EXPLORATIONS IN LEXICAL REPETITION ANALYSIS: THE OUTCOMES OF MANUAL VS. COMPUTER AUTOMATED RESEARCH METHODS

doi.org/10.61425/wplp.2013.07.1.28

Mária Adorján
Language Pedagogy PhD Program, Eötvös Loránd University, Budapest
mary@adorjan.com

Abstract: This paper describes two exploratory studies applying Károly's (2002) theory-based discourse analytical framework, whose pedagogically oriented taxonomy and methodological procedures were originally devised to investigate the text-structuring role of lexical repetition in a Hungarian EFL academic context. The first study consists of entirely manually conducted data collection and analysis, whereas the second study – in both of its stages – uses computer assistance. The latter study poses particular problems that are recorded with a view to possible future automation. The application of the analytical framework was also extended from argumentative essays to summaries and comparison-and-contrast essays. Results reveal that the analytical tool is capable of extracting rich data in the newly investigated genres, too. Further research is necessary, however, into the application of the tool for large corpora.

Keywords: Lexical repetition, discourse analysis, summary, comparison-and-contrast essay, computerized analysis

1 Introduction

Cohesion and coherence in text have become widely researched areas within the field of discourse analysis, and a great deal of attention has been given to the subjects of lexical cohesion and lexical repetition due to their significant discourse function (e.g., Halliday, 1985; Halliday & Hasan, 1976; Hoey, 1991; Reynolds, 1995; Tyler, 1994, 1995). 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. 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, p. 9). In a more recent corpus linguistic study Teich and Fankhauser claimed that nearly fifty percent of a text's cohesive ties consist of lexical cohesion devices (2004, p. 327), thus making lexical cohesion the most pronounced contributor to semantic coherence.

It is important to note that in Hoey's (1991, pp. 51-75) repetition model, lexical repetition is described in a broader sense than usual. It is used synonymously with lexical cohesion because it comprises reiteration (repeating the same word), and also paraphrase, i.e., repeating through other lexical items which are semantically related, such as antonymy, synonymy or superordinates.

The study of lexical cohesion is relevant as the role of lexical repetition patterns in English written text is controversial among native and non-native language users and linguists alike. If we examine reiterations, for instance, Connor (1984) found that repeating words is a sign of limited vocabulary or poor text structuring, which most teachers would agree with. However, the problem is more complex because lexical choice depends on various factors. According to Reynolds (2001), for example, lexical repetition used by writers changes in relation to writing topic, cultural background, and development of writing ability, the third being the most determining factor. Scientific articles, on the other hand, require more reiterations than popular articles (Myers, 1991) because exact concepts cannot be replaced by synonyms. This complexity calls for more research into lexical cohesion in general, and into texts produced by language learners on various topics and genres.

Lexical cohesion is studied both in text linguistics (discourse analysis)[1] and corpus linguistics. In the former field, 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. Reviewing state-of-the-art 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).

Several manual and computer-aided methods are available to analyse 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. Károly's (2002) theory-based analytical tool designed for the study of lexical repetition, which was the starting point for the present research project, is one of the frameworks devised to offer a manual analytical method for studying the text-structuring role of lexical repetition. It is a revised version of a comprehensive analytical model created by Hoey (1991) to reveal the organizing function of lexical repetition in texts.

Károly's (2002) research results showed that her theory-driven "objective" analytical tool not only offered a descriptive function, but with her analytical measures, the tool was capable of predicting the "intuitive" assessment of teachers evaluating the essays with regard to the structure of EFL academic argumentative essays. The results of her analysis proved that the texts, which

[1] The term text linguistics and discourse analysis cover related but not the same kind of approaches to the study of text. The analysis of the two concepts fall beyond the scope of this study. For a detailed discussion of the differences of the two, see Widdowson (1996) and de Beaugrande (1997).

had previously been rated high or low by experienced university instructors, differed significantly in both repetition types and patterns. Post-tests conducted with another group of teachers confirmed these findings, thus indicating that the analytical measures devised are reliable and the results may be generalized for a wider sample.

Even though Hoey's (1991) framework has been researched extensively for decades, and there is a visible revival of his analytical model from a corpus linguistics angle, few studies have addressed Károly's (2002) model. Seidl-Péch's (2011) study, which is a computer-aided application of Károly's (2002) taxonomy on large corpora, is one of these, indicating that it is feasible to use Károly's framework for computerized analysis. Seidl-Péch, however, limited her explorations to the taxonomy, and did not follow Károly's analytical steps due to methodological decisions based on a different research focus, namely studying the quality of translation.

Intending to fill this gap, I undertook two studies (Adorján, 2011a; Adorján, 2011b) aimed at testing the applicability of Károly's (2002) lexical repetition analysis framework to two genres: summaries and comparison-and-contrast essays. While drawing on the results of these studies, another aim of my research was to gain insights into theoretical and methodological questions on the application of the employed analytical steps in the context of larger corpora, given that such a method – to my knowledge – does not exist.

2 Background and aims

The theoretical background of the study focuses on the following issues: first, Hoey's (1991) and Károly's (2002) lexical repetition models are presented, followed by the description of Károly's empirical investigation, which was the starting point of this research project. Next, some examples follow of how these models have been applied on larger corpora.

2.1 Lexical cohesion and lexical repetition within the study of coherence and cohesion

Lexical organization and its role in establishing coherence have been the focus of several influential studies (e.g., Halliday & Hasan, 1976, 1985; Hoey, 1991; Reynolds, 1995; Sinclair, 1998; Tyler, 1994, 1995). Hoey focused his research on cohesion, claiming that markers of cohesion appear in the text as observable features, while studying coherence is out of the scope of textual analysis because it "is a facet of the reader's evalutation of a text" (1991, p. 12). The dominant role of lexical cohesion within cohesion types as "the only type of cohesion that regularly forms multiple relationships" in text (p. 10), also made it relevant to explore. He maintained that lexical repetition is suitable for "objective" analysis, and as such, countable, categorizable and "capable in principle of automatic recognition" (1991, p. 12).

2.2 Hoey's (1991) model for analyzing the text organizing role of lexical repetition

Hoey (1991) was the first to provide a comprehensive analytical model which reveals the organizing function of lexical repetition in texts. In his view, the role of grammatical cohesion is less significant than that of lexical cohesion, therefore, he focused on words with lexical

meaning. He analyzed newspaper articles to demonstrate how patterns of lexical repetition work between adjoining sentences and over considerable distances within a given text. He devised a new taxonomy of lexical repetition types and observed their patterns in text formation. These categories are presented in Table 1.

REPETITION			**Example**
I. Lexical repetition	simple		*bear-bears*
	complex		*drug-drugging*
II. Paraphrase	simple		*produce-cause*
	complex	antonymy	*hot-cold*
		link triangle	*writer-author-writing*
		the "mediator" missing	*writer-(author)-writing*
		other	superordinates *(biologists-scientists)* co-reference *(Augustus-the Emperor)*
III Non-lexical repetition	substitution links		e.g. *personal pronouns, modifiers*

Table 1. Hoey's (1991) taxonomy of lexical repetition types (the table and the examples are based on Károly, 2002, p. 80)

Hoey (1991) defines the key concepts of his taxonomy in the following way:

- Simple lexical repetition occurs "when a lexical item that has already occurred in a text is repeated with no greater alternation than is entirely explicable in terms of a closed grammatical paradigm" (p. 55).
- Complex lexical repetition occurs "either when two lexical items share a lexical morpheme, but are not formally identical, or when they *are* formally identical, but have different grammatical functions" (p. 55).
- 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" (p. 62).
- Complex paraphrase occurs when "two lexical items are definable such that one of the items includes the other, although they share no lexical morpheme". This category is broken down into three subcategories: 1. Antonymy, 2/a Link triangle, 2/b The "mediator" missing, 3. Other types of complex paraphrase: superordinates and co-reference (p. 64).

Hoey claimed that "lexical items form *links,* and sentences sharing three or more links form *bonds*" (p. 91). Bonded sentences lead to nets, which organize text, in a manner similar to Hasan's (1984) identity and similarity chains[2]. Hoey found that bonded sentences are central to

[2] According to Hasan (1984), cohesive chains occur when an element in text refers back to a previous element. The first type, *identity chain*, shares the same referent (*girl, she, she*), while the second type, *similarity chain*, includes other types of reoccurrences in text, not necessarily with the same referent, e.g., *went out, got … home* (examples from Hasan, 1984, p. 212).

text, as they contain the main information (macropropositions[3]). Sentences with no links or (for some texts) few links, contain merely additional information, and can be considered marginal (Hoey, 1991, p. 91). Hoey's main claim that links created via lexical repetition may form bonds which highlight significant sentences was later reaffirmed by Reynolds (1995).

