Lexical Repetition in Academic Discourse
A Computer-aided Study of the Text Organizing Role of Repetition

PhD Dissertation Summary
PhD Programme in Language Pedagogy
Doctoral School of Education
Eötvös Loránd University, Budapest, 2015

Candidate: Adorján Mária
Supervisor: Károly Krisztina, PhD, habil.
Committee:

Chair: Medgyes Péter, DSc.
Internal Opponent: Pohl Uwe, PhD.
External Opponent: Prószéky Gábor, DSc.
Secretary: Tartsayné Németh Nóra, PhD.
Members: Pődör Dóra, PhD
        Jasmina Sasdowska, PhD
        Tankó Gyula, PhD
1. Introduction

Due to the various functions and diverse attitudes to lexical repetition in discourse, it is an aspect of cohesion which creates difficulty for raters when assessing L2 academic written texts. Most probably, English teachers would agree with Connor (1984), who found that repeated words in students’ writings were both a sign of limited vocabulary and of poor text structuring. Lexical choice, however, depends on a wide range of factors, for example individual differences among students, such as language level or cultural background, or differing requirements of various subjects (Reynolds, 2001). Previous research suggests for instance, that science articles require more reiterations than popular articles because scientific terminology cannot be replaced by synonyms (Myers, 1991). Therefore, lexical repetition in academic writing is a highly relevant field of study, which cannot be analyzed in isolation, without considering its contextual dimensions.

Lexical repetition is studied both in text linguistics (discourse analysis) and corpus linguistics, which two terms cover related but not identical approaches to the study of text. In discourse analysis, first the various cohesive devices are categorized according to semantic relatedness criteria, and a theoretical framework is built, which is later tested on a small number of texts. Lexical repetition patterns are analyzed quantitatively and manually (e.g., the researcher counts how many times certain categories are represented in the text) as well as qualitatively (e.g., conclusions are drawn observing the types, location and lexical environment of repeated words). The main problem with this type of analysis is that only a small number of texts can be observed; therefore, the data gained do not permit generalizations.

The other approach to lexical cohesion analysis is offered by corpus linguistics, which allows for automated analysis of large linguistic data. A disadvantage of this method is that individual differences within texts in a corpus cannot be observed qualitatively. Reviewing best practice in text-based research, Graesser, McNamara, and Louwerse (2011) maintain that the recent shift in discourse analysis is characterized by moving from “theoretical generalizations based on empirical evidence observing a small corpus to large-scale corpus-based studies” (p. 37), and the results have changed “from deep, detailed, structured representations of a small sample of texts to comparatively shallow, approximate, statistical representations of large text corpora” (p. 37).

Current computer-aided lexical cohesion analysis frameworks built for large-scale assessment fail to take into account where repetitions occur in text and what role their patterns play in organizing discourse. This study intends to fill this gap, by drawing on Hoey’s (1991)
theory-based analytical tool devised for the study of the text-organizing role of lexical repetition, and its refined and extended version, Károly’s (2002) lexical repetition model, which was found to be capable of predicting teachers’ perceptions of argumentative essay quality with regard to its content and structure.

2 Research aims

In the first two stages of this research the aim is to test the applicability of Károly’s (2002) LRA (lexical repetition analysis) model to the academic summary and the compare/contrast genres by manual and partially computer-aided analyzes, in order to test whether her analytical tool can predict teachers’ judgement regarding discourse quality in the case of these two genres, too. In the third stage of the research, the aim is to alter the tool to enable large-scale analysis of EFL student corpora in order to propose a more complex, computer-aided analytical instrument that may be used to directly assess discourse cohesion through the study of lexical repetition.

The main questions guiding this research are therefore the following:

(1) Is Károly’s (2002) theory-based lexical repetition model, a revised version of Hoey’s (1991) repetition model, applicable to the study of summaries and compare/contrast essays written by Hungarian EFL university students?

(2) What modifications are necessary in Károly’s (2002) theory-based lexical repetition model to be applicable to large-scale EFL learner corpora?

In the last part of the dissertation, by employing the theoretical, empirical and methodological results gained from the corpora, several new analytical steps are proposed first, and these are arranged in a modular format. Next, in order to better align the newly proposed computer-aided method to its manual version, a parallel is drawn between the analytical processes of the new LRA model and an existing socio-cognitive framework.

