the superficial variations” and we can look at phraseology in a systematic way.

Therefore, extending the size of the corpus could lead to a richer analysis of lexical cohesion and could further reveal more differences between native and non-native English writing of argumentative essays. Also, there is a need in future to collect more electronic data by Arabic speakers of English to compensate for the absence of this group in the area of learner corpora. As Granger et al. (2015: 2) point out “these collections are usually quite large and are collected from a great number of learners, they are arguably more representative than smaller data samples involving a limited number of learners”.

Chapter 4 further demonstrated that the Arab L2 corpus and the LOCNESS corpus were different in respect to the situational variables. The first variable is the time under which the writing task has been conducted. For example, the Arab L2 writers completed their writing task at home without time restriction whereas the British writers wrote their essays within a restricted time framework. Thus, a question remains whether the time limitation imposed on these essays encouraged the use of repetition as a cohesive strategy, as Reynolds (2002) observes. Another difference is the proficiency level of the two groups. Proficiency was not important in the present study because the analysis did not intend to relate the differences in the use of lexical cohesion to this measure. However, these variables such as time and proficiency are recommended to be controlled in any future study to ensure that they do not have an impact on the obtained results.

As for the LRNetM model I suggested in this thesis, it is important to acknowledge that LRNetM provides only a starting point for the analysis of lexical cohesion from a corpus-linguistic perspective. Also, this model has not been applied before and it therefore leaves room for further improvement, particularly in the type of cohesive categories it analyses, and in the way how it works to quantify cohesive categories. LRNetM has proven to be rewarding in quantifying simple and derived repetition, at least for the purpose of this study. That is because as Thornbury (2010) observes, lexical repetition is amenable to analysis using corpus tools. However, LRNetM was less efficient in analysing textual categories of lexical cohesion such as signalling nouns, which were analysed using a text analysis. LRNetM also did not incorporate into the wordlist other lexical cohesive networks that include more complex relations such as synonyms and hyponyms. Each one of these categories has its own characteristics and whose exploration can yield interesting findings and help provide the

complete picture of lexical cohesion in text. Nevertheless, such categories also appear in the wordlist and can be grouped into networks of semantically related words. But this classification is not as straightforward as simple and derived repetition. Rather, these categories can be highlighted in the wordlist and then transferred with their word frequency into a table to be counted using my procedure of counting. The analysis of lexical cohesion applying LRNetM can also be extended by studying collocates of the lexical items in the lexical repetition network either in individual texts or in the complete corpus as one discourse community. For example, if a simple repetition network in an essay from ALEC or LOCNESS contains these lexical items: *access...access...accessibility...accessible*, the LRNetM model can be applied to study collocates of these items. This analysis will help identify collocational cohesion in a text and add a further depth to the analysis of lexical cohesion.

With reference to the method of counting that LRNetM adopted, it was only suggestive but it was applied consistently and met the purpose of the present study which was identifying frequencies in NS and NNS writing. As demonstrated in Chapter 5, the counting method used the frequency data as a starting point to assist in counting the tokens of simple and derived repetition in each lexical network in the text under analysis. For simple repetition, LRNetM assumed one of the members in the simple repetition network as a prime word, and therefore it excluded it from the counting. For a derived repetition network, occurrences of simple repetition were deducted from the length of the lexical network in question. Other researchers can test this counting method on other data set in order to ensure that this method can be replicated. In addition, researchers can suggest software that can group a text into lexical cohesive relations, which can be visualised as networks of lexical items that relate with each

other cohesively. *GraphColl* (Brezina et al. 2015), is an example of such software. However, it is only designed to build collocational networks. It would be interesting if this software is developed to further include lexical cohesive relations.

In Chapter 7, the number of lexical items used to analyse lexical cohesion in terms of semantic preference and prosody was small. This small number was due to the small size of both corpora used in this thesis. Also the fact that the analysis of semantic preference and prosody is subtle and needs more space has made this small number sufficient within the context of the present study. Future studies, therefore, could investigate many more items. It is hoped that results from such research may help further investigate this neglected aspect of language use and awareness, and that such data may inform theoretical discussion of lexical cohesion much more than has been the case to date.

There are other important points in this study that should not be considered as limitations but they are under emphasised because they lie outside the scope of this research. These points thus could also be developed in future research because they are directly related to the findings of this thesis. The first one is the analysis of lexical cohesion through ‘thematic progression’. I have already looked briefly at this in Chapter 6. Thematic progression proved to be a key measure in analysing lexical cohesion by showing how arguments are developed to build up a coherent text. However, more research is needed to address this theory in a systematic way and examine how Arab L2 learners are balancing between old and new information in the text to establish lexical cohesion.

Another area of future research is analysing signalling nouns that take the form of ‘lexical couplets’ or ‘strings’. My research showed that this linguistic feature is distinctive to the Arab

L2 writers compared to the British native writers. Nevertheless, my corpus is small, and this linguistic phenomenon needs to be further examined in a larger corpus to verify whether it is significant or not in the English writing of Arab writers. Additionally, my analysis of lexical cohesion through signalling nouns revealed how British native and Arab non-native writers structure their texts differently in terms of the problem-solution pattern. Therefore, it would be useful to carry out a more systematic analysis of this pattern and its contribution to the cohesion of text.

8.6 Conclusion

The present study compared frequency and function of lexical cohesion in argumentative writing of Arab speakers of English and English native speakers. This comparison provided a comprehensive picture of the difficulties that Arab L2 writers have in their use of lexical cohesion. The frequency analysis confirmed what Carter (1987: 95) observed: that counting cohesive devices will not explain the reason why some written texts are considered to be better organised than others. Granger (1996: 17) advises that in studying learner language, it is crucial to combine a quantitative and a qualitative approach, comparing frequency and semantic/syntactic use. Reviewing the available literature in the present study has shown that there is an existing gap in the scholarship regarding how lexical cohesion functions at the levels of paradigmatic and syntagmatic in learner writing. Therefore, the contribution of the present study was to address this gap by supplementing the frequency analysis of lexical cohesion with analysing the function of lexical cohesion. I have shown that text analysis and corpus linguistics can be integrated for a unique insight into the explanation of the function of lexical cohesion. Both approaches confirmed that cohesion is a means to an end, not the end itself. Hartnett (1986: 152) concluded that the success of a writer's composition depends on

much more than successful use of any cohesive devices. Nevertheless, these features and their different uses can help us to describe how readers understand and writers control the textual structure that expresses rhetorical development in written discourse. The integration of text-based analysis and corpus linguistic approaches further pinpoint that learning individual words and their meanings does not suffice to achieve great fluency in second language writing.

The results of this thesis suggest that we indeed should change the way we conceptualise lexical cohesion, and start approaching it as a textual concept rather than a lexical one. These results are especially important for EFL/ESL teachers who do not seem to think of lexical cohesion in this way. The new insight into the nature of lexical cohesion will certainly help them become more aware of the different characteristics of lexical cohesion and reach a better understanding of this important type of linguistic features in writing. However, these findings seem to be only a first step towards recognising how lexical cohesion functions in text. It is left for future research to explore this field further. There is a definite need for more precise methods to analyse cohesive devices in discourse and corpus linguistics more meaningfully.