Hoey (1991) used newspaper articles and a non-narrative book for his analysis. Based on the results taken from the short sample, he showed how abridgements/summaries could be created from a longer text by deleting marginal sentences, collecting central sentences, or selecting topic opening and topic closing sentences. He admitted that these modes would summarize different aspects of the original text with shifts in meaning. Although Hoey presented these models as possible means to create summaries, he did not give guidance on how to distinguish between their quality. He argued, however, that lexical repetition patterns revealed by his analytical tool can indicate differences in text quality.

Tyler (1995) criticized Hoey on the grounds that quantitative investigation of lexical repetition alone cannot capture the difference between well and badly formed texts: qualitative analysis is necessary to explore how repetition is used. Tyler's (1992) empirical study indicated that repetition in itself was not sufficient to cause cohesion, the perceived quality difference of native and non-native speakers' language production is influenced by *what* and *how* is repeated. Nevertheless, she did not contradict Hoey's main claim regarding the function of bonds as text-building devices.

2.3 Károly's (2002) lexical repetition model

Hoey's (1991) taxonomy was revised in Károly's (2002) study, putting it into a wider perspective. She pointed out that the original model contained three weaknesses: (1) theoretical problems with the taxonomy, such as several obscure category labels, and the unclear definition of the basic unit of analysis, (2) weaknesses of the method of analysis (such as not examining intra-sentential repetition, or the missing theoretical foundation for choosing the number of bonds to be seen as significant connections), (3) research methodological problems (such as making strong claims based on a single text type).

Károly (2002) introduced the term *lexical unit* as the basic unit of analysis. This is a unit "whose meaning cannot be compositionally derived from the meaning of its constituent elements" (Károly, 2002, p. 97), i.e., together the individual words placed one after the other mean something different than each word means standing alone. A *lexical unit* can be a one-word unit, an idiom or a phrasal compound (words expressing a unique concept, e.g., *Non-Native English Speaking Teachers, non-NEST-s*). She also proposed a new taxonomy of the lexical repetition types, as indicated in Table 2. As the table shows, Károly operates with more traditional grammatical terms. Her *instantial relations* category introduces a semantic category which is temporarily bound by context, and resembles Hasan's (1994) *instantial lexical cohesion* category, which was originally broken down to *equivalence*, *naming* and *semblance.*

[3] See van Dijk (1977), for a detailed description.

Categories		**Examples**
Lexical relations I. Same unit repetition		
1. repetition	simple	*writers – writers*
	derived	*writers – write*
II. Different unit repetition		
2. synonymy	simple	*to exercise – (after) working out*
	derived	*built - construction*
3. opposites	simple	*small - major*
	derived	*hatred - like*
4. hyponymy		*languages - English*
5. meronymy		*hands - fingers*
Text-bound relations 6. instantial relations		*manager – O'Leary*

Table 2. Summary of the categories of repetition (based on Károly, 2002, p. 104)

Károly (2002) also introduced a number of new analytical measures related to the combination of links and bonds to extend the research capacity of the analytical tool. For instance, the *length of bonds* category indicating how far apart bonded sentences are located from each other, and the distinction between *adjacent bonds* and *non-adjacent bonds* to indicate which sentences form mutual relationships. *The strength of bonds* was calculated to reveal how many links connect sentences in the given text. Appendix A lists the analytical measures.

2.4 Conceptual differences between the two models

If we examine the conceptual differences between Hoey's (1991) and Károly's (2002) taxonomies in terms of how suitable each one is for computerized analysis, we find that the former operates with two categories: the *Link triangle* and *The mediator missing*, which might pose problems for computerized research, even in the eyes of a non-expert. While Károly's system records only one-to-one relationships between intersentential lexical units, Hoey's *Link triangle* category inadvertently confuses his own taxonomy by looking for connections between more than two elements at the same time. While data on frequencies and locations of intersentential relationships between lexical units can be observed and analyzed relatively easily, triangle-type relationships would be more difficult to detect and record. Triangle frequencies could also prove to be impossible to interpret alongside the other types of data. If, we find three links between the first and the 20th sentence, according to Hoey, we can claim that these two sentences are bonded. In other words, there is a strong semantic and structural relationship between them with a distance of over 20 sentences. However, it is not described what procedure should be followed if there is another word in sentence 21, which may be a candidate for a link triangle: how this would influence the overall number and length of bonds within the text.

The other category, *The mediator missing*, is also difficult to empirically observe, since it looks for links in places where there are none overtly present. Such missing mediators should be

manually tagged, on a case-by-case basis. This would seriously slow down the analysis, if not make it impossible. Moreover, mediator words are usually nouns, which could be substituted by pronouns. Therefore they would be excluded from the analysis, in accordance with Hoey's theoretical decision.

2.5 Károly's (2002) empirical investigation

Károly (2002) investigated the organization of ten high-rated and ten low-rated argumentative EFL essays. Her main hypothesis was that her revised tool is able to differentiate between well and badly-structured essays, based on the role lexical repetition plays in structuring texts. Her method of analysis focused on new aspects of bonds, such as their position, length, and strength between sentences with special discourse function (SDF), such as the title, the thesis statement, the topic sentences and the concluding sentences. Her results indicated that (1) high-rated essays contained more repetition links, including more same unit repetition links, (2) the number of bonds connecting SDF sentences was higher in high-rated essays. Károly's first result revealed thus far hidden dimensions of lexical repetition, such as (1) above, which means that even high-rated essays contained many reiterations, although it is common teachers' practice to advise against it in texts. The analytical measures are described in Appendix A and a sample analysis containing the steps is provided in the Method section.

2.6 Applying Hoey's (1991) lexical cohesion model to a large corpus

At the time of his research, Hoey did not have access to a computer program specifically designed to assist his analysis of longer texts. In his book on lexical repetition patterns, he analyzed a 5-sentence long article in detail, as well as the first forty sentences of the first chapter of a non-narrative book. He concluded that *theoretically* it is possible to create summaries of texts of "unlimited length" applying his repetition model, but he did not give instructions on how to do so in practice. Furthermore, the process of comparing extremely long texts with their summaries was not examined.

The first text-processing computer application based on Hoey's model (Tele-Pattan) was created by Benbrahim and Ahmad in 1994 (de Oliveira, Ahmad, & Gillam, 1996). It represented a computer implementation of two of Hoey's four lexical repetition categories: Simple Repetition and Complex Repetition (de Oliveira et al. 1996). The program created five summaries of the same stock exchange news, which were then evaluated by four traders and five university students. The outcome was that 60% of the raters felt that essential information was missing, and that participants evaluated the summaries differently. As the texts were not available in the research paper, the results cannot be confirmed. However, it can be argued that the text-processing program was limited in use because (1) it incorporated only two of Hoey's categories, and perhaps as a consequence (2) the resulting summaries were rated differently, even though the type of text (stock exchange news) did not allow for a wide variety of lexical choice and sentence structure. Additionally, for a non-expert it would seem relatively easy to summarize such a functionally 'predictable' genre.

As the size of available and searchable corpora increased significantly, British Telecom financed a project, lead by Hoey and Collier, to design a "software suite" for the abridgement of electronic text (Collier, 1994), by automatically selecting central sentences, i.e., sentences containing the macropropositions in text. The program was able to create a matrix of links in seconds, but again, only for the two basic repetition categories: simple and complex repetition. According to Collier (1994), thesaural links were added manually to analyze antonyms, but this step resulted in only a minor improvement in the program. His research plan lists several semantic and structural difficulties in automating central concordance line selection and he concludes that further research is necessary into these areas. Two programs evolved from the original version: a document similarity tracer (Shares), and an automatic document summarization/abridgement system (Seagull). A demo version of both can be accessed at the Birmingham City University Research and Development Unit for English Studies website.

As Collier described above, the automated identification of repetition links was attempted using a concordance selector. Due to the extremely labourious nature of data collection, many studies utilize a concordance program (e.g., AntConc, Concordance) to search discourse data in the area of investigating lexical repetition patterns. As data is textual, a frequency analysis software is helpful in counting how many times certain words appear in the text. It is also possible, using a concordancing application, to count how many times certain pairs are repeated. The software is able to show in which sentence the repetitions occur. It cannot evaluate qualitative data, however, without a human observer to process information (Hunston, 2002).