This study is multidisciplinary in nature, aiming to contribute to the fields of (a) applied linguistics, more closely to discourse analysis and corpus linguistics; (b) language pedagogy, especially to the teaching and evaluating EFL academic writing; and (c) computer science, to enhance educational software development. The newly proposed model may be useful for teachers assessing discourse cohesion, for instance, by highlighting the lexical net created by semantic relations among sentences in text. Alternatively, it can be used as a self-study aid by students of academic writing.
3 Theoretical Framework

Coherence and cohesion in text have become widely researched areas within the field of discourse analysis, with special focus on lexical cohesion due to its significant discourse function (e.g., Halliday, 1985; Halliday & Hasan, 1976; Hoey, 1991; Reynolds, 1995, 2001; Tyler, 1994, 1995). Given that there is disagreement in the literature about coherence and cohesion, for the purposes of this research project, the following two similar sets of definitions will be used for cohesion and coherence:

(1) Coherence is “the understanding that the reader derives from the text” (Crossley & McNamara, 2010, p. 984).

(2) “Cohesion refers to the presence or absence of explicit clues in the text that allow the reader to make connections between the ideas in the text” by (Crossley & McNamara, 2010, p. 984). and;

(1) “[C]oherence is the quality that makes a text conform to a consistent world picture and is therefore summarizable and interpretable.” (Enkvist, 1990, p. 14)

(2) “Cohesion is the term for overt links on the textual surface []” (Enkvist, 1990, p. 14)

Lexical cohesion was defined by Hoey (1991) as “the dominant mode of creating texture” because it is “the only type of cohesion that regularly forms multiple relationships” in text (p.10). He called these relationships lexical repetition, using repetition in a broader sense, referring not only to reiterations but also various other forms of semantic relatedness, such as synonyms, antonyms, meronyms, as well as other paraphrases. Based on Halliday and Hasan’s (1976) empirical investigation of cohesive ties in various text types, Hoey concluded that lexical cohesion accounted for at least forty percent of the total cohesion devices (1991). In a more recent corpus linguistic study Teich and Fankhauser (2004) claimed that nearly fifty percent of cohesive ties consist of lexical cohesion devices, thus making lexical cohesion the most pronounced contributor to semantic coherence.

3.1 Hoey’s Repetition Model

Several manual and computer-aided methods exist to analyze lexical features in text. Of particular interest are frameworks capable of not only identifying and classifying linguistic elements but also providing information on their patterns and roles in structuring text. Hoey’s (1991) theory-based analytical tool designed for the study of lexical repetition is the first one of the frameworks devised to offer a manual analytical method for studying the text-structuring role of lexical repetition. This framework explores the semantic network (links, bonds and the
**lexical net** of text and distinguishes between central and marginal sentences by finding lexical repetition patterns. With this method it is possible to summarize certain types of discourse.

Hoey (1991) claimed that “lexical items form links when they enter into semantic relationships” (p. 91). These links, however, are only realized between two sentences, not inside a sentence. Therefore, if two words are repeated within one sentence, these will not be analyzed. The reason for this, according to Hoey, is that intra-sentential repetitions do not play a role in structuring discourse, even if they have an important function, e.g., emphasis (*They laughed and laughed and laughed uncontrollably.*). Hoey differentiated his concept of link from Hasan’s cohesive tie in two aspects. Firstly, his categories were greatly different from those of Hasan’s. Secondly, he emphasized that links have no directionality. Hoey’s important claim is that certain sentences play a more central role in organizing discourse than others. Sentences sharing three or more links are significant for the organization of discourse because they form bonds, a higher level connection. Marginal sentences, with fewer than three links, do not contribute essentially to the topic, therefore if omitted, do not disrupt the flow of the discourse.

Bonded sentences lead to nets, which ultimately organize text, in a manner similar to Hasan’s (1984) identity and similarity chains. Hoey found that bonded sentences are central to text, as they are the core bearers of information resembling the concept of macropropositions described by van Dijk and Kintsch (1983). Hoey’s main claim that links created via lexical repetition may form bonds which subsequently create significant sentences, was later reaffirmed by Reynolds (1995) and by Teich and Fankhauser (2004, 2005).

### 3.2 Károly’s revised and extended LRA model

Hoey’s (1991) comprehensive analytical model was later revised by Károly (2002) who made significant changes in the categories (see Table 1). Károly also extended the model by introducing several new analytical steps to reveal the organizing function of lexical repetition in texts. Hers was the first application of Hoey’s model in a Hungarian higher education setting. Károly’s (2002) research results showed that her theory-driven ‘objective’ analytical tool not only offered a descriptive function, but with her analytical procedures the tool was capable of predicting the ‘intuitive’ assessment of teachers judging argumentative essay quality with regard to its content and structure. Given that in holistic scoring teachers assign more weight to content and organization than to any other components (Freedman, 1979), and given that these two components comprise the concepts of cohesion and coherence, which are responsible for textuality (Halliday & Hasan, 1976), it is of little surprise that lexical repetition analysis (LRA) can detect the difference between valued and poor writing.
Table 1. Types of lexical relations in Károly’s taxonomy with examples (examples based on Károly, 2002, p. 104, and the two corpora of this research)