Appendices

Appendix 1: Essay prompts and writing instructions template

1. Essay Prompts

Please read the essay prompts for the five topics below and select one of them to write about. Read the writing instructions below.

Topic 1: Computer vs. brain

As computers are becoming smarter and people, of different ages, start relying more and more on them in different areas of life, the power of the human brain will surely diminish. To what extent do you agree or disagree with this opinion?

Topic 2: Immigration

For different reasons, a recent increase in the number of immigrants to many developed countries, including the UK, has taken place. People argue that immigration should be stopped or at least controlled. However, others argue that if immigration is ceased, it might have economic and cultural consequences. To what extent do you agree or disagree with this opinion?

Topic 3: Traffic

Cheap petrol in some countries increases the number of cars in the road and this leads to many traffic problems. Government should impose higher fees for petrol to reduce such number of cars and use the money to improve public transportation. To what extent do you agree or disagree with this opinion?

Topic 4: Free education

The fact that, in some countries, education is free or at least not expensive, particularly in the case of higher degrees, limits the quality of education. To what extent do you agree or disagree with this opinion?

Topic 5: Dangerous sports

Many people play sports that are dangerous such as motor racing, and they might get injured or even dead. Dangerous sports should, therefore, be banned. To what extent do you agree or disagree with this opinion?

2. Writing Instructions

- a) Write an argumentative (discussion) essay of approximately 500 – 1000 words. Try to present supporting evidence and reasoning for your argument wherever possible, and use, as much as possible, formal and academic language as if you were writing for an IELTS exam.
- b) Please provide examples from your own knowledge and experience to support your argument.
- c) You could do background reading surfing the internet but please DO NOT copy and paste from articles on the internet. Use your own style to write in an argumentative way without looking at a specific article or trying to summarise it.
- d) You could consult a dictionary in case you need to find a specific word.
- e) Try to type your essay using a computer (if possible).
- f) Once you complete your essay, please fill in a learner profile.

Appendix 2: A learner profile template

Please answer the following questions which are useful to build up a learner profile for each participant. This information will be used for research purposes.

Part 1: Participant Background information

1. Country of origin/nationality.....
 2. Age 18-25 26-35 36-45 45 and above
 3. Gender M F
 4. Field of study.....
 5. Degree of study: MA MSc MPhil PhD
Other (please specify).....
 6. Educational institution (If possible).....
 7. IELTS score or any other English equivalent test including ‘a pre-sessional course’
(if applicable).....
- (If you feel this question is personal or you have not sit a language test before, please provide the English language entry requirement specified by your school to be offered a place to study your current course.....)
8. For how long have you been studying English?
 - Years of English at school..... months/years
 - Years of English at university till now..... months/years

9. For how long have you been in the UK?

1 month - 6 months 6months – 2years above 2 years

10. For how long have you been in an English speaking country other than UK (please add all the staying periods together, if any)?

11. Do you use British English or American English when you write?

British English American English

Part2: Text profile

Please tick the appropriate box after you finish writing your essay.

1. I used a dictionary to check up some words YES NO

If YES, please provide the type of the dictionary:

- Bilingual dictionary
- English monolingual dictionary
- Other(s) :.....

2. I used the web YES NO

3. Other references that you used while writing your essay (please specify).....

4. I did not use any references. I just relied on my background knowledge to write.

YES NO

Thank you for your participation

Appendix 3: Participant Consent Form Template

To participant,

Researcher details

I am a PhD student on Applied Linguistics & English Teaching programme at the University of Nottingham, UK.

The purpose of the PhD study

This research study is a corpus-linguistic study, which attempts to collect as much as written essays as possible to analyse specific linguistic features with the help of corpus-linguistic software. More specifically, I mainly aim to compile a small corpus of argumentative essays by postgraduate Arab students in different disciplines over the UK. Besides, this study intends to compare the linguistic features of the writing of the above mentioned students with British native students' writing to shed light on features of non-nativeness in learner writing.

The participants' ability to write in an argumentative mode will be studied from a linguistic perspective only. The topics will be about general life issues and do not entail specific knowledge from students. The collected essays will represent an electronic corpus of Arab students. It is possible that this corpus is used by other researchers in future for research purposes, particularly in the field of corpus linguistics.

Each participant will be instructed to fill in a learner profile after performing a writing task either at home or at university based on students' preference and time. The results will be presented in my PhD thesis and none will be published online without any consent from you.

Declaration of Consent

It is a university requirement that all respondents give their formal consent to take part in any research. For this reason could you please sign and date the declaration below after you give your consent to the following points:

Consent to the use of a metadata questionnaire and performing a writing task

YES NO I confirm that the purpose of the study has been explained and that I have understood it.

YES NO I have had the opportunity to ask questions and they have been successfully answered.

YES NO I understand that my participation in this study is voluntary and that I am free to withdraw from the study at any time, without giving a reason and without consequence.

YES NO I understand that all data are anonymous and that there will not be any connection between the personal information provided and the data.

YES NO I understand that there are no known risks or hazards associated with participating in this study.

YES NO I consent to my data being transcribed and wish to be referred to anonymously.

YES NO I consent that my writing will be kept in an 'electronic corpus' for research use.

YES NO I confirm that I have read and understood the above information and that I agree to participate in this study.

Should require any further information, please do email me using the following email address:

██

Participant's Name and Signature: _____

Researcher's Signature: _____

Date: _____

Appendix 4: ALEC and LOCNESS British A-level students' essays sub-corpus CD-ROM

Appendix 5: A complete essay from ALEC showing the intense use of commas

Human brain is the most powerful thing existed // and the reason for the computer to be available because of the human brain. // It is true that computer can run faster, operate pre-defined functions faster and then execute these functions upon the defined parameters. // The computer can not act on its own, // the human brain have to define and determine what function to be executed. // The article about agreeing or disagreeing to certain levels that the power of the computer and new technologies raising will affect out brains and decrease it functions.

To some certain aspects I agree the computer run faster than the human brain when working with numbers and long formulas, and processing data faster when it comes to precision of transmitting and receiving data. // As of this topic I am sure and 100% that the computer will not overcome the human brain functionality. // Next to any computer or machine there must be a human to check the operation of that computer. // Computer can not work on their on // they need some kind of lines to be defined, some tasks which is programmed by human then embedded into the machine to do the operation. // In autopilot the technology still faces a big deal in letting the computer do everything from taking off to landing, the two very complicated tasks are the taking off and the landing of the plane. // I am not talking about small plane, but about an airbus plane with passengers on board // these have big deal to be worked by human. // We need the computer in our daily life // and we feel like we can not live a day without using our mobile phones, tablets, or laptop. // This because we feel like connected to each other all the time just by using social media network. // This has effects on our communication skills and leads to lack decrease these skills // and comparing to computers it can not feel anything it has no emotions or whatsoever. The functions of the computer we trying to pushed forward and embedded these into all of our lifestyle. // The use of “Internet of Things” becoming widely open to all and sharing and working our ways to be connected all the time by this huge project, // Internet of Things works on providing all the connected nodes together and deliver tasks and operation upon need. // This smart technology invented by human brains and people working together and developed such network. // The use of computer and its addiction can be good when spending a lot of time on computers, // it

faster our research and looking deep into many different topics apart of the topic intended for research, // some said using computers in the right way as for studies and carrying homework will develop the ability in thinking more and be adoptable to new information comes along as we serve the internet looking for particular task // and in result we went through many tasks and articles.