2.7 A corpus-based investigation using Károly's (2002) taxonomy

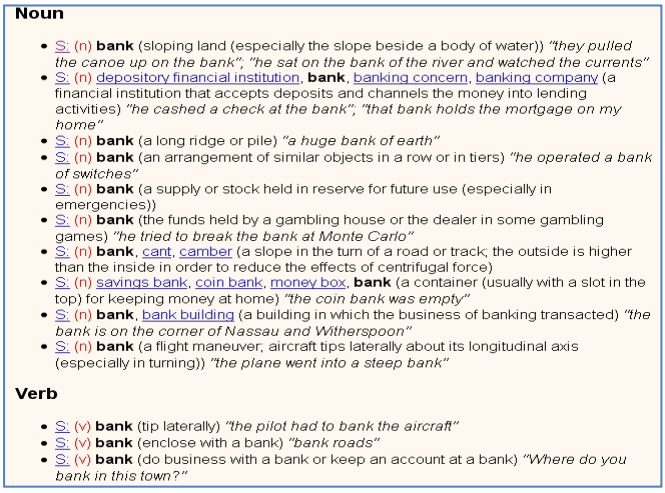
A recent empirical investigation based on Károly's (2002) taxonomy aimed to compare shifts in lexical cohesion patterns between translated and authentic Hungarian texts (Seidl-Péch, 2011). Seidl-Péch found that authentic Hungarian and translated Hungarian texts differ in lexical cohesion patterns. Her quantitative analysis was facilitated by language technology modules provided by Orosz and Laki (Laki, 2011; Novák, Orosz, & Indig, 2011), whose linguistic parser (analyser) program helped to automate the analysis. The Hungarian WordNet Program (Prószéky & Mihaltz, 2008) was used to explore semantic links between sentences.

Although Seidl-Péch's study was the first to utilize Károly's lexical repetition analysis framework for a multilingual corpus-based investigation, it cannot be considered as a model for further research for several reasons. Firstly, due to the limitations of the HunNet application, the scope of Seidl-Péch's research was described as limited to investigating only nouns. In the results section, however, the screenshots revealed that the software also analyzed pronouns (*azt, arra,* p. 135). It is possible that the number of pronoun repetitions was also included in the sum of repetitions. If not, it is not explained how they were discarded.

Secondly, she did not provide enough details on how the application analyzed the texts exactly. She did not explain, for example, how lexical sense disambiguation occurred precisely. An English example would be this: How did the application decide which were the synonym sets for *bank*? As lexical repetition analysis sets out to identify semantic relations, and honomymy is frequent in English, the key methodological question is whether the software offered this word for the researcher to manually choose the right meaning in the given context, or the application

selected the right meaning on its own, entirely automatically. In the latter case, it is of key importance to examine how the program decided which meaning was relevant.

To explore this feature in the English version of WordNet, on which the HunWordNet was based, I experimented with the word *bank* to find out which meaning is considered first: the most frequent, the most likely[4] or some other factors are considered? The result was that *bank* as *sloping land* was offered before *bank* as *financial institution* (as shown in Figure 1 below), which might mean that the WordNet application was trained on literary texts as a database to calculate frequencies, and not on business (or even news) texts. This raises the question of how synonyms or antonyms were counted by HunWordNet in Seidl-Péch's (2011) research.



Noun

- S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"*, *"he sat on the bank of the river and watched the currents"*
- S: (n) depository financial institution, **bank**, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) *"he cashed a check at the bank"*, *"that bank holds the mortgage on my home"*
- S: (n) **bank** (a long ridge or pile) *"a huge bank of earth"*
- S: (n) **bank** (an arrangement of similar objects in a row or in tiers) *"he operated a bank of switches"*
- S: (n) **bank** (a supply or stock held in reserve for future use (especially in emergencies))
- S: (n) **bank** (the funds held by a gambling house or the dealer in some gambling games) *"he tried to break the bank at Monte Carlo"*
- S: (n) **bank**, cant, camber (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
- S: (n) savings bank, coin bank, money box, **bank** (a container (usually with a slot in the top) for keeping money at home) *"the coin bank was empty"*
- S: (n) **bank**, bank building (a building in which the business of banking transacted) *"the bank is on the corner of Nassau and Witherspoon"*
- S: (n) **bank** (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)) *"the plane went into a steep bank"*

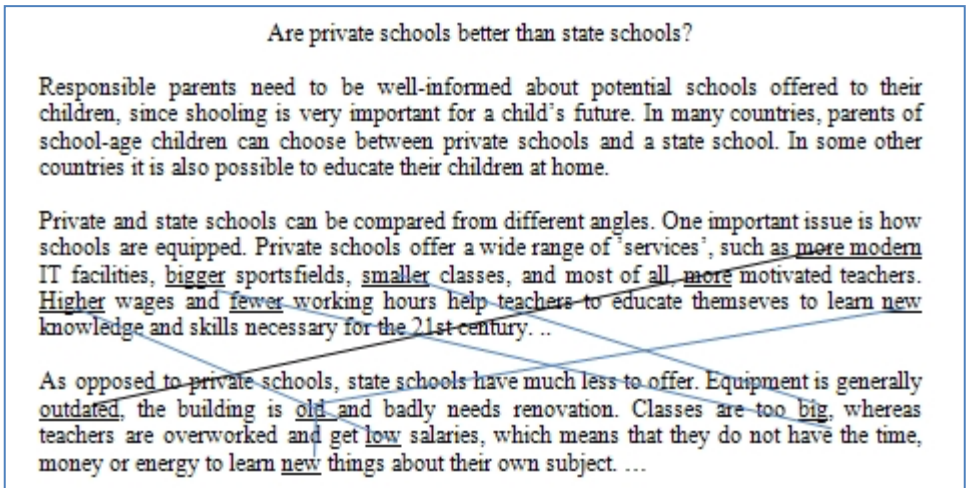
Verb

- S: (v) **bank** (tip laterally) *"the pilot had to bank the aircraft"*
- S: (v) **bank** (enclose with a bank) *"bank roads"*
- S: (v) **bank** (do business with a bank or keep an account at a bank) *"Where do you bank in this town?"*

Figure 1. Synonyms offered for the word *bank* in WordNet

Another issue to consider is that Seidl-Péch limited the scope of her research to nouns. It seems appropriate to do more research into texts which contain more adjectives or verbs than usual to explore their text structuring significance. One genre where adjectives and adverbs are also frequently compared with their opposites is the comparison-and-contrast essay. Figure 2 on the next page shows part of such an essay, presented here as an example of how much lexical cohesion would be lost without the analysis of adjectives and adverbs. (Note: In reality, many more adjectival/adverbial repetition links exist within these two paragraphs than indicated in Figure 2. They are not visible now because the pairs they connect with only appear in later paragraphs.)

[4] I looked up several words in the WordNet dictionary related to the meaning of *bank – as institution* to find out whether the software 'remembers' the previous requests when I asked it to define *bank*. It did not remember. (This was only an unorthodox trial-error test to explore this feature, it is not based on literature.)


Are private schools better than state schools?

Responsible parents need to be well-informed about potential schools offered to their children, since shooling is very important for a child's future. In many countries, parents of school-age children can choose between private schools and a state school. In some other countries it is also possible to educate their children at home.

Private and state schools can be compared from different angles. One important issue is how schools are equipped. Private schools offer a wide range of 'services', such as more modern IT facilities, bigger sportsfields, smaller classes, and most of all, more motivated teachers. Higher wages and fewer working hours help teachers to educate themseves to learn new knowledge and skills necessary for the 21st century. ..

As opposed to private schools, state schools have much less to offer. Equipment is generally outdated, the building is old and badly needs renovation. Classes are too big, whereas teachers are overworked and get low salaries, which means that they do not have the time, money or energy to learn new things about their own subject. ...

Figure 2. Three paragraphs of a sample comparison-and-contrast essay indicating some of the lexical repetition links (adjectives/adverbs)

In light of the previous theoretical and empirical considerations, the present research project has two purposes. Firstly, to gain insights into theoretical and methodological questions based on Károly's (2002) lexical repetition analysis framework by further testing the applicability of her analytical tool on the genres of summary and comparison-and-contrast essay. The second aim is to observe and record theoretical and practical decisions which might arise during the process of manual and computer-aided data collection and analysis, bearing in mind that the tool is intended for computerized corpus analysis.

3 Research questions

This study will be guided by four research questions. Two of them attempt to shed light on theoretical and empirical aspects of studying lexical repetition, focusing on testing the applicability of Károly's (2002) analytical method on further genres in a similar academic context:

(1) How can Károly's (2002) theory-based repetition analysis framework be extended to the study of summaries written by Hungarian EFL university students?
(2) How can Károly's (2002) theory-based lexical repetition analysis framework be extended to the study of comparison-and-contrast academic essays written by Hungarian EFL university students?

The research also focuses on methodological issues, which is reflected by two additional research questions. Manual and partially computerized data collection and analysis are explored. The ultimate aim is to apply the methods to a large corpus. Therefore, possible limitations of computer automated analysis, for example, observing which aspects of text need human context-based decisions, are also investigated with the following two research questions:

(3) How can a concordancing software application (*Concordance, Version 3.3)* facilitate the analysis of the text-organizing function of lexical repetition patterns?

(4) How does the examined corpus need to be prepared for a computer-assisted analysis based on Károly's lexical repetition analysis tool?

4 Methods

4.1 The corpora

More diverse features of the text organizing role of repetition can be observed if we apply the original model to various other genres. The texts selected for the corpora are summaries and comparison-and-contrast essays.