<table>
<thead>
<tr>
<th>Categories of lexical relations</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Same unit repetition</td>
<td></td>
</tr>
<tr>
<td>1. repetition</td>
<td>simple</td>
</tr>
<tr>
<td></td>
<td>derived</td>
</tr>
<tr>
<td></td>
<td>writers – writers</td>
</tr>
<tr>
<td></td>
<td>writers – write</td>
</tr>
<tr>
<td>II. Different unit repetition</td>
<td></td>
</tr>
<tr>
<td>2. synonymy</td>
<td>simple</td>
</tr>
<tr>
<td></td>
<td>derived</td>
</tr>
<tr>
<td></td>
<td>to exercise – (after) working out</td>
</tr>
<tr>
<td></td>
<td>built – construction</td>
</tr>
<tr>
<td>3. opposites</td>
<td>simple</td>
</tr>
<tr>
<td></td>
<td>derived</td>
</tr>
<tr>
<td></td>
<td>small – major</td>
</tr>
<tr>
<td></td>
<td>hatred – like</td>
</tr>
<tr>
<td>4. hyponymy</td>
<td>simple</td>
</tr>
<tr>
<td></td>
<td>derived</td>
</tr>
<tr>
<td></td>
<td>languages – English</td>
</tr>
<tr>
<td>5. meronymy</td>
<td>simple</td>
</tr>
<tr>
<td></td>
<td>hands – fingers</td>
</tr>
<tr>
<td>6. instantial relations</td>
<td>simple</td>
</tr>
<tr>
<td></td>
<td>derived</td>
</tr>
<tr>
<td></td>
<td>manager – O’Leary</td>
</tr>
</tbody>
</table>

The results of Károly’s (2002) analysis proved that the texts, which had previously been judged by experienced university instructors, differed significantly in both repetition types and patterns. Post-tests conducted with another group of teachers confirmed these findings indicating that the analytical measures devised can reliably predict how teachers perceive essay quality, and the results may be generalized for a wider sample.

Hoey (1991) found predictable lexical repetition patterns in news articles, whereas Károly (2002) studied the academic argumentative essay genre in this respect. Due to the fact that summary and compare/contrast essay are the two most commonly used genres across the disciplines at universities (Bridgeman & Carlson, 1983; Moore & Morton, 1999), these integrative (reading into writing) tasks also deserve such thorough investigation. Therefore, research needs to be extended to the predictive power of Károly’s model in genres most likely faced by EFL students across universities in Hungary, however, at the moment of writing no such study exists.

3.3 Existing large-scale essay assessment

Large-scale essay assessment applications (such as E-rater, Criterion, Intelligent Essay Assessor) have been in use for decades to test essay writing skills in the EFL context. However, these applications were developed by major testing agencies and are not available to the public. These essay scoring programs measure cohesion and coherence quantitatively, using statistical methods and natural language processing (NLP) techniques. Their methods focus on cohesion and coherence on the local level, mainly by comparing adjacent sentences semantically, on the assumption that words in adjacent sentences form semantic chains which can be identified for
topic progression. This can be called the “lexical chain principle”. However, these chains are linear in nature, and indicate cohesion on the local level, whereas discourse also shows global cohesion. If text is considered to create (or to be created by) lexical nets, it is necessary to observe semantic links between all the sentences in the text, even if they are located far from each other; in other words, the “lexical chain principle” need to be switched for the “lexical net principle”.

4. Research design and methodology

In order to test the applicability of Károly’s model on other academic genres, two small corpora of thirty-five academic summaries and eight compare/contrast essays were collected from English major BA students at Eötvös Loránd University. The lexical repetition patterns within the corpora were analyzed manually in the case of the summaries, and partially with a concordance program in the case of the compare/contrast essays.

The study used a mixed methods design including both qualitative and quantitative methods, as suggested by Creswell (2007); and a sequential mixed design paradigm, as described by Tashakkori and Teddlie (2003). The rationale behind using qualitative and quantitative methods is that, according to Tyler (1995) and Károly (2002), quantitative analysis alone cannot inform research about the real role of lexical repetition in organizing discourse. In the first stage of this study, the model was applied to the summary genre. The second stage utilized results gained from the first stage and continued to test the model on compare/contrast essays. In this second stage, a concordance analyzer\(^1\) was used at the initial step of the analysis.

In the third stage, the theoretical, empirical and methodological results of the previous stages formed the basis of the design of the new, semi-automated analytical tool.