Even spending too much time on strategical games help on pushing the limits of how we think and be smarter, // many computer games has different strategies and new ideas that a gamer can walkthrough and finished on new time breaker each time. // The computers can affects some people who are not capable of taking this powerful technology and worked it in the right way in developing the skills in researching and thinking wider and outer the box, whoever uses the technology in the wrong way as of addiction like in gaming that has no strategical ideas and has no reasoning for playing or just spending too much time on watching TV media, Drama and gossiping shows. // It will affects the brain functionality // and it will surely diminish it's power. // These technologies were invented to help us on working and carrying a simplified ways to finish the work in better quality and in good time management. // It is all about time and quality, // people needs computers to help on their work and computers needs human brain to develop and integrate new functions based on the operation raised. // It is us who think for computers // and it is us who embed and program the computer to run some certain operations. // So in conclusion the computer can diminish human brain when it intended for wrong ways other than the thinking beyond computer and try to make computer does the work for us.

Appendix 6: An analysis of lexical cohesion in a complete essay from ALEC applying Halliday & Hasan's (1976) model

T-unit No.	No. of ties	Cohesive item	Type	Distance	Presupposed item
2	1	brains	Simple Rep.	0	brains
3	8	exercises (n) complex problems X 2 reliance computers solving solve	Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep.	0 1 1 1 1 1 1	exercising (v) complex problems rely computers solve solve
4	5	computers provide solve complex problems	Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep.	0 2 0 0 0	computers provide solve complex problems
5	9	easy access reliant reliable computers provide efficient answers brains	Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep.	0 0 1 1 0 0 3 3 2	easy access reliance reliance computers provide efficient answers brains
6	3	various complex problems	Simple Rep. Simple Rep. Simple Rep.	4 1 1	various complex problems
7	8	solve problems Power brain relied computer access internet	Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep.	2 0 4 1 1 1 1 1	solve problems power brains reliant computers access internet

8	8	Accessibility information easier identify reliable solutions problems solve	Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep.	0 3 2 0 0 0 1 0	access information easy identify relied solve problems solve
9	6	computer internet power brain solve problems	Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep.	1 1 1 1 0 0	computer internet power brain solve problems
10	1	computers	Simple Rep.	0	Computer
12	3	mathematical power brain	Simple Rep. Simple Rep. Simple Rep.	8 2 2	mathematical power brain
13	3	rely calculate numbers	Simple Rep. Simple Rep. Simple Rep.	5 9 9	reliable calculators numerical
14	7	rely X 2 brains X 2 computers X 2 powerful	Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep.	0 1 3 1	rely brains computers power
15	3	simple mathematical calculation	Simple Rep. Simple Rep. Simple Rep.	2 2 1	simple mathematical calculate
16	4	number quicker multiples calculate	Simple Rep. Simple Rep. Simple Rep. Simple Rep.	2 14 2 0	numbers quick multiples calculation
17	1	younger	Simple Rep.	0	Younger
18	3	rely computers calculators	Simple Rep. Simple Rep. Simple Rep.	3 3 1	rely computers calculate

19	3	access computers internet	Simple Rep. Simple Rep. Simple Rep.	10 0 9	accessibility computers internet
20	7	rely brain solutions problems numerical complex simple	Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep.	1 5	rely brain solve problems numbers complex simple
21	6	rely computers exercise brain X 2 powerful	Simple Rep. Simple Rep. Simple Rep. Simple Rep. Simple Rep.	0	rely computers exercises brain powerful

**Appendix 7: An analysis of lexical cohesion in a complete essay from ALEC applying
Hoey's (1991b) repetition matrix CD-ROM**

References

- Ädle, A. (2006). *Metadiscourse in L1 and L2 English*. Amsterdam: Benjamins.
- Ädle, A. (2008). Involvement features in writing: Do time and interaction trump register awareness? *Language and Computers* 66: 35–53.
- Adorjan, M. (2013). Explorations in lexical repetition analysis: The outcomes of manual vs. computer automated research methods. *WoPaLP* 7: 1-28.
- Aktas, R. N & V. Cortes. (2008). Shell nouns as cohesive devices in published and ESL student writing. *Journal of English for Academic Purposes* 7 (1): 3-14.
- Al-Jubouri, A. (1984). The role of repetition in Arabic argumentative discourse. In: Swales J, Mustafa H. (eds.) *English for Specific Purposes in the Arab World*. Birmingham: Language Service Unit, Aston University, 99-117.
- Allen, H.B. (1970). A monotonous dialogue. In F. Larudee (ed.) *TESL in the Middle East*. Cairo: American University in Cairo Press, 93-102.
- Alonso, S., & McCabe, A. (2003). Improving text flow in ESL learner compositions. *The Internet TESL Journal* 9 (2): 1-10. Available at <<http://iteslj.org/Articles/Alonso-ImprovingFlow.html>>
- Altenberg, B. & Granger, S. (2002). *Lexis in Contrast: Corpus-based Approaches*. Amsterdam/Philadelphia: John Benjamins.
- Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, 22 (2): 173-195.
- Altenberg, B. & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (ed.) *Learner English on computer* London and New York: Longman, 80- 93.
- Álvarez de Mon y Rego, I. (2006). A contrastive study of encapsulation and prospection in written scientific text. In J. Flowerdew & M. Gotti (eds.) *Studies in Specialized Discourse* Bern: Peter Lang, 21-40.
- Aziz, Y. Y. (1988). Theme-Rheme Organization and Paragraph Structure in Standard Arabic. *Word* 39 (2): 117-128.
- Baker, P. (2004). Querying Keywords: Questions of difference, frequency and sense in keywords analysis, *Journal of English Linguistics* 32: 346-59.
- Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

- Baker, P. (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Baker, M. (2011). *In Other Words: A Coursebook on Translation* (2nd ed.). London: Routledge.
- Barlow, M. (2005). Computer-based analyses of learner language. In R. Ellis & G. Barkhuizen (eds.) *Analysing Learner Language*. Oxford, UK: Oxford University Press, 335-357.
- Beaugrande, R. de & Dressler, W. (1981). *Introduction to Text Linguistics*. London: Longman.
- Bell, J. (2005). *Doing Your Research Project: A Guide for First-Time Researchers in Education and Social Science*, 4th edition. London: Open University Press.
- Benitez-Castro M-A. (2015). Coming to Grips with Shell-nounhood: A Critical Review of Insights into the Meaning, Function and Form of Shell-noun Phrases, *Australian Journal of Linguistics* 35 (2): 168-194.
- Benitez-Castro M-A & Thompson, P. (2015). Shell-nounhood in academic discourse: A critical state-of-the-art review. *International Journal of Corpus Linguistics* 20 (3): 378-404.
- Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. & R. Quirk. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.
- Biber, D., S. Conrad & R. Reppen. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Biber, D. (2006). *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.
- Bolton, K., Nelson, G. & Hung, J. (2002). A corpus-based study of connectors in student writing: Research from the International Corpus of English in Hong Kong (ICE-HK). *International Journal of Corpus Linguistics* 7 (2): 165-182.
- Brezina, V, McEnery, T. & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20 (2): 139-173.
- Brezina, V. and M. Meyerhoff. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics* 19 (1): 1-28.