The *summary* corpus consists of 10 summaries, written by a group of English major BA students. The students' task was to read an internet-based text about the history and business model of a company and summarize it in at least one but no more than two pages. The original text consisted of 7,405 words. The information considered important for the summaries was selected by the teacher and had to be incorporated into the summaries by the students. As it was a pedagogically motivated exercise in an ESP course, the teacher also provided some general business vocabulary to be used in the summaries. The evaluation of the summaries was based on assessing (1) whether they contained the information required by the task, (2) whether they omitted irrelevant information, and (3) whether the summaries had logical structuring of topics. The task sheet for the summaries is provided in Appendix B.

The summaries were evaluated by two English teachers, according to two criteria: content and organization. Accuracy and appropriacy were not examined, as they are not relevant to investigating the organizing function of lexical repetition in text (Károly, 2002, p. 127). The content criterion was met by the presence of the required information, as well as by the lack of irrelevant information. Organization was evaluated by the presence or lack of logical structuring of topics. The texts were rated on five-point scales in both categories.

The inter-rater assessment showed 100% agreement on the quality of the summaries. Based on the evaluation criteria, five summaries were assessed as high-rated, and five as low-rated. (It was an interesting feature of the evaluation that neither rater gave more than a one-point difference in scores for content vs. organization for the same summary. In other words, if raters gave 3 points for content, they never gave as much as 5 for organization, and vice versa. A conclusion might be that the two criteria are connected in summaries.)

The *comparison-and-contrast academic essay* corpus consisted of eight texts, written by the same pool of participants. Their task was to write a comparison-and-contrast academic essay on a subject related to applied linguistics of approximately 600 words. The contents of the papers covered three areas: language and communication, culture, and education. Each essay had a title and consisted of 4-7 paragraphs, containing a separate introduction and conclusion paragraph. The corpus contained 4,971 words. This data size, due to the complex nature of manual analysis, was similar to that of international and Hungarian empirical studies in this specialized field of discourse analysis. The essays in this corpus were given a percentage score by the tutor: four were high-ranked, and four were low-ranked.

4.2 Justification for using the particular texts as corpora

The selected texts were assignments during two academic EFL courses. The main concern was to gain data from the same pool of participants, just as Károly (2002) did for her original analytical model. As the usage of lexical cohesion devices is greatly determined by language level (Reynolds, 2001), it was important to collect texts from learners who had already passed the Proficiency Test in English administered by the university (CEFR level C1). The writers of both groups of essays fulfil this criterion, thus data gained from the analysis of their essays can be compared to Károly's (2002) results.

A special feature of this particular group of summaries was that they contained an unusually large number of instantial relations because the subject was a specific agent (Ryanair) and the available synonyms to be used instead of the proper noun were limited (airline, company, firm, carrier). As, according to Károly (2002), the lexical relations loosen from the same unit repetition being the strongest, to instantial relations being the most distant, analyzing this aspect of the sample might reveal how this semantic feature affects lexical cohesion. It is also relevant to examine this from an automation perspective because instantial relations cannot be established without human intervention.

By contrast, the comparison-and-contrast essays were based on theoretical, scientific topics. Therefore, the results gained from these two samples might answer some questions raised by Reynolds (2001) and Myers (1991) about lexical repetition differences according to topic. Comparison-and-contrast essays also lend themselves to experimentation with the elimination or exclusion of adjectives and/or adverbs from the analysis, as discussed in the Theoretical background section in Seidl-Péch's (2011) research, in order to draw conclusions regarding the ratio of noun vs. adjectival/adverbial links within this sample.

Finally, the importance of summaries and comparison-and-contrast essays in an academic writing context has also been described in the literature (e.g., Hammann & Stevens, 2003; Hidi & Anderson, 1986; Hyland, 2006; Spivey, 1990). Their frequency in higher education and academic discourse, cognitively demanding nature, and accessibility in a Hungarian academic EFL context were also taken into consideration, as my research results have pedagogical implications.

4.3 Preparation of the corpus for computer automated analysis

The manual analysis of summaries was replaced by a partially computerized analysis in the case of the comparison-and-contrast essays. These essays were written using Microsoft Word in the format required by the teacher. In order to utilize the *Concordance* software application, however, alterations were necessary in the text format and structure. First, each sentence was broken into a new line. This was necessary for the program to handle the data sentence by sentence. The title was also treated as one sentence because it was also searched for concordances. Next, misspelt words had to be corrected, because the built-in headword recognition dictionary would not have recognized them. Multi-word phrases were united by placing a hyphen between them, otherwise the program would have counted each word separately. For instance, one essay compared *Native and Non-Native English Speaking Teachers*, later referring to them as *NESTs* and *non-NESTs*. As the original intention of the writer was to

use the abbreviation (*NEST*) as a simple repetition for *Native English Speaking Teacher*, the uploading of the four-word expression had to reflect reference to one lexical unit. This was especially important because the basis of the analysis was the 'lexical unit' (Károly, 2002), described in Section 2.3. in detail.

The essays were saved as text files (.txt) and loaded into the concordance program. Each text was annotated to retain its basic structural features such as *title* and *paragraphs*, to facilitate further research of sentences at paragraph boundaries. Tagging also contained a description of paragraphs for further analysis where possible, naming them *introductory paragraph, paragraph(s) describing similarities, paragraph(s) describing differences,* and *summary paragraph.* This step was non-compulsory, as the printed versions could also have been used to determine paragraph features.

4.4 Procedures of data analysis in Károly's (2002) analytical framework

Károly's lexical repetition analysis framework employs both quantitative and qualitative measures. Appendix A summarizes the quantitative measures used in the analysis in a table format. The methods of establishing the various types of repetitions and their frequencies and of establishing the combination of links and bonds are further explained by the sample analysis using the summary samples as a data source.

1. The four basic quantitative characteristics of the summaries were investigated (as in: **Basic measures,** Appendix A).

2. A coding matrix was created, in which cells represented possible intersentential links.

3. The types of the links were determined, and written into the matrix cells (see capitalized abbreviations in Table 3, such as S, R, SS, etc.). Table 3 indicates part of the repetition matrix for Text 3 as an example, with the repetition links itemized and classified.

		S1	S2
S1	*Ryanair- Ryanair* SR		
S2		*airline- airline* SR, *company- carrier* SS, *low-cost - low-cost* SR	
S3	*Ryanair- Ryanair* SR	*Ryanair - Ryanair* SR, *runs -operates* SS	*Ryanair - airline* SS

Table 3: A detail of the repetition matrix of Text 3, itemized and classified
Abbreviations: SR: simple repetition, SS: simple synonym, 0: the title, S1, S2: sentences

4. Another matrix was drawn, indicating the number of links in each cell. Cells with three or more links (=bonds) were shaded accordingly (Figure 3).

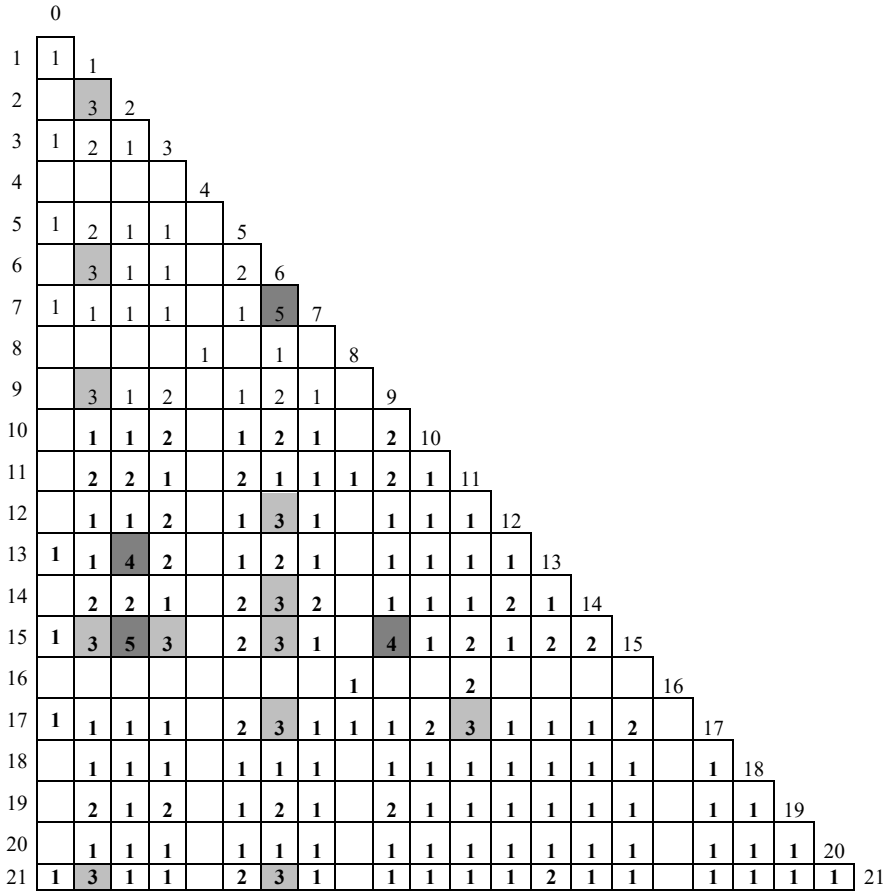


Figure 3: The repetition matrix of Text 3, indicating the number of links.