In the first task the students were instructed to summarize Ryanair’s history and business model, highlighting its strengths and weaknesses. The constructs in this summary task were (1) company history, (2a) strengths of the business model, (2b) weaknesses of the business model. The input text was an Internet source, a Wikipedia entry (an electronic genre) describing the history of a company. The features of the input text were assessed partly by observation, partly by utilizing the Coh-Metrix (McNamara, Louwerse, Cai, & Graesser, 2005) textual analysis program: Text Easibility Assessor measures. It analyzed five features of the input text: Narrativity, Syntactic Simplicity, Word Concreteness, Referential Cohesion, and Deep

---

\(^1\) a program which displays every instance of a specified word with its immediate preceding and following context
Cohesion. The further assessment methods are described in Table 2. Next, the lexical repetition analysis was carried out manually by two coders following Károly’s (2002) analytical steps.

<table>
<thead>
<tr>
<th>Type of assessment</th>
<th>Method of assessment</th>
<th>Aim</th>
</tr>
</thead>
<tbody>
<tr>
<td>class teacher evaluated the summaries</td>
<td>holistic, one grade multiplied by 4 to gain a 20point scale.</td>
<td>evaluation of content, organization, language and grammar/mechanics</td>
</tr>
<tr>
<td>content similarity evaluation with text similarity program</td>
<td>(a) testing a text similarity online application and (b) checking summaries for direct copying</td>
<td>text originality, how similar text is to the source text, eliminating too similar texts from the corpus to leave only texts in the corpus suitable for lexical repetition analysis</td>
</tr>
<tr>
<td>two other teachers who did not read the original text: judgement of discourse control</td>
<td>analytical, checking overt cohesive devices</td>
<td>assessing textual coherence and cohesion, observing texts as self-contained pieces of discourse</td>
</tr>
<tr>
<td>automated textual analysis</td>
<td>Coh-Metrix cohesion</td>
<td>using an objective tool to assess cohesion</td>
</tr>
</tbody>
</table>

Table 2. Overview of assessments, methods and their aims in Stage 1

The compare/contrast essay corpus consisted of eight texts from the same pool of students, whose task was to write an academic essay on an applied linguistics topic of approximately 600 words. The source texts (books and journal articles) were selected by the students. The same two coders analyzed the corpus, again following Károly’a (2002) analytical steps.

5. Results

5.1 Empirical results

This study revealed that Károly’s model is suitable for the analysis of the text-organizing role of lexical repetition in both genres. It also revealed that the structures of high-rated summaries and compare/contrast essays were different from low-rated ones: in both genres the main ideas were organized into sentences with special discourse functions, such as the theses and topic sentences.

5.2 The newly proposed LRA model: the three modules of the analysis

The new analytical phases considered necessary for a computer-aided lexical repetition analysis (LRA) were identified during Stages 1 and 2 of this research. These phases follow a
strict sequence. An overview of the proposed model for computer-assisted LRA is shown in Figure 1.

<table>
<thead>
<tr>
<th>1 Preparation of the corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Plagiarism check</td>
</tr>
<tr>
<td>• Identifying sentence boundaries</td>
</tr>
<tr>
<td>• Annotating the text for title, paragraph boundaries</td>
</tr>
<tr>
<td>• L2 special corpora treatment / Error treatment</td>
</tr>
<tr>
<td>• POS tagging of sentences</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2 Establishing links</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Establishing 'key terms'</td>
</tr>
<tr>
<td>• Identifying lexical units (manual or computerized)</td>
</tr>
<tr>
<td>• Locating same unit repetition links</td>
</tr>
<tr>
<td>• Locating different unit repetition links</td>
</tr>
<tr>
<td>• Semantic relations analysis / Disambiguation of meaning</td>
</tr>
<tr>
<td>• Creating visual output for repetition links</td>
</tr>
<tr>
<td>• Calculating links</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3 Establishing bonds</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Locating bonds</td>
</tr>
<tr>
<td>• Creating visual output for bonds</td>
</tr>
<tr>
<td>• Calculating bonds</td>
</tr>
</tbody>
</table>

Figure 1 The steps of the new LRA model

Designing the first phase of the new LRA model, **Preparation of the corpus**, became necessary due to the change from manual to computer-aided analysis. The second and third phases, **Establishing links** and **Establishing bonds**, were based on Hoey’s (1991) original LRA framework which was further developed by Károly (2002). These are also modified, however, only to a lesser extent. The phases are considered modules because they can be independently developed and comprise different actions. They need to be linked to be part of a LRA computer program. Several of these steps can utilize existing computer applications, however, they all need further testing on larger data.