- Brown, J. D. (2011). Quantitative research in second language studies. In E. Hinkel (ed.) *Handbook of research in second language teaching and learning*. New York, NY: Routledge, 190-206.
- Brown, G. & Yule, G. (1983). *Discourse Analysis*. Cambridge: Cambridge University Press.
- Bublitz, W. (2011). Cohesion and Coherence. In J. Zienkowski., J-O Östman., & J. Verschueren (eds.) *Discursive pragmatics* Amsterdam/Philadelphia: John Benjamins Publishing Company, 37-49.
- Caldwell, C. (2009). *Lexical Vagueness in Student Writing: Are Shell Nouns the Problem?* Saarbrücken: VDM Verlag Dr. Muller.
- Carrell, P. L. (1982). Cohesion is not coherence. *TESOL Quarterly* 16: 479-488.
- Carter, R. (1987). *Vocabulary: Applied Linguistic Perspectives*. London: Allen & Unwin.
- Carter, R. and McCarthy, M. (1997). *Exploring Spoken English*. Cambridge: Cambridge University Press.
- Chafe, W. (1986). Evidentiality in English conversation and academic writing. In: W. Chafe & J. Nichols (eds.) *Evidentiality: The linguistic coding of epistemology*, Norwood, Newjersey: Ablex, 261-272.
- Cheng, W. (2009). Describing the extended meanings of lexical cohesion in a corpus of SARS spoken discourse. In J. Flowerdew & M. Mahlberg (eds.) *Lexical Cohesion and Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 65-83.
- Collier, A. (1994). *A System for automating concordance line selection*. Available at <http://rdues.bcu.ac.uk/publ/AJC_94_02.pdf>.
- Connor, U. (1984). A study of cohesion and coherence in English as a second language students' writing. *International Journal of Human Communication* 17 (3): 301-316.
- Connor, U. (1987). Argumentative Patterns in Student Essays: Cross Cultural Differences. In C., Ulla and Kaplan, Robert B. (eds.) *Writing across Languages: Analysis of L2 Text*. Reading, Mass.: Addison Wesley, 57-71.
- Connor, U. (1996). *Contrastive Rhetoric: Cross-Cultural Aspects of Second Language Writing*. Cambridge: Cambridge University Press.
- Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics*, 6th edition. Wiley-Blackwell.
- Daneš, F. (1974). Functions of Sentence Perspective and the Organization of the Text. In: F. Dand, (ed.) *Papers on Functional Sentence Perspective*. Prague: Academia, 106-128.

- De Cock, S. (2003). *Recurrent sequences of words in native speaker and advanced learner spoken and written English: a corpus-driven approach*. Unpublished PhD thesis. Louvain-la-Neuve: Université catholique de Louvain.
- Dickins, J., Hervey., S & Higgins, I. (2002). *Thinking Arabic Translation*. Routledge Taylor and Francis group. London and New York.
- Dodd, B. (2000). Introduction: the relevance of corpora in German studies. In B. Dodd (ed.) *Working with German Corpora*. Birmingham: University of Birmingham Press, 1-39.
- Dornyei, Z. (2007). *Research Methods in Applied Linguistics*. New York: Oxford University Press.
- El-Gazzar, N. (2006). *Lexical cohesive devices in Arab students' academic writing: Implications for teaching vocabulary*, MA Thesis, American University of Sharjah.
- Emmott, C. (1989). *Reading between the lines: building a comprehensive model of participant reference in real narratives*. Ph.D. Thesis, University of Birmingham.
- Enkvist, N. E. (1978). Coherence, pseudo-coherence, and non-coherence. In I. O. Ostman (ed.) *Cohesion and semantics*. Abo: Abo Akademi Foundation, 109-128.
- Evtushenko and Butuzova. (2014). Punctuation of Cohesive Devices: Theory and Practice. *Procedia - Social and Behavioral Sciences* 154: 391-394.
- Fakaude, E. and Vargs, L. (1992). Cohesion and text creation. *Language Learning Journal*. No. 5 Hertford shire: Stephen Austin & Sons Ltd.
- Firth, J.R. (1957). Modes of meaning. In J. R. Firth (eds.) *Papers in linguistics 1934-1951*. Oxford: Oxford University Press, 190-215.
- Fligelstone, S. (1992). Developing a scheme for annotating text to show anaphoric relations. In Leitner, G. (ed.) *New directions in English language corpora. Methodology, results, software developments*. Berlin: Mouton de Gruyter, 153-70.
- Fletcher, W. (2003/8). *Phrases in English (PIE)*. Available at <www.usna.edu/LangStudy/PIE/> (accessed 7 April 2016).
- Flowerdew, J. (2002). Introduction: Approaches to the analysis of academic discourse in English. In J. Flowerdew (ed.) *Academic discourse*. Harlow: Longman, 1-17.
- Flowerdew, J. (2003). Signalling nouns in discourse. *English for Specific Purposes* 22 (4): 329-346.

- Flowerdew, J. (2006). Use of signalling noun in a learner corpus. *Lexical Cohesion and Corpus Linguistics (Special Issue), International Journal of Corpus Linguistics* 11 (3): 227-247.
- Flowerdew, J., & Mahlberg, M. (2009). *Lexical Cohesion and Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Flowerdew, J. (2009). Use of signalling nouns in a learner corpus. In J. Flowerdew & M. Mahlberg (eds.) *Lexical Cohesion and Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 85-102.
- Flowerdew, J. (2010). Use of signalling nouns across L1 and L2 writer corpora. *International Journal of Corpus Linguistics* 15 (1): 36-55.
- Flowerdew, J., and Forest, R. W. (2015). *Signalling Nouns in English: A Corpus-based Discourse Approach*. Cambridge: Cambridge University Press.
- Flowerdew, L. (2004). The argument for using English specialised corpora to understand academic and professional language. In U. Connor and T. Upton (eds.) *Discourse in the Professions*. Amsterdam: John Benjamins, 11-33.
- Flowerdew, L. (2008). *Corpus-based Analyses of the Problem-Solution Pattern: A Phraseological Approach*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Forutan A. & Nasiri, S. R. (2011). Signalling nouns in English and Persian: A contrastive study. *Theory and Practice in Language Studies* 1 (10): 1273-1283.
- Francis, G. (1986). *Anaphoric Nouns* (Discourse Analysis Monograph 11). English Language Research, Birmingham: University of Birmingham.
- Francis, G. (1988). The teaching of techniques of lexical cohesion in an ESL setting. In V. Bickley (ed.) *Language in a Bilingual or Multilingual Setting* Hong Kong: Institute of Language in Education, 325-338.
- Francis, G. (1993). A corpus-driven approach to grammar – principles, methods and examples. In M. Baker, G. Francis, & E. Tognini-Bonelli (eds.) *Text and Technology In Honour of John Sinclair*. Amsterdam Benjamins, 137-156.
- Francis, G. (1994). Labelling discourse: an aspect of nominal-group lexical cohesion. In M. Coulthard (ed.) *Advances in Written Text Analysis*. London: Routledge, 83-101.
- Gablasova, D., Brezina, V. & McEnery, T. (2017). Exploring learner language through corpora: comparing and interpreting corpus frequency information. *Language Learning* 67: S1: 130-154.