5. A table was created to indicate the position and direction of bonds. Table 4 shows the number of bonds pointing backward and forward within Text 3.

Sentence	Number of bonds pointing backward and forward	The bonded sentences (No. of links in brackets)
0 (title)	- ; 0	1-2 (3)
1	0 ; 5	1-6 (3)
2	1 ; 2	1-9 (3)
3	0 ; 1	1-15 (3)
4	0 ; 0	1-21 (3)
5	0 ; 0	2-13 (4)
6	1 ; 6	2-15 (3)
7	1 ; 0	3-15 (3)
8	0 ; 0	6-7 (5)
9	1 ; 1	6-12 (3)
10	0 ; 0	6-14 (3)
11	0 ; 1	6-15 (3)
12	1 ; 0	6-17 (3)
13	1 ; 0	6-21 (3)
14	1 ; 0	9-15 (4)
15	5 ; 0	11-17 (3)
16	0 ; 0	
17	2 ; 0	
18	0 ; 0	

19	0 ; 0	
20	0 ; 0	
21	2 ; -	

Table 4: Bonded sentences in Text 3

6. The span of bonds and the cumulative bond span had to be determined, as illustrated in Figure 4 below.

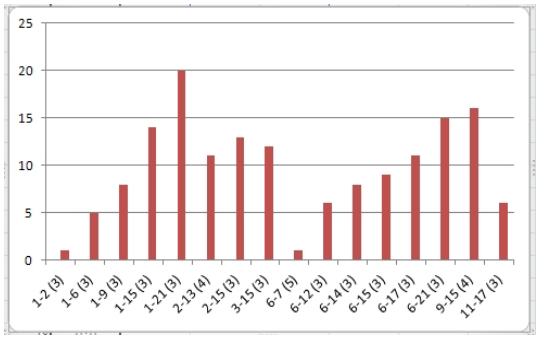


Figure 4: The span of bonds in Text 3

Figure 4 indicates the bond span in Text 3. The first two numbers (e.g., 1-2) show the two sentences connected by links, and the number in brackets shows the number of bonds connecting the sentences. The shortest span involves two adjacent sentences (e.g. sentences 1-2), and the longest span is between the first and the last sentence of the text (sentences 1 and 21).

7. The strength of connection was determined and illustrated as shown in Figure 5.

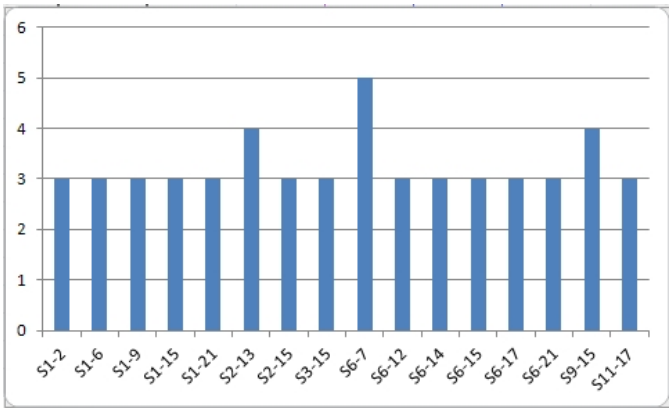


Figure 5: The strength of connection between bonded sentences in Text 3

Figure 5 shows that in Text 3 the average connection between bonded sentences consists of three links. The strongest connections are between sentences 6 and 7, with five links.

8. After the above described procedures had been repeated with each summary, the frequencies and ratios were calculated (Appendix A: **Measures related to repetition type**). As the number of sentences differed in the summaries, frequency counts were necessary. The number of each repetition type was divided by the number of sentences.

9. According to **Measures related to the combination of links and bonds**, the frequency of links and bonds were analyzed. After each text was studied, the criterion of significant connection was determined and set to be three links. (The largest number of links was 5.)

4.5 Establishing the various types of repetitions and their frequencies using the concordance program

First, the basic measures were calculated automatically by the concordance program: *number of words (types), number of words (tokens), type/token ratio, number of sentences*, out of which only the *number of sentences* and *number of tokens* data were used in this study. Next, a full search was carried out on concordances in order to establish simple repetitions. Although the program contained a built-in lemmatization tool, it distinguished between lemmas as types and tokens, and made no distinction between inflections and derivations, therefore it could not be used to distinguish between simple or derived repetitions. As the program contained no built-in parser, thorough examination was necessary to distinguish between cases such as *uses* – plural noun vs. *uses* – present tense verb, third person singular. This step had to be taken so as not to distort the simple /derived repetition ratio in the analysis.

4.5.1 Types of concordances eliminated from the headword lists

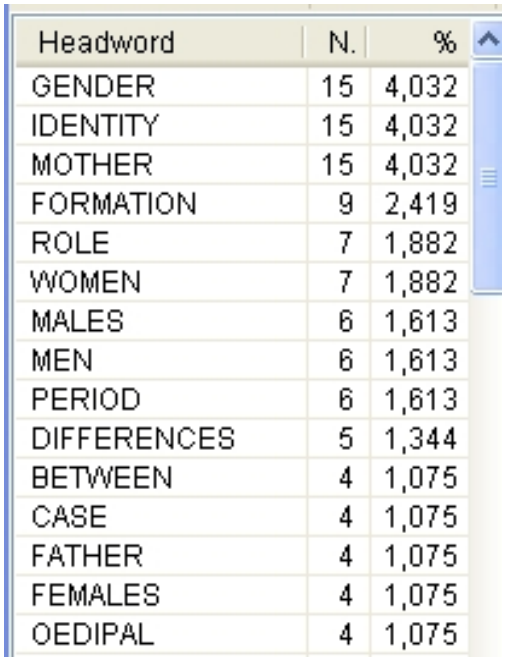
- Concordances of non-content (grammatical) words, as they were not within the scope of the examination. They were ignored by the program by loading them into the 'stop list', an inside dictionary which can be extended by the user.
- Nouns which were part of a discourse organizer phrase (lexical bundle), whenever they served the purpose of conjunctions in the paragraphs were also eliminated, such as *hand* in the expression of *on the one hand / on the other hand*. In the same fashion, when the noun *summary* was part of the introductory phrase *in summary* in the conclusion paragraph, it was eliminated from the wordlist however, when the concordance *summary /summarize* appeared as content word-pairs in one of the essays, it was treated as relevant to the analysis and kept as part of the list.
- According to Hoey's (1991) and Károly's (2002) analysis, concordances within sentences (repetition inside sentences) do not contribute to the organizing function of lexical cohesion of texts, therefore, these concordances were ignored
- In the cases of non-integral citation, the name of the author and the date was deleted, but the name was kept when an integral citation was used. The reason for this was that in non-integral citations the author and the date were indicated as additional information (in brackets), and not as an integral part of the sentence, whereas the integrally cited author could have been part of an instantial repetition link.

4.5.2 Types of concordances added to the headword lists

- inflected words, such as singular and plural forms of the same noun (regular and irregular forms: *situation, situations, man, men*),
- possessive cases in singular and plural (*child, child's*),
- verbs conjugated (third person singular and plural, simple present and simple past tense forms, regular and irregular).

4.5.3 Listing concordances (**Measures related to repetition type**, Appendix A.):

First, simple repetition frequency lists were drawn up. Figure 6 indicates the wordlist counted by the program, organized according to frequency. It this text the words *gender, identity* and *mother* appeared most.