### 5.2.1 L2 special corpora treatment / Error treatment

While teachers correct essays, they look for lexical, grammatical, structural and mechanical errors. Before a learner corpus is handed over to a program for textual analysis, however, decisions need to be made about error corrections: whether to change anything in the
text or not; and if errors are corrected: what and how to correct. The aim of this research was to find semantic links in the learner texts, therefore, several errors had to be eliminated. To this end, the following treatment is suggested for computer-aided LRA:

- Errors of mechanics regarding the sentence should be corrected: each sentence initial word needs to be capitalized, each sentence should end in a period.
- Spelling errors need to be corrected otherwise links will be missed.
- Multiword units which the writer clearly meant as a one-word unit need to be hyphenated or otherwise signaled to the program to interpret as a single word (e.g. non-native-speaker-English-teacher = non-NEST).
- Semantic errors should not be altered (e.g. errors of style, register)
- Lexico-grammatical mistakes need to be considered on a one-by-one basis.
- Errors in syntax do not need treatment.
- Errors on the discourse level should not be altered.

5.2.2. POS tagging

POS tagging is an important part of sense disambiguation because the English language contains a great number of polysemous and homonymous nouns. In order to find the appropriate link pairs, the right meaning of the word needs to be selected. A POS tagger, which marks the syntactical function of the words may assist in the selection process. For instance, the word *dance* in the sentence *this was our last dance* stands as a noun, in *we dance all night* stands as a verb, and in *I would like to dance* stands as an infinitive. Thus, the coded link pairs could range from simple same unit repetition (exact mention) to derived same unit repetition (e.g., when the word class changes).

So as to be able to suggest a POS tagger available for the analysis, I made a brief experiment with three online tagger applications\(^2\). I tested which program is able to recognize some ‘problematic cases’ in derivations. The following short sentences were entered into the taggers: *This was a good read. John's painting is hung in the hall. John's careful painting of the wall made me jealous. John carefully painting the wall made me jealous. Painting can be dangerous. You can bank on it.*

The first two applications misidentified sentence 3 in which *painting* referred to an ongoing action. The first POS tagger did not recognize that *read* in that context is a noun

---

premodified by a determiner and an adjective. This tagger uses the Penn Treebank tagset\(^3\), whose accuracy is about 97.1 percent on ordinary texts. According to the tagset description, VBG means verb, gerund or participle, making no distinction between the three. As a consequence for our research, this tagger cannot distinguish between derived and inflected verbs therefore, cannot distinguish between same unit repetition and derived unit repetition. The second POS tagger misinterpreted painting in a sentence initial position, identifying it as a proper noun. The third tagger (CLAWS) was developed by UCREL (University Centre for Computer Corpus Research on Language, Lancaster). Based on the results, this proved to be the most reliable of the three, thus this is suggested for the analysis.

5.2.3 POS tagging for lower level L2 texts

The summary and the compare/contrast essay corpora were written by near-proficiency level language learners. However, texts written by lower level students might contain errors which can be an obstacle for the POS taggers we analyzed above. Dickinson and Ragheb (2013) give a detailed description of their recommended POS annotation practice for a learner corpus with the annotation tool Brat\(^4\) devised for learner corpus. Their assumption is that the linguistic violations in any learner text are characteristic of the learner’s interlanguage (i.e. the stage of language development the language learner is at the moment of writing), therefore, annotation should focus on the context and not on the mistake. Their goal is to mark syntactic and morpho-syntactic information with as little error encoding as possible. Their advice for annotators is:

“Try to assume as little as possible about the intended meaning of the learner. ... Specifically, do the following: (1) fit the sentence into the context, if possible; (2) if not possible, at least assume that the sentence is syntactically well-formed (possibly ignoring semantics) and (3) if that fails, at least assume the word is the word it appears to be (i.e., do not substitute another word in for the word which is present).” (p.3).

In other words, do not try to guess what the learner wanted to mean because this is prone to mistakes. Given that in some cases several interpretations of the same sentence are possible, a suggested annotation practice is illustrated in Table 3. For our purposes this suggested treatment of linguistic violations seems a viable option, because if we put extra words into the text on assumption of what the writer intended, we might put a link into the text where it was

\(^3\) https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

\(^4\) http://brat.nlplab.org/
originally not intended. Further detailed descriptions of error coding in learner corpora can be found for instance, in a Cambridge Learner Corpus error analysis study by Nicholls (2003).