- Garside R., G. Leech & A. McEnery (eds.). (1997). *Corpus Annotation. Linguistic Information from Computer Text Corpora*. Longman: London.
- Garside, R., and Smith, N. (1997). *CLAWS part-of-speech tagger for English*. Available at <<http://ucrel.lancs.ac.uk/claws/trial.html>> (last accessed May 2015).
- Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (eds.) *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies*. Lund: Lund University Press, 37-51.
- Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes* 15 (1): 17-21.
- Granger, S. (1998a). The computer learner corpus: a versatile new source of data for SLA research. In S. Granger (ed.) *Learner English on Computer*. London & New York: Longman, 3-18.
- Granger, S. (ed.). (1998b). *Learner English on Computer*. Addison Wesley/New York: Longman.
- Granger, S. (2002). A bird's eye view of learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (eds.) *Computer Learner Corpora, Second Language Acquisition, and Foreign Language*. Amsterdam/Philadelphia: John Benjamins, 3-13.
- Granger, S. (2004). Computer learner corpus research: Current status and future prospects. In U. Connor & T. Upton (eds.) *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi, 123-145.
- Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1 (1): 7-24.
- Granger, S., Gilquin, G., & Meunier, F. (2015). *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press.
- Granger, S. & Paquot, M. (2009). Lexical verbs in academic discourse: A corpus-driven study of learner use. In M. Charles, D. Pecorari, & S. Hunston (eds.) *Academic writing. At the interface of corpus and discourse*. London: Continuum, 193-214.
- Gilquin, G. and M. Paquot. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction* 1: 41-61
- Granger, S. & Tribble C. (1998). Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning. In G. Sylviane (ed.) *Learner English on Computer*. London & New York: Addison Wesley Longman, 199-209.

- Gray, B. & V, Cortes. (2011). Perception vs. evidence: an analysis of this and these in academic prose. *Journal of English for Academic Purposes* 30: 31-43.
- Greaves, C. (2009). *ConcGram 1.0*. Amsterdam: John Benjamins.
- Gutwinski, B. (1976). *Cohesion in Literary Texts*. The Hague: Mouton.
- Halliday, M., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
- Halliday, M. (1985). *An Introduction to Functional Grammar*. London: Arnold.
- Halliday, M. (1994). *An Introduction to Functional Grammar*, 2nd ed. London: Edward Arnold.
- Halliday, M. & Hasan, R. (1989). *Language, Context, and Text: Aspects of Language in a Socialsemiotic Perspective*. Oxford: Oxford University Press.
- Halliday, M. and R. Hasan. (1985). *Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective*. Oxford: Oxford University Press.
- Halliday, M. A. K. & Matthiessen, C. (2004). *An Introduction to Functional Grammar* (3rd edition). London: Arnold.
- Hamdan, A.S. (1988). *Coherence and Cohesion in Texts Written in English by Jordanian University Students*. Unpublished Ph.D. Thesis, Manchester University, England.
- Hasan, R. (1984). Coherence and cohesive harmony. In J. flood (ed.) *Understanding reading comprehension*. Delaware: International Reading Association, 181-219.
- Hasselgård, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics* 4: 237- 259.
- Hasselgård, H. (2012). Facts, ideas, questions, problems, and issues in advanced learners' English. *Nordic Journal of English Studies* 11 (1): 22-54.
- Hatim, B & I. Mason. (1997). *The Translator as Communicator*. London: Routledge.
- Hartnett, C.G. (1986). Static and dynamic cohesion: signals of thinking in writing. In ed. B. Couture (ed.) *Functional approaches to writing: Reserch Perspectives*. London: Frances Pinter, 142-153.
- Hawes, T. (2015). Thematic progression in the writing of students and professionals. *Ampersand* 2: 93-100.
- Hinkel, E. (2001). Matters of cohesion in L2 academic texts. *Applied Language Learning* 12 (2): 111-132.

- Hinkel, E. (2002). *Second Language Writers' Text: Linguistic and Rhetorical Features*. Mahwah: Lawrence Erlbaum.
- Hirsch, E. D. (1975). Stylistics and Synonymity. *Critical Inquiry*, 1 (3): 559-579.
- Hoey, M. (1983). *On the Surface of Discourse*. London: George Allen and Unwin.
- Hoey, M. (1991a). Another perspective on coherence and cohesive harmony. In Ventola, E. (ed.) *Functional and Systemic Linguistics: Approaches and Uses*. The Hague: Mouton de Gruyter, 385-413.
- Hoey, M. (1991b). *Patterns of Lexis in Text*. Oxford: Oxford University Press.
- Hoey, M. (1993). *Data, Description, Discourse: Papers on the English Language in Honour of John McH Sinclair*.
- Hoey, M. (2001). *Textual Interaction: An Introduction to Written Discourse Analysis*. Routledge: London & New York.
- Hoey, M. (2005). *Lexical Priming. A new Theory of Words and Language*. London: Routledge.
- Hoffmann, S., Evert, S., Smith, N., Lee, D. Y. W. & Berglund Prytz, Y. (2008). *Corpus Linguistics with BNCWeb – A Practical Guide*. Frankfurt am Main: Peter Lang.
- Huang, L. F. (2011). *Discourse markers in spoken English: A corpus study of native speakers and Chinese non-native speakers*, Unpublished PhD dissertation, University of Birmingham.
- Hunston, S. & Francis, G. (2000). *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam/Philadelphia: John Benjamins.
- Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Hunston, S. (2007). Semantic prosody revisited. *International Journal of Corpus Linguistics* 12 (2): 249-268.
- Hunt, K. W. (1965). *Grammatical Structures Written at Three Grade Levels*. Research Report No. 3. Champaign, IL: National Council of Teachers of English.
- Hunt, K. W. (1970). Do sentences in the second language grow like those in the first? *TESOL Quarterly* 4 (3): 195-202.
- Hyland, K. (2005). Stance and engagement: a model of interaction in academic discourse. *Discourse Studies* 7 (2): 173-192.
- Hyland, K. & J. Milton. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing* 6 (2): 183-205.