Headword	N.	%
GENDER	15	4,032
IDENTITY	15	4,032
MOTHER	15	4,032
FORMATION	9	2,419
ROLE	7	1,882
WOMEN	7	1,882
MALES	6	1,613
MEN	6	1,613
PERIOD	6	1,613
DIFFERENCES	5	1,344
BETWEEN	4	1,075
CASE	4	1,075
FATHER	4	1,075
FEMALES	4	1,075
OEDIPAL	4	1,075

Figure 6: A detail of the headword frequency list (Text 3).
Headwords are listed according to occurrence. (N= number, % = the percentage of the occurrence in the text)

The chart in Figure 7 (next page) was prepared manually to show the sentential position of the words. Columns B-Z indicate the sentences in which the given headword occurs. Number 1 represents the title. With the help of the table, conclusions can be drawn for all types of repetition links. This way intrasentential links, such as the simple synonyms of *men* and *male* (Rows 4 and 6) can be eliminated from the link-count, as they both occur in sentence 23.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	language/s	1		3						9						15	16	17			20	21		23	24	25
2	gender/s																		18	19			22			
3	women, women's		2			5		7	8			11		13		15								23		25
4	men, men's		2				6	7	8						14	15								23		25
5	female	1		3						9											20			23	24	
6	male	1		3						9											20			23	24	
7	example		2																18					23		
8	devices											11	12	13	14					19						
9	hedging											11			14											
10	Japanese																16	17	18	19						
11	speaker																			19			22			
12	talking					5					10															
13	speech		2		4																			23		
14	talk						6	7	8		10															
15	difference/s	1			4											15	16	17			20	21		23		25
16	different		2	3															18						24	
17	forms																			19		21	22			
18	lexical															15		17			20					
19	morphological															15		17			20					
20	use/d											11	12	13	14				18							
21	aspects		2	3																						
22	cars						6	7														21	22			
23	certain																									
24	choice															15									24	
25	computers						6	7																		
26	topic/s				4			7								15									24	
27	modern																16	17								
28	person										10															
29	phonetical																				20			23		
30	relationships					5		7																		
31	distictions															15					20		22			
32	verb																					21	22			
33	explains											11										21				
34	states					5					10															
35	prefer/red				4	5																				
36	preference/s		2					7																		
37	rule/s								8														22			

Figure 7. Another detail of the headword frequency list (Text 3).
The numbers represent the sentences in which the words occur. (No.1 = the title)

For other lexical repetition types, a "classic" coding matrix was created, in which cells represented possible intersentential links. The types of the links were determined, and put into the matrix cells, where all links were itemized, and their type of repetition was identified. The results indicated by the frequency table and the matrix representing all types of repetitions were checked to give the final number of repetition links of each essay. Finally, the frequency of all repetition types were calculated, and the results were presented in a table. Establishing the combination of links and bonds followed the procedures described above in section **4.4 (**from point 4).

5 Results and discussion

In what follows, the presentation of the results is organized according to the four research questions:
(1) How can Károly's (2002) theory-based repetition analysis framework be extended to the study of summaries written by Hungarian EFL university students?
(2) How can Károly's (2002) theory-based lexical repetition analysis framework be extended to the study of comparison-and-contrast academic essays written by Hungarian EFL university students?
(3) How can a concordancing software application (*Concordance, Version 3.3)* facilitate the analysis of the text-organizing function of lexical repetition patterns?

(4) How does the examined corpus need to be prepared for a computer-assisted analysis based on Károly's lexical repetition analysis tool?

5.1 Results related to extending Károly's analytical tool in summaries

Károly's analytical tool was capable of detecting structural differences between the lexical organization of high and low-rated summaries. According to the results, high-rated summaries contain more repetition links in general. A tendency can be observed towards more frequent use of simple repetition, simple synonymy, and derived opposites in high-rated summaries. This group also contains a higher number of derived repetitions. The mean frequency of derived repetitions is **0.08** in low-rated summaries, while it is as high as **0.434** in high-rated summaries. There is also a difference between the two groups in the usage of instantial relations, with high-rated summaries containing more. The results imply that there is a clear difference between the two groups concerning same unit repetition (SUR) because the average frequency in low-rated summaries is **2.71**, whereas in high-rated summaries it is more than double (**5.45**). The average frequency of different unit repetition (DUR) in low-rated texts is **2.042**, and in high-rated summaries it is **3.658**, as indicated in Table 6 below.

Code	SUR	DUR
L1	1.6	1.45
L2	3.08	2.58
L3	1.6	1.58
L4	3.03	2.13
L5	4.24	2.47
mean frequency	2.71	2.042
H1	3.39	2.26
H2	5.71	2.28
H3	3.06	2.46
H4	9.45	6.31
H5	5.66	4.98
mean frequency	5.45	3.658

Table 6: The difference between the mean frequencies of SUR and DUR in high-and low-rated summaries
Abbreviations: SUR: Same unit repetition, DUR: Different unit repetition

These results are in line with the outcome of Károly's (2002) lexical repetition analysis on argumentative essays, the results of which indicated that high-rated essays employed more same unit repetition, particularly derived repetition, simple opposites and instantial relations.

Results of **Measures related to the combination of links and bonds** (Appendix A) slightly differed from results of Károly's (2002) original study on argumentative essays. In the case of argumentative essays, links and bonds of the title with the body text were good indicators of a well-organized structure, whereas in the present study, no bonds were found between the title and the body text in nine of the ten summaries. The reason was that the writers of these summaries used a single word, the company name, as the title (*Ryanair*), therefore, the title contained only one link with any of the sentences. The one summary title which consisted of more words (*Ryanair, Europe's largest low-cost carrier*) created only one bond with the first

sentence. Therefore, no conclusions could be drawn on this aspect of the summaries. Moreover, as the structure of the summaries did not contain a thesis statement similar to one characteristic of argumentative essays, this aspect could not be analyzed either. This meant that five of Károly's (2002) measures related to the combination of bonds could not be investigated due to structural differences between summaries and argumentative essays. (see Appendix A. **Measures 3**.)

The frequency of links and bonds, as well as the density of bonds were calculated, as indicated in Table 7. For calculating the frequency of links, the number of links was divided by the number of sentences. For calculating the frequency of bonds the same was repeated. The number of bonds was divided by the number of cells to calculate the density of bonds.

Code	Frequency of links	Frequency of bonds	Density of bonds
L1	3.36	0.05	0.008
L2	5.55	0.055	0.002
L3	3.2	0.06	0.008
L4	5.24	0.034	0.002
L5	6.88	0.05	0.006
mean	4.846	0.05	0.0052
H1	5.46	0.266	0.033
H2	8	0.71	0.09
H3	5.53	0.266	0.033
H4	12.7	0.9	0.056
H5	10.66	0.76	0.69
mean	8.47	0.58	0.18

Table 7: Frequency of links and bonds and density of bonds

As Table 7 shows, in the case of the investigated summaries, several differences can be observed in both frequency and density. In all three categories, high-rated summaries contained higher values. In high-rated texts, the frequency of links and bonds, as well as the density of bonds were higher. A reason for this might be that good summaries contained more 'compact' sentences because the writers had arranged the information from the original text in a logical order, with introductory sentences. This also indicates that even though not all summaries were organized into paragraphs (in some cases no classic paragraph boundary sentences could be observed), there were still sentences with information content which could be considered as topic sentences.

A qualitative investigation into the location of the bonded sentences showed that in high-rated summaries these were more likely to appear in the first three-quarters of the texts. This could have been caused by the type of information the students had to collate (see Appendix A), as the task students were given consisted of organizing two types of information. The first topic they had to summarize (basic information on the company) involved several facts and figures, where listings and bullet point arrangements were also permitted in textual representation, which was in fact the method chosen in low-rated summaries. This meant fewer cohesive devices, as adjacent sentences did not rely on each other. Good summaries, on the other hand, contained this information with introductory sentences followed by lists.

To illustrate the different structuring of important information, two examples are given.
(1) He introduced quick turn-around times, 'no frills' and no business class, as well as collecting ancillary revenue. (from a low-rated summary)
(2) His measures have become models for low-fare airlines, which included quick turn-around times, no business class, only a single model of aircraft in operation and the use of regional airports instead of international ones. (from a high-rated summary)

As can be observed, both sentences contain almost the same information. The first sentence is marginal, as it has no links or bonds with any other sentences. However, if it were not present, essential information would be lost. The underlined units in the second sentence contain links to other sentences, and the sentence is also bonded to two others.

Up to this point, it is possible to conclude that bonded SDV (special discourse value) sentences contain the most important information in good summaries. Therefore, Károly's (2002) analytical tool devised for argumentative essays can be extended to distinguish between high- and low-rated summaries. It was seen in Table 7 that high-rated summaries had a higher frequency of bonds and it was illustrated with the two examples that in good summaries it was bonded sentences which contained the important information.

However, there are some cases when bonding between sentences is superfluous. The following two sentences (sentences 2 and 13) are taken from the same summary.
(S2) The airline was founded in 1985 in Ireland by Christopher Ryan, and it has become today's largest low-cost carrier in Europe.
(S13) By 2003, Ryanair was among the largest carriers in Europe.

The sentences have four links in common and both are bonded with other sentences in the text. The problem is that the second sentence contains hardly any new information compared to S2, and could have been incorporated into another sentence. According to the framework, however, it is included in the net of bonding, frequencies, cumulative bond span and density of bonds, adding to their values. It might be possible that another type of content-based element should be incorporated into the present framework in order to investigate the flow of the information content and to filter unnecessary repetitions between sentences.