Interpretation | Analysis of interpretation
---|---
*I think that my life is mine. In Korea, however Parents’s think is very important.*

**not suggested:**
In Korea, however, what parents think is very important.

parents (common noun, plural, nominal)
think (verb, present simple, singular)
*Parents’s* misformed possessive
*what* missing word

**suggested:**
In Korea, how parents think is very important

parents (common noun, plural, nominal)
think (verb, present simple, singular)
’*s* minor orthographical issue
*however* lexical issue; used instead of *how*
the sentence is well-formed:
*Parents’s* = subject of *think*
*however* = adjunct of *think*

Table 3. Two interpretations of a learner sentence and their analyzes (based on Dickinson & Ragheb, 2013)

5.2.4 Using WordNet with the existing taxonomy

For establishing the types of links other than same unit repetition, WordNet is suggested as an online option. Still, the types of links in the existing LRA model need further study. Károly (2002) already revised Hoey’s (1991) taxonomy in order to make it less ambiguous and to better serve the coding process. However, the question arises, whether the taxonomy should be aligned with the WordNet thesaural categories for the sake of making the analysis more viable by computerized means. This change to Károly’s categorization, nevertheless, will raise difficulties in determining the ‘right’ ratio for the different link categories because the WordNet thesaurus defines more types of semantic relations (Fellbaum, 1998) than Károly’s original model does. Whether employing more categories would help or hinder the LRA, will have to be resolved by further research.

5.2.5 Using WordNet with errors in a learner corpus

How WordNet can be used with a learner corpus needs further testing. As it is described in the Frequently Asked Questions section of WordNet, the program looks for the base forms of words “stripping common English endings until it finds a word form present in WordNet”. It also “assumes its input is a valid inflected form. So, it will take “childes” to "child”, even
though "childes" is not a word"\(^5\). What is important for this research is that although pre-treatment of the corpus is necessary for learner errors, some spelling mistakes such as overgeneralizing certain existing grammar rules (e.g., the -\(s\) vs. -\(es\) plural suffix) may not cause problems for WordNet in the identification of words.

5.3 Theoretical considerations

5.3.1 Altering the taxonomy

A key decision is whether the taxonomy should be changed or should remain intact. This needs to be decided in light of the results in Stages 1 and 2 in connection with the ratio of lexical repetition types. Even though the sample size did not make it possible to find significant differences, both Károly’s (2002) results based on argumentative essays, and the results of this research (in both stages) indicate that there is a strong positive correlation between the frequency of derived repetition and cohesion, and as a consequence, between the frequency of derived repetition and coherence, thus having a major influence on discourse quality. An example for this could be the following two adjacent sentences:

(S5) O’Reily decided to change flight timetables in order to increase turnaround times.
(S6) This decision resulted in more income for the company.

These two sentences are connected by a derived repetition link which is very common in academic writing. More typical examples in adjacent sentences could be e.g., \(X\) claimed that ... -- This claim .... As mentioned above, this structure as a cohesive device is typical between two adjacent sentences in academic discourse. In Károly’s (2002) model, there is a clear distinction between simple and derived repetition, the basis of which is the distinction between inflected and derived word forms. These appear as two distinct categories in the taxonomy. She claims that high-rated essays contain more derived repetitions, therefore the revised taxonomy to be developed for large-scale analysis should keep the original two categories, unless large-scale research does not suggest otherwise. The concordancing program used in Stage 2 however, could not make a distinction between the two repetition types.

If no alternative concordancer is found which can distinguish between inflection and derivation, two possible solutions are suggested: (1) either sacrifice this distinguishing function of the model, or (2) trace for derived repetition links between adjacent sentences with a sentence parser in order to observe whether they contain such elements. Following this, the ratio of such links should be collated with the higher category same unit repetition links (derived repetition links in adjacent sentences divided by ratio of same unit repetition links). Large-scale

\(^5\) https://wordnet.princeton.edu/faq
investigations then will be able to inform us whether the observation of this discourse feature indicates statistically significant differences between high and low-rated texts or not.

5.3.2 Introducing the concept of ‘key term’ into the coding process

In Stage 1, one of the first issues to be solved was how to code the proper noun Ryanair because the frequency of the category instantial relations would have grown ‘out of proportion’ compared to the other types of links in the summaries. Examining summary tasks across various disciplines revealed that it is not uncommon that students are asked to summarize a source text in which the main topic is specific and is referred to with a proper name (i.e., the name of a person, a company, an action, a theory, a model): in other words, a specific instead of a general concept. Given that topic specificity needs a unified treatment, the following analytical decision is suggested: if the topic is referred to as a proper noun, this noun should be considered a key term, and lexical repetition links this term enters into should be considered simple repetition (in cases of word-for-word repeating) or simple synonym (in cases of proper noun—equation mentioning). Only one proper noun lexical unit should be treated key term in each text.