- Ivanič, R. (1991). Nouns in search of a context: a study of nouns with both open- and closed-system characteristics. *International Review of Applied Linguistics in Language Teaching* 29: 93-114.
- Johnstone, B. (1983). Arabic lexical couplets and the evolution of synonymy. *General linguistics* 23 (1): 51-61.
- Johnstone, B. (ed.). (1994). *Repetition in Discourse: Interdisciplinary Perspectives*. New Jersey: Alex publishing.
- Johnstone, B. (1991). *Repetition in Arabic Discourse: Paradigms, Syntagms and the Ecology of Language*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher* 33 (7): 14-26.
- Johnson, R. B., Onwuegbuzie, A., & Turner, L. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research* 1: 112-133.
- Kaderavek, J. (2015). *Language Disorders in Children: Fundamental Concepts of Assessment and Intervention*. (2nd ed.). Upper Saddle River, N.J., Pearson Education Inc.
- Kai, J. (2008). Lexical Cohesion Patterns in NS and NNS Dissertation Abstracts in Applied Linguistics: A Comparative Study. *The Linguistic Journal* 3 (3): 132-159.
- Kaplan, R. B. (1966). Cultural thought patterns in intercultural education. *Language Learning* 16: 1-2.
- Kaplan, R. B. (1987). Cultural thought patterns revisited. In U. Connor & R. B. Kaplan (eds.) *Writing across languages: Analysis of L2 text*. Reading, MA: Addison-Wesley, 9-21.
- Károly, K. (2002). *Lexical Repetition in Text*. Frankfurt am Main: Peter Lang.
- Khalil A. (1989). A study of cohesion and coherence in Arab EFL College students' writing. *System* 17 (3): 359-71.
- Kohn, K. (1986). The analysis of transfer. In E. Kellerman and M. Sharwood Smith (eds.) *Crosslinguistic Influence in Second Language Acquisition*. New York: Pergamon Press, 21-34.
- Kunz, A. (2010). *Variation in English and German Nominal Coreference: A Study of Political Essays*. Peter Lang.
- Kubota, R. (1999). Japanese culture constructed by discourses: Implications for applied linguistics research and English language teaching. *TESOL Quarterly* 33: 9-35.

- Kubota, R., & Lehner, A. (2004). Toward critical contrastive rhetoric. *Journal of Second Language Writing* 13: 7-27.
- Lautamatti, L. (1978). Observations on the development of the topic of simplified discourse. In U. Connor & R. B. Kaplan (eds.) *Writing across Languages: Analysis of L2 Text*. Addison-Wesley, Reading, MA, 87-114.
- Leech, G. (1998). Learner corpora: what they are and what can be done with them. In S. Granger (ed.) *Learner English on Computer*. London & New York: Longman, xiv-xx.
- Lieber, P. E. (1979). *Cohesion in ESL students' expository writing: A descriptive study*. New York University. PhD thesis.
- Liebman, J. D. (1992). Toward a new contrastive rhetoric: Differences between Arabic and Japanese rhetorical instruction. *Journal of Second Language Writing* 1: 141-165.
- López Samaniego, A. (2011). *La categorización de entidades del discurso en la escritura profesional* [The categorisation of discourse entities in professional writing] Unpublished PhD thesis, University of Barcelona, Barcelona.
- Lorenz, G. (1999). Learning to cohere: causal links in native vs. non-native argumentative writing. In W. Bublitz, U. Lenk and E. Ventola (eds.) *Coherence in Spoken and Written Discourse. How to create it and how to describe it*. Amsterdam & Philadelphia: John Benjamins Publishing Company, 55-75.
- Mahlberg, M. (2005). *English General Nouns: A Corpus Theoretical Approach*. Amsterdam/Philadelphia: John Benjamins.
- Mahlberg, M. (2009). Lexical cohesion. Corpus linguistic theory and its application in English language teaching. In J. Flowerdew & M. Mahlberg (eds.) *Lexical Cohesion and Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 103-122.
- Martin, A. V. (1989). *Bridging vocabulary: An essential component of ESL proficiency*. Paper presented at the 23rd Annual TESOL Convention, San Antonio, Texas.
- Martin, J. R. (1992). *English Text: System and Structure*. Amsterdam: John Benjamins.
- Mastropierro, L. & Mahlberg, M. (2017). Key words and translated cohesion in Lovecraft's *At the Mountains of Madness* and one of its Italian translations. *English Text Construction* 10 (1): 78-105.
- Mauranen, A. (1993). *Cultural Differences in Academic Rhetoric*. Frankfurt: Peter Lang.
- McArthur, T. (1996). *The Oxford Companion to the English Language*. Oxford: Oxford University Press.

- McCarthy, M. (1991). *Discourse Analysis for Language Teachers*. Cambridge: Cambridge University Press.
- McCarthy, M. & Handford, M. (2004). Invisible to us: a preliminary corpus-based study of spoken business English. In U. Connor and T.A. Upton (eds.) *Discourse in the Professions: Perspectives from corpus linguistics*. Amsterdam: John Benjamins, 167-201.
- McCarthy, M., & Carter, R. (2014). *Language as Discourse: Perspectives for Language Teaching*. London/New York: Routledge.
- McEnery, T. (2006). *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. Abington, UK: Routledge.
- McEnery, T., & Kifle, A. (2002). Epistemic modality in argumentative essays of second-language writers. In J. Flowerdew (ed.) *Academic Discourse*. London: Pearson Education Limited, 182-195.
- McEnery, T., & Wilson, A. (2001). *Corpus Linguistics* (2nd ed.). Edinburgh: Edinburgh University Press.
- McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.
- McGee, I. (2008). Traversing the lexical cohesion minefield. *ELT Journal* 63 (3): 212-220.
- Meunier, F., De Cock, S., Gilquin, G. & Magali, P. (2011). (eds.). *A Taste for Corpora. In Honour of Sylvaine Granger*. Amsterdam/Philadelphia: Benjamins Publishing Company.
- Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Oxford: Clarendon Press.
- Morgan, J. L. and Sellner, M. B. (1980). Discourse and linguistic theory. In R. J. Spiro, B. C. Bertranm, and W. F. Brewer (eds.) *Theoretical issues in reading comprehension*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Morley, J. (1999). Sticky business: a case study of cohesion in the language of politics in the Economist. In L. Lombardo, L. Haarman, J. Morley & C. Taylor. *Massed Medias. Linguistic Tools for Interpreting Media Discourse*. Milano: LED, 18-83.
- Morley, J. (2009). Lexical cohesion and rhetorical structure. In J. Flowerdew & M. Mahlberg (eds.) *Lexical Cohesion and Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 5-22.
- Morley, J. & Partington, A. (2009). A few frequently asked questions about semantic – or evaluative – prosody. *International Journal of Corpus Linguistics* 14 (2): 139-158.

- Mohamed-Sayidina, A. (2010). Transfer of L1 cohesive devices and transition words into L2 academic texts: The case of Arab students. *RELC Journal* 41 (3): 253-266.
- Mohamed, A. H, Omer M. R. (2000). Texture and culture: cohesion as a marker of rhetorical organization in Arabic and English narrative texts. *RELC Journal* 31 (2): 45-75.
- Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* 17 (1): 21-48.
- Morris, J., and Hirst G. (2006). The subjectivity of lexical cohesion in text. In J. C. Chanahan, Y. Qu, & J. Wiebe (eds.) *Computing attitude and affect in text: Theory and Applications*. Springer, Dodrecht, The Netherlands, 41-47.
- Mukherjee, J., & Rohrbach, J. (2006). Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research. In B. Kettemann & G. Marko (eds.) *Planning, gluing and painting corpora: Inside the applied corpus linguist's workshop*. Frankfurt: Peter Lang, 205-232.
- Murphy, L. (2003). *Semantic Relations and the Lexicon: Antonymy, Synonymy, and Other Paradigms*. Cambridge University Press, Cambridge, UK.
- Murphy, T. (2001). The Emergence of Texture: An Analysis of the Functions of the Nominal Demonstratives in an English Interlanguage Corpus. *Learning Language and Technology* 5 (3): 152-173.
- Neff, J., F. Ballesteros, E. Dafouz, F. Martínez and J.P. Rica. (2004). The expression of writer stance in native and non-native argumentative texts. In R. Facchinetti and F. Palmer (eds.) *English Modality in Perspective*. Frankfurt am Main: Peter Lang, 141-161.
- Nesi, H. & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics* 11 (3): 283-304.
- Nystrand, M. (1986). *The Structure of Written Communication: Studies in Reciprocity between Writers and Readers*. Academic Press, Orlando, FL.
- Oakes, M. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Ong, W. J. (2003). *Orality and Literacy: The Technologizing of the Word*. London and New York: Routledge.
- Ostler, SE. (1987). English in parallels: A comparison of English and Arabic prose. In U. Connor & R.B. Kaplan (eds.) *Writing across languages: Analysis of L2 text* Reading, MA: Addison-Wesley, 169-185.
- Ouaouicha, D. (1986). Contrastive rhetoric and the structure of learner-produced argumentative texts in Arabic and English. *Dissertation Abstracts International* 47 (9), 3339A.