One low-rated summary was of particular interest in this sense, considering that it used an unusual number of lexical redundancy intrasententially. This redundancy was caused by using a very high number of equivalences, where the same referent (*Ryanair*) was circumlocuted in various ways (*company, airline, carrier, firm* and wrongly **aircraft*) within sentences. This type of repetition is not observed within the present framework, as only intersentential repetitions are examined. It was, however, noticeable to raters and was therefore included in the pedagogical evaluation of this summary.

Another doubt concerning the importance of central sentences is based on Tyler's (1995) critique of Hoey (1991), claiming that several marginal sentences were essential in a good summary of a given text. This investigation showed that this might be true when information in the original text is listed and appears only once, but has considerable content value, and without which the information of the original would be partially lost. Therefore, such sentences are not to be considered merely holding additional information. According to Tyler's (1995) findings, cited

also by Károly (2002), 'the articulation of relevant concepts is the property that makes a sentence peripheral or central to the text, not the number of lexical repetitions or links it contains' (p. 272). If there had been a strict word limit for the summaries, the one containing sentence (1) might have been rated higher than (2), see illustration in this section above.

The values of the final two measures: the mean of the cumulative bond span and the strength of bonds were higher in high-rated summaries than in low-rated ones. However, as one of the low-rated texts contained no bonds, and two summaries contained one bond, the mean values provide limited information concerning the quality of texts. The sample summary in the Method section should not be considered typical, insofar as it contained the highest number of bonds. The mean value of the strength of bonds was 3.4 for high-rated summaries, which indicates that three links was the average to connect bonded sentences. This figure is less than two (1.8) in low-rated summaries, and can be explained by Table 7, as both the frequency of links and the frequency of bonds are lower.

5.2 Results related to extending Károly's analytical tool to comparison-and-contrast essays

As Table 8 shows, all of the essays contain various repetition types. In line with other previous research results (Károly, 2002, Hoey, 1991), the most frequent type is simple repetition, but derived repetition is also frequent. Furthermore, a clear tendency can be observed towards the more frequent use of simple synonyms and simple opposites, as opposed to derived ones. In some cases, the ratio of repeated words is surprisingly high, especially if we consider the fact that only content words were calculated in the study and grammatical words were entirely ignored. Hyland (2006) explains this phenomenon by noting that a high proportion of content words in relation to grammar words is characteristic of the academic register, thus adding to the high lexical density of such texts.

code	Frequency of types of repetition										
	Same unit repetition			Different unit repetition							
			SUR	Synonymy		Opposites		Hyponymy	Meronymy	Instantial relations	DUR
	Simple	Derived		Simple	Derived	Simple	Derived				
H1	0.35	0.67	**1.02**	0.03	0.02	0.06	0.007	0.02	0	0	**0.137**
H2	0.49	0.19	**0.68**	0.06	0.01	0.05	0.05	0.12	0	0.002	**0.292**
H3	0.62	0.07	**0.69**	0.02	0.01	0.01	0.01	0.02	0	0	**0.07**
H4	0.5	0.03	**0.53**	0.04	0.1	0.06	0.07	0.06	0	0	**0.33**
L1	0.38	0.02	**0.4**	0.009	0.008	0.02	0.006	0.02	0	0	**0.063**
L2	0.47	0.08	**0.55**	0.039	0.2	0.066	0.05	0.02	0.001	0.066	**0.442**
L3	0.48	0.01	**0.49**	0.042	0.07	0.03	0.008	0.033	0	0	**0.183**
L4	0.45	0.05	**0.5**	0.02	0.2	0.066	0.066	0.04	0	0	**0.356**

Table 8: The frequency of types of repetition in high- and low-rated essays. Abbreviations: SUR: Same unit repetition, DUR: Different unit repetition

Another feature of all the essays is the relatively high number of hyponyms and hyperonyms used, which could also be a feature of their academic-related content. The essays

compare abstract academic topics such as gender identity formation or native language acquisition models and therefore often employ clarifications or give definitions of terms. An initial assumption would suggest that hyperonyms appear in the introductory paragraph, where the topic is explained and in the final paragraph, where conclusions are drawn and ideas summarized. No clear evidence of this is visible however, rather a scattered positioning of hyponyms and hyperonyms can be observed. A closer investigation reveals that this type of repetition appears in initial and closing sentences within paragraphs. Although high-rated essays contain slightly more repetitions than low-rated ones, Károly's (2002) previous study revealed that quantitative measures alone cannot predict textual quality.

Results in relation to the combination of links and bonds indicated that both high-rated and low-rated essays contain more bonded sentences than marginal sentences, but high-rated essays have a higher frequency of bonds. The ratio of marginal/bonded sentences also differs between the two groups (the average ratio is 0.3 and 0.215 in high and low rated essays, respectively). The main bond-related differences fall into two areas: the relative use of bonds at paragraph boundaries (especially in the introductory and concluding paragraphs), and in the span of bonds between sentences. High-rated essays connect paragraph-initial and paragraph-final sentences more frequently to each other, providing a more structured framework to the topic explained within the paragraph. Furthermore, lexical links of the topic sentences in the introductory paragraph reappear in high-rated essays in one of the concluding sentences, typically in the opening sentence of the concluding paragraph. Figure 8 indicates the location of bonded sentence-pairs in the highest-rated essay. The two numbers (e.g., 2-3) show the two sentences connected by bonds. It also reveals the span of bonds between sentences, which means how far apart the connected sentences are located from each other in the text. The shortest span involves two adjacent sentences (e.g., Sentences 2-3), and the longest span is between the first and the last sentence of the summary (Sentences 2 and 25).

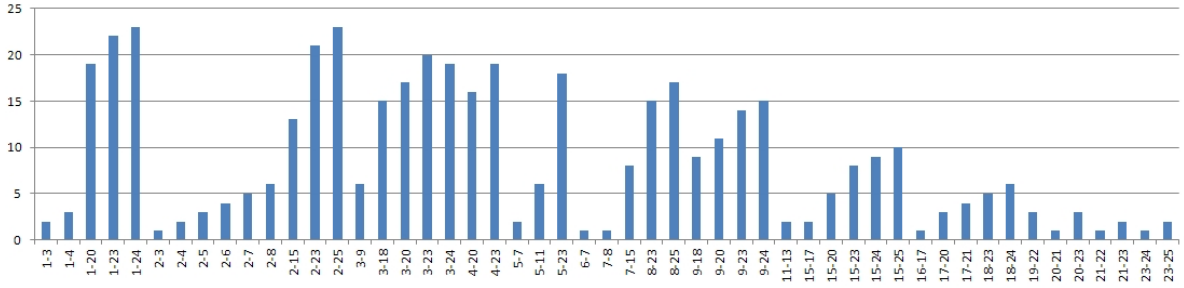


Figure 8: The span of bonds in Text 1. The two numbers (e.g. 2-3) show the two sentences connected by bonds.

Seven essays contain bonds between the title and the body of the text, however, high-rated essays contain more bonds (average = 4.3). The highest number of bonds is 7, while one low-rated essay contains only 2 links between the title and the essay but no bonds at all. Figure 8 indicates that Text 1 has 5 bonds with the title, 2 with introductory sentences (sentences 3 and 4, and 3 with sentences towards the end of the essay, one of which is the summary statement, No. 23.). Sentences No. 2 and 23 are the key sentences, the former having 9 bonds pointing forward, and the latter containing 11 bonds referring backward, providing a structural framework for the essay. Therefore, it can be argued that there is a relationship between the position of bonded

sentences and the span of bonds, and together they influence text quality. In the case of low-rated essays, the tendency described above is not as clearly represented.

5.3 Using the Concordance program for analyzing lexical repetition patterns

The third research question referred to how the Concordance Text Analysis program can facilitate researching the text-organizing function of lexical repetition patterns. It certainly assisted in organizing the vast amount of data collected from the corpus, especially in the first phase of the analysis when the number of words, sentences and paragraphs were counted. The program, if used in further research, might also add to the depth of the analysis with a new element, namely establishing the sentence density or, as the program defines it: *the density of words per sentence*. The Sorting Lemmatized Headwords function also considerably reduced the number of overlooked concordances.

Unfortunately, there were limitations to the program with regard to counting concordances within the scope of the current research, for three reasons. Firstly, the specific categories of Károly's analytical tool did not exactly match the categories of the concordancer. Especially problematic was the type/token count instead of the inflection/derivation distinction required by the present analytical tool. Secondly, the table of links, which was drawn on the basis of the computerized headword frequency count (Table 3), although served as a good means of visual representation for the simple repetition links, in fact doubled the workload. This was due to the fact that it could not represent any other repetition types. As a consequence, these had to be itemized and classified in a cell matrix, as well as collated in another matrix in a number format. (The latter two matrices were used by Hoey (1991) and Károly (2002), and also in Adorján (2011) as a compulsory element of the analysis.) Thirdly, the program was not suitable for establishing bonds between sentences.