5.3.3 Lexical unit identification in the case of multiword units

Proper identification of the units of analysis is key to gaining valid and reliable results in lexical repetition analysis. We distinguish between one-word lexical units and multi-word lexical units, the most problematic for this research being noun compounds. The English language produces a great number of noun compounds and each new domain brings along more of their specific new noun compounds not present in WordNet, which was trained on the Brown Corpus (a general corpus of English) not on special corpora for academic English or even more specific subcorpora based on professional registers. Specialist dictionaries and handbooks are necessary, just as it was done ‘intuitively’ in the coding phase. Prószéky and Földes (2005) describe their comprehension assistant software as capable of identifying multiword expressions. This is achieved by using several big capacity dictionaries and restructuring their entries by splitting up original entries, thus separating single and multiword expressions; the latter then could be entered as new headwords. This way the program became capable of detecting multiword expressions, which might seem a solution to the problem of identifying context-specific multiword units.

5.3.4 Connecting the new LRA model to a cognitive framework

In order to contextualize our refined model, a parallel can be drawn between the mental processes during reading described in Khalifa and Weir’s (2009) cognitive framework and the steps of our newly proposed computer-aided lexical repetition model. See Table 4 for a detailed explanation of how each analytical step corresponds to the stages of human reading.
<table>
<thead>
<tr>
<th>Processes of reading</th>
<th>Operationalized processes (adapted by Bax, 2013)</th>
<th>Explanation of processes</th>
<th>Steps of the new, computer-aided LRA model</th>
</tr>
</thead>
<tbody>
<tr>
<td>word recognition</td>
<td>word matching, word-level</td>
<td>reader identifies same word in question and text</td>
<td>identifying lexical units</td>
</tr>
<tr>
<td>lexical access</td>
<td>synonym, antonym and other related word matching, word-level</td>
<td>identifying word meaning and word class</td>
<td>establishing same unit repetition links</td>
</tr>
<tr>
<td>syntactic parsing</td>
<td>grammar/syntax parsing, clause/sentence-level</td>
<td>reader disambiguates word meaning and identifies answer</td>
<td>POS tagging of sentences</td>
</tr>
<tr>
<td>establishing propositional meaning</td>
<td>establishing propositional meaning, sentence-level</td>
<td>reader establishes meaning of a sentence</td>
<td>disambiguation of meaning</td>
</tr>
<tr>
<td>inferencing</td>
<td>inferencing, sentence/paragraph / text level</td>
<td>beyond literal meaning to infer further significance</td>
<td>~ establishing bonds</td>
</tr>
<tr>
<td>building a mental model</td>
<td>building a mental model, text-level</td>
<td>using several features of text</td>
<td>~ establishing bonds</td>
</tr>
<tr>
<td>creating a text level representation</td>
<td>understanding text function, text-level</td>
<td>using genre knowledge to identify text structure and purpose</td>
<td>~ establishing bonds</td>
</tr>
<tr>
<td>creating an intertextual representation</td>
<td>not in test situation, between texts</td>
<td>comparing texts</td>
<td>not relevant, but present in Hoey’s (1991) original model: establishing links/ bonds between texts</td>
</tr>
</tbody>
</table>

Table 4. The contextualization of the new LRA model in this study: the parallel processes between the model and the cognitive processes during reading (Khalifa & Weir, 2009), with explanations

5.4 Visual representation of links and bonds

Identified links and bonds can be illustrated in various ways. During the manual analysis connectivity matrices and tables were used to represent the links and bonds within a text. In Teich and Fankhauser’s (2005) study the matrix format was substituted by indexing each sentence with link location numbering. An even more sophisticated option could be representing the actual text with either various lexical repetition link types highlighted, or central sentences highlighted. This way the building blocks of the text, such as the paragraphs,
their boundaries, and the sentences with special discourse function could also be represented within their discourse functions, and the teachers could comment immediately, for instance, on ill-formed, unconnected sentences by referring to their missing centrality which was identified by the model.

6. Conclusions
6.1 Special use of the model for academic summary writing

The new model can have two distinct uses in the academic summary writing process. Firstly, it can be used during the input text selection phase, when the main ideas need to be extracted from the source text, according to Hoey’s (1991) original idea. By applying the steps (Establishing links and Establishing bonds), the central sentences can be collected and the resulting abridged version of the input text can substitute for the manual collection of the main ideas, which so far have been generated by teachers.

Alternatively, the main points of the text can be collected by teachers in the traditional way, and their decisions can be ‘objectively’ tested by applying the tool on the text. It has to be noted, however, that this kind of main idea collection gives only valid results if the students need to write a whole-text summary. If the task is guided summary writing, the content points students need to select will probably differ from the main ideas of the input text, and in this case the central sentences in the input text will not coincide with the information required.