- Palmer, A., Sporleder, C. & Linlin, L. (2012). So to Speak: A Computational and Empirical Investigation of Lexical Cohesion of Non-Literal and Literal Expressions in Text. *Discourse: A journal of linguistics, psycholinguistics and computational linguistics* 11: 1-24.
- Parsons, G. (1991). Cohesion Coherence: Scientific Texts. In Ventola E. (ed.) *Functional and systemic linguistics, approaches and uses*. Berlin: Houton de Gruyter, 415-429.
- Partington, A. (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.
- Partington, A. (2004a). Utterly content in each other's company: Semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9 (1): 131-56.
- Partington, A. (2004b). Corpora and Discourse: a most congruous beast. In A. Partington, J. Morley & L. Haarman (eds.) *Corpora and Discourse*. Bern: Peter Lang, 11-20.
- Petch-Tyson, S. (2000). Demonstrative expressions in argumentative discourse: A computer corpus-based comparison of non-native and native English. In S. Botley & T. McEnery (eds.) *Corpus-based and Computational Approaches to Discourse Anaphora*. Amsterdam, Netherlands: John Benjamins, 43-64.
- Pueyo, I. G. and Val S. (1996). The construction of technicality in the field of plastics: a functional approach towards teaching technical technology. *English for specific purposes* 15 (4): 251-278.
- Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
- Renouf, A. (1986). Lexical resolution. In W. Meijs (ed.) *Corpus Linguistics and Beyond: Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora*. Amsterdam: Rodopi, 121-131.
- Reynolds, D. W. (1995). Repetition in nonnative speaker writing: More than quantity. *Studies in Second Language Acquisition* 17 (2): 185-209.
- Reynolds, D. W. (2002). Language in the Balance: Lexical Repetition as a Function of Topic, Cultural Background, and Writing Development. *Language Learning* 51 (3): 437 - 476.
- Rieschild, V. R. (2006). Emphatic repetition in spoken Arabic. In I. Mushin (ed.) *Proceedings of the 2004 Conference of the Australian Linguistic Society*, 1-23. Available at <<http://hdl.handle.net/2123/274>> (Accessed July 2016).
- Rimmon-Kenan, S. (1980). The paradoxical status of repetition. *Poetics Today* 1 (4): 151-159.

- Sa'Addedin, M. A. (1989). Text Development and Arabic-English Negative Interference. *Applied Linguistics* 10 (1): 36-51.
- Salkie, R. (1995). *Text and Discourse Analysis*. London/New York: Routledge.
- Sardinha, T. (2000). Semantic prosodies in English and Portuguese: A contrastive study. *Cuadernos de Filologi'a Inglesa* 9 (1): 93-110.
- Saussure, F. de. (1983). *Course in General Linguistics*; edited by Charles Bally and Albert Sechehaye, in collaboration with Albert Reidlinger; translated from the French by Wade Baskin. London: Owen, 1960.
- Scarcella, R. (1984). Cohesion in the writing development of native and non-native English speakers. (Doctoral Dissertation, University of Southern California). *Dissertation Abstracts International* 45, 1386A.
- Schiffrin, D. (1987). *Discourse Markers*. Cambridge: Cambridge University Press.
- Scott, M. (1997). PC Analysis of Key Words and Key Key Words. *System* 25 (1): 1-13.
- Scott, M. (2012). *WordSmith Tools (Version 6)* [Computer software]. Stroud: Lexical Analysis Software.
- Scott, M. & Tribble, C. (2006). *Textual Patterns: Keyword and Corpus Analysis in Language Education*. Amsterdam: Benjamins.
- Schmid, H. J. (2000). *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Berlin, New York, NY: Mouton de Gruyter.
- Schmitt, N and Zimmerman, C. B. (2002). Derivative Word Forms: What Do Learners Know? *TESOL Quarterly* 36 (2): 145-171.
- Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learning driven data. In S. Granger, J. Hung & S. Petch-Tyson (eds.) *Computer learner corpora, second language acquisition and foreign language teaching*. Philadelphia: John Benjamins, 213-234.
- Shakir, A. (1991). Coherence in EFL Student-Written Texts: Two Perspectives. *Foreign Language Annals* 24 (5): 399- 411.
- Shalaby, N. Yahya, N. & M. El-Komi. (2009). Analysis of Lexical Errors in Saudi College Students' Compositions. *Ayn, Journal of the Saudi Association of Languages and Translation* 2 (3): 65-93.

- Silva, T. (1990). Second language composition instruction: developments, issues, and directions in ESL. In B. Kroll (ed.) *Second language writing research: Research insights for the classroom*. Cambridge: Cambridge University Press, 11-23.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, J. M. (1993). Written discourse structure. In J. M. Sinclair, M. Hoey & G. Fox (eds.) *Techniques of description: spoken and written discourse, a festschrift for Malcolm Coulthard* London, New York, NY: Routledge, 6-31.
- Sinclair, J. M. (1996). The search for units of meaning. *Textus* 9 (1): 75-106.
- Sinclair, J. M. (1998). The lexical item. In E. Weigand (ed.) *Contrastive Lexical Semantics*. Amsterdam/Philadelphia: John Benjamins, 1-24.
- Sinclair, J. M. (2004). *Trust the Text. Language, Corpus and Discourse*. London: Routledge.
- Sinclair, J., & Renouf, A. (1988). A lexical syllabus for language learning. In R. Carter & M. McCarthy (eds.) *Vocabulary and language teaching*. Harlow: Longman, 140-158.
- Spencer, A. (1991). *Morphological Theory. An Introduction to Word Structure in Generative Grammar*. Cambridge, Mass: Blackwell.
- Stewart, D. (2010). *Semantic Prosody: A Critical Evaluation*. London/New York: Routledge.
- Stotsky, S. (1981). The vocabulary of essay writing: can it be taught? *College Composition and Communication* 32: 317-326.
- Stotsky, S. (1983). Types of Lexical Cohesion in Expository Writing: Implications for Developing the Vocabulary of Academic Discourse. *College Composition and Communication* 34 (4): 430-446.
- Strauss, S. & Fieze, P. (2014). *Discourse Analysis: Putting our Worlds into Words*. New York: Routledge.
- Stubbs, M. (1983). *Discourse Analysis: The Sociolinguistic Analysis of Natural Language*. Oxford: Basil Blackwell.
- Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative study. *Functions of Language* 2 (1): 23-55.
- Stubbs, M. (2001a). Computer-assisted text and corpus analysis: lexical cohesion and communicative competence. In D. Schiffrin, D. Tannen & H. E. Hamilton (eds.) *The Handbook of discourse Analysis*. Malden Massachusetts, Oxford: Blackwell, 304-320.
- Stubbs, M. (2001b). *Words and Phrases. Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

- Tadros, A. (1985). *Prediction in Text* (Discourse Analysis Monograph 10). Birmingham: English Language Research, University of Birmingham.
- Tadros, A. (1994). Predictive categories in expository text. In M. Coulthard (ed.) *Advances in written text analysis*. New York: Routledge, 69-82.
- Tannen, D. (1989). *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse*. Cambridge England. New York: Cambridge University Press.
- Tanskanen, S. K. (2006). *Collaborating towards Coherence: Lexical Cohesion in English Discourse*. Amsterdam, Philadelphia: John Benjamins.
- Teich, E., & Fankhauser, P. (2004). WordNet for Lexical Cohesion Analysis. In Soijika, P., Pala, K., Smrz, P., Fellbaum, C., & Vossen, P. (eds.) *Proceedings of the 2nd International WordNet Conference*. Masaryk University, Brno, Czech Republic, 326-331.
- Teich, E., & Fankhauser, P. (2005). Exploring lexical patterns in text: lexical cohesion analysis with WordNet. *Working papers of the SFB 632, Interdisciplinary studies on information structure* (ISIS). University of Bibliothek Frankfurt am Main.
- Teubert, W. (2001). Corpus Linguistics and Lexicography. *International Journal of Corpus Linguistics* 6: 125-53.
- Thibault, P. J. (1991). Grammar, Technocracy, and the Noun: Technocratic Values and Cognitive Linguistics. In Ventola, E. (ed.) *Functional and Systematic Linguistics: Approaches and Uses*. Berlin and New York: Mouton de Gruyter, 281-306.
- Thornbury, S. (2010). What can a corpus tell us about discourse? In A. O’Keeffe & M. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 270-287.
- Ting, F. (2003). An investigation of cohesive errors in the writing of PRC Tertiary EFL students. *STETS Language and Communication Review* 2 (2).
- Tirkkonen-Condit, S. (1984). Towards a description of argumentative text structure. In H. Ringbom & M. Rissanen (eds.) *Proceedings from the Second Nordic Conference for English Studies*. Abo: Meddelanden fran Stiftelsens for Abo Akademi Forskningsinstitut nr. 92, 221-233.
- Tognini Bonelli E. (1996). *Corpus Theory and Practice*. Birmingham: TWC.
- Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam/Philadelphia: Benjamins.

- Trabasso, T., Secco, T., & Van den Broek, P. (1984). Causal cohesion and story coherence. In H. Mandl, H., Stein, N. L., & Trabasso, T. (eds.) *Learning and comprehension of text*. Hillsdale, N.J: Lawrence Erlbaum, 83-112.
- Twardzisz, P. (1997). *Zero Derivation in English: A Cognitive Grammar Approach*. Lublin: Wydawnictwo.
- Ventola, E. (1996). Packing and unpacking of information in academic texts. In Ventola, E., Mauranen, A. (eds.) *Academic Writing: Intercultural and Textual Issues*. John Benjamins, Amsterdam, 153-194.
- Voss, E. (2008). Review: Mahlberg, M. 2005. English General Nouns: A Corpus Theoretical Approach. *Corpora* 3 (2): 227-230.
- Waelateh, B. (2016). The comparison of pronunciation in English with Thai, Malay and Arabic by minority Malay Muslims in Thailand. In Vopava, J. (ed.) *Proceedings of the 6th MAC 2016*. Prague: MAC Prague consulting Ltd., 187-202.
- Wang, H., & Wang, T. (2005). A contrastive study on the semantic prosody of CAUSE. *Modern Foreign Language* 28 (3): 297-307.
- Wang, L. (2007). Theme and rheme in the thematic organization of text: Implications for teaching academic writing. *Asian EFL Journal* 9 (1): 164-176.
- Watt, R. J. C. (1999-2009). *Concordance (Version 3.3)* [Computer software].
- Wei, N.X. (2006). A corpus-based contrastive study of semantic prosodies in learner English. *Foreign Language Research* 132: 50-54.
- Werlich, E. (1976). *A Text Grammar of English*. Quelle & Meyer, Heidelberg.
- Widdowson, H. G. (1978). *Teaching Language as Communication*. Oxford: Oxford University Press.
- Widdowson, H.G. (1997). EIL, ESL, EFL: Global issues and local interests. *World Englishes* 16 (1): 135-146.
- Willis, D. (1990). *The Lexical Syllabus: A New Approach to Language Teaching*. London: HarperCollins.
- Winter, E.O. (1974). *Replacement as a function of repetition: A study of some of its principal features in the clause relations of contemporary English*. Unpublished Ph.D. dissertation, University of London.
- Winter, E. O. (1977). A clause-relational approach to English texts: a study of some predictive lexical items in written discourse. *Instructional Science* 6: 1-92.

- Winter, E. O. (1979). Replacement as a fundamental function of the sentence in context. *Forum linguisticum* 4 (2): 95-133.
- Winter, E. O. (1992). The notion of unspecific versus specific as one way of analysing the information of a fund-raising letter. In W. C Mann & S. A Thompson (eds.) *Discourse Description: diverse linguistic analyses of a fund-raising text* Amsterdam, Philadelphia, PA: John Benjamins, 131-170.
- Williams G. (1998). Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles. *International Journal of Corpus Linguistics* 3 (1): 151- 171.
- Williams, M. P. (1989). *A Comparison of the Textual Structures of Arabic and English Written Texts: A Study in the Comparative Orality of Arabic*. PhD thesis. University of Leeds.
- Witte S., & Faigley, L. (1981). Coherence, cohesion and writing quality. *College Composition and Communication* 32 (2): 189-204.
- Whitsitt, S. (2005). A critique of the concept of semantic prosody. *International Journal of Corpus Linguistics* 10 (3): 283-305.
- Wynne, M. (2005). *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books.
- Wynne, M. (2005). Archiving, Distribution and Preservation. In M. Wynne, (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, 71-78. Available at <<http://ota.ox.ac.uk/documents/creating/dlc/> [Accessed June 2017].
- Xiao, Z., & McEnery, A. (2006). Collocation, semantic prosody and near synonymy: A cross-linguistic perspective. *Applied Linguistics* 27 (1): 103-129.
- Yamasaki, N. (2008). Collocations and colligations associated with discourse functions of unspecific anaphoric nouns. *International Journal of Corpus Linguistics* 13 (1): 75-98.
- Zamel, V. (1997). Toward a model of transculturation. *TESOL Quarterly* 31: 341-352.
- Zhang, W. (2008). *In search of English as foreign language (EFL) teachers' knowledge of vocabulary instruction*. Unpublished doctoral dissertation, Georgia State University.
- Zhang, W. (2009). Semantic prosody and ESL/EFL vocabulary pedagogy. *TESL Canada Journal* 26 (2): 1-12.