The second use of the model is similar to the one described in Stage 1 of this research, when it was used to distinguish the quality of summaries, observing the quantity and types of links and bonds within texts. However, using the computer-aided model it will also be possible to compare the organizations of both the input and output texts with the same analytical tool, which might reveal so far hidden similarities or differences in their lexical patterning.

6.2 Limitations

It is important to emphasize that this lexical repetition analysis model was designed as an aid intending to gain results in connection with cohesion, a text-internal concept. It is out of the scope of this model to address coherence, a concept which is text-external. In other words, the lexical repetition links within the text, whose patterns the model is attempting to capture, are overt, ‘countable’ cohesive links. On the other hand, coherence is described in this study as the interaction between the reader and the text, and is therefore seen as a subjective concept, and as such, the model does not intend to interpret it.

For the same reason, the design is computer-aided rather than fully automated, observing lexical links as data, and disregarding other aspects of discourse quality, such as
syntactical, grammatical and stylistic variables, or their interactions. However, given that human readers’ overall quality judgements on texts are influenced by cohesion, the model aims at positively correlating with readers’ overall quality judgements in this respect.

Two main factors limited this study: the small size of the sample and the variables of the task. Due to the fact that the original model was devised for manual analysis, only a limited number of student writing could be analyzed. Therefore, although several interesting results were revealed by the model, statistical significance could not be calculated due to the small sample size, only certain tendencies could be observed. Such finding was, for instance, that perhaps contrary to assumption, high-rated summaries contained not only a higher number of simple synonymy and derived opposites, but also more simple repetition links.

As far as task variables are concerned, the analytical tool in its present form may predict subjective perceptions of the quality of the type of summary observed, however, no conclusions could be drawn on its reliability and validity in cases when the original document to be abridged uses a different narrative form. Similarly, the influence of the length limit of summaries were not examined either. In the case of the compare/contrast essay corpus, the uneven ratio of the two essay patterns (block and point-by-point patterns) made it impossible to draw reliable conclusions on whether the model can distinguish between the two patterns with regard to the clustering of bonds according to patterning. These limitations motivated the study to investigate how to apply the model on larger corpora.

6.3 Suggestions for further research

The first issue that would deserve further investigations is the treatment of collocations. These are non-systemic semantic relations which are excluded from the analysis. Morris and Hirst (2004) report a study where readers had to identify words in a general news article, which, in their view, were semantically related. With 63% agreement the result showed that the readers identified word groups and not word pairs when they had to give labels for the relations. This finding might suggest that perhaps cohesion is perceived in a formation which is different from links. Hoey (1991) already attempted to describe the nature of these formations with the link triangle/the mediator missing concept, but did not further elaborate on this idea.

Another interesting finding in Morris and Hirst (2004) is that most identified word pairs were collocations. These relations, which represent world knowledge, are perceived strongly by the reader as semantically related. Even though it is a fact that collocations mostly appear within the same sentence, it would be interesting to study to what extent such (rather intra-sentential) semantic relationships influence discourse cohesion, and how these could be incorporated into the proposed model.
The next theoretical issue to consider is the language level of the writers and its consequences for the written product with regard to the present lexical repetition model. Among the many possible difficulties that might arise from learner errors, only the issue of faulty sentence creation is mentioned now, as a problem area. Several language learners violate the two basic formal rules of sentence building, namely that sentences should start by a capital letter and end in a period, question mark, or exclamation mark. This type of error is prominent particularly below levels IELTS 5.5 / CEFR B1, and has to be manually corrected, although the sentence boundaries are not always certain.

A further reason for not gaining valid results when observing inter-sentential links might be that sentence creation is only partially based on fixed compulsory elements: there is also room for writer creativity. The same information content can be packaged into one sentence or divided between two sentences. For example, a compound sentence with and can be rewritten as two separate sentences with the same meaning spread out in two sentences using a connective, such as moreover, furthermore, etc. This is a key issue if we want to analyze text with a tool based on inter-sentential relations. The same problem seemed to arise in the case of a research study utilizing Latent Semantic Analysis (Landauer, Laham, & Foltz, 2003), which analytical tool assesses semantic relatedness between adjacent sentences using vector-based similarity.

The last area for further research is connected to language technology. During this research a number of existing programs were analyzed, typically those which have already been reviewed in scientific journals, or the ones that offer publicly available manuals. Such were, for instance, Concordancer 3.3; Coh-Metrix; several POS taggers; or WordNet. It is possible, however, that other commercially available programs exist, perhaps in modular format, which might be suitable for certain steps of this analytical process. If not, such a modular program can be built aligned with the newly designed LRA model. Especially if it graphically represents links and bonds, such program can be sold as a self-contained product or online writing tool to assist academic writing teachers and their EFL students.
References of the summary


**Computer software**