5.4 Preparation of the corpus for computer-based analysis

First, the steps necessary for the concordancing software to collect data from the essays were repeated, as detailed in 4.3. Subsequently, the **Basic measures** and **Measures related to repetition type** (Appendix A) were quantitatively analyzed for each text, when certain areas proved to be problematic from a coding perspective. The first decision concerned the identification of lexical unit boundaries. As lexical units are the basic unit of analysis, careful investigation was necessary to find what belongs to one lexical unit. The case of *Native and Non-Native English Speaking Teachers*, mentioned in Section 2.4 previously, was an extreme example because it involved 6 (or possibly 7) orthographic words. Other examples which also required individual decisions were highlighted in Section 4.4.1 (e.g., the nouns *hand* and *summary).* Hyland (2012) also reports similar concerns regarding the subjectivity in the treatment of such recurrent word sequences in discourse analysis. The unique nature of each text would require individual decisions from the analyst, while it is counterproductive from a computer automation perspective.

Conclusions

This paper aimed to investigate Károly's (2002) theory-based analytical tool from two perspectives. Two research questions focused on whether the tool can be applied to two other

academic genres: summaries and comparison-and-contrast essays, whereas two further questions were concerned with research methodology. Károly's (2002) repetition model, together with a content-based approach, was capable of detecting structural differences between high-rated and low-rated texts in both genres. Particularly informative were the analytical steps related to the position and span of bonded sentences. Results indicated that there is a relationship between the position of bonded sentences and the span of bonds, and together they influence text quality.

Conclusions based on results regarding the usage of computerized data analysis indicate that Concordance, Version 3.3 is unable to capture certain important features of the framework, for example, except for reiterations, it is unable to locate other lexical repetition categories. Furthermore, the application is unable to store data or collate data into charts or matrices. Specific categories of the present analytical framework require the development of a software application with matching categories to enable research using large corpora.

Further research in the following areas is necessary: (1) how to identify lexical units reliably, (2) how to ensure disambiguation of meaning before links are connected, (3) how to establish a satisfactory treatment of instantial relations.

Proofread for the use of English by Sean Murphy (freelance)

References:

Adorján, M. (2011a). *Predicting rater appeal of summaries based on Károly's (2002) study of the text-organizing role of lexical repetition* (Unpublished seminar paper). Language Pedagogy PhD Program, Eötvös Loránd University, Budapest, Hungary.

Adorján, M. (2011b). *Can lexical repetition patterns predict perceived discourse quality? An exploratory study on EFL comparison-and-contrast essays* (Unpublished seminar paper). Language Pedagogy PhD program, Eötvös Loránd University, Budapest, Hungary.

Birmingham City University, Research and Development Unit for English: SHARES. System of Hypermatrix Analysis, Retrieval, Evaluation and Summarisation [Computer software]. *Birmingham City University, Research and Development Unit for English Studies Website:* http://rdues.bcu.ac.uk/sharesguide/index.shtml

Collier, A. (1994). *A System for automating concordance line selection.* Retrieved May 5, 2011, from Birmingham City University, Research and Development Unit for English Studies Website: 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. *Papers in Linguistics: International Journal of Human Communication, 17,* 301-316. https://doi.org/10.1080/08351818409389208

de Beaugrande, R. (1997). *New foundations for a science of text and discourse: Cognition, communication, and freedom of access to knowledge and society.* New Jersey: Ablex Publishing Corporation.

de Oliveira, P. C. F., Ahmad, K., & Gillam, L. (1996). *A Financial News summarisation system based on lexical cohesion.* Retrieved November 5, 2011 from http://antigo.univille.br/arquivos/4694_Paper3.pdf

Graesser, A., McNamara, D., & Louwerse, M. (2011). Methods of automated text analysis. In M. L. Kamil, D. Pearson, E. Moje, & P. Afflerbach (Eds.), *Handbook of reading research. Volume IV* (pp. 34-54). New York: Routledge.

Halliday, M. A. K. (1985). *Spoken and written language.* Oxford: Oxford University Press.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* London: Longman.

Hammann, L. A., & Stevens, R. (2003). Instructional approaches to improving students' writing of compare-contrast essays: An experimental study. *Journal of Literacy Research, 35*(2), 731-756. https://doi.org/10.1207/s15548430jlr3502_3

Hasan, R. (1984). Coherence and cohesive harmony. In J. Flood (Ed.), *Understanding reading comprehension* (pp. 181-219). Delaware: International Reading Association.

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of Educational Research, 56*(4), 473-493. https://doi.org/10.2307/1170342

Hoey, M. (1991). *Patterns of lexis in text.* Oxford: Oxford University Press.

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Hyland, K. (2006). *English for academic purposes. An advanced resource book.* London: Routledge.

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics, 32,* 150-169. https://doi.org/10.1017/s0267190512000037

Károly, K. (2002). *Lexical repetition in text.* Frankfurt am Main: Peter Lang.

Laki, L. J. (2011). *Statisztikai gépi fordítási módszereken alapuló egynyelvű szövegelemző rendszer és szótövesítő* [Monolingual text analyzer and lemmatizer based on statistical computational methods.] Paper presented at the 8th Hungarian Computational Linguistics Conference, Szeged, Hungary.

Miller, G. (1995). WordNet: A Lexical Database for English. [Computer software]. Retreived in November, 2011 from http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=8106

Myers, G. (1991). Lexical cohesion and specialized knowledge in science and popular science texts. *Discourse Processes, 14*(1), 1–26. https://doi.org/10.1080/01638539109544772

Novák, A., Orosz, Gy., & Indig, B. (2011*). Javában taggelünk.* [We are a-tagging.] Paper presented at the 8th Hungarian Computational Linguistics Conference, Szeged, Hungary.

Prószéky, G., & Mihaltz, M. (2008). Magyar WordNet: az első magyar lexikális szemantikai adatbázis. [HunWordNet: the first Hungarian lexical semantic database]. *Magyar Terminológia, 1*(1), 43-57.

Reynolds, D. W. (1995). Repetition in nonnative speaker writing: More than quantity. *Studies in Second Language Acquisition, 17*(2), 185-209.

Reynolds, D. W. (2001). Language in the balance: Lexical repetition as a function of topic, cultural background, and writing development. *Language Learning, 51*(3), 437-476. https://doi.org/10.1111/0023-8333.00161

Seidl-Péch, O. (2011). Fordított szövegek számítógépes összevetése. [Computerized comparison of translated texts] (Doctoral dissertation). Eötvös Loránd University, Budapest.

Sinclair, J. (1998). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Spivey, N. N. (1990). Transforming texts: Constructive processes in reading and writing. *Written Communication, 7,* 256-287.

Teich, E., & Fankhauser, P. (2004). *Exploring lexical patterns in text: Lexical cohesion analysis withWordNet*. Retrieved in November, 2011, from http://www.sfb632.uni-potsdam.de/publications/isis02_7teich-fankhauser.pdf

Tyler, A. (1992). Discourse structure and specification of relationships. A crosslinguistic analysis. *Text, 12*(1), 671-688. https://doi.org/10.1515/text.1.1992.12.1.1

Tyler, A. (1994). The role of repetition in perceptions of discourse coherence. *Journal of Pragmatics*, *21*, 671-688. https://doi.org/10.1016/0378-2166(94)90103-1

Tyler, A. (1995). Co-constructing miscommunication: The role of participant frame and schema in cross-cultural miscommunication. *Studies in Second Language Acquisition, 17,* 129-152.

van Dijk, T. A. (1977). *Text and context.* London: Longman.

Watt, R. J. C. (1999-2009). Concordance (Version 3.3 July, 2003) [Computer software]. Retrieved in November, 2011, from http://www.concordancesoftware.co.uk/concordance-software-download.htm

Widdowson, H. (1978). *Teaching language as communication*. Oxford: Oxford University Press.

Widdowson, H. (1996). *Linguistics.* Oxford: Oxford University Press.

APPENDIX A

The summary task

1. Please check the information available about Ryanair on the Internet: http://en.wikipedia.org/wiki/Ryanair

2. Imagine that you work for an unsuccessful airline company as a tourism and travel expert. Your boss admires the management of Ryanair and wants to copy its profitable business model, therefore he asks you to summarize Ryanair's history and business model, highlighting its strengths and weaknesses.

Here are some key points for you to mention:

1. Basic information about the company:
   - foundation (when, where, founder, manager)
   - key dates and events
   - number of aircraft, flights, passengers
2. Political changes (domestic, British, EU) and their effects on the company
3. O'Leary's decisions and their effects (good and controversial ones)

Write at least one page but no more than two (double spaced, Times New Roman 12 or similar).

Possible words and expressions to use: revenues, operating expenses, turn-around times, outsourcing, increase, expansion, optional extras

APPENDIX B

Quantitative measures in the analysis (Károly, 2002, p. 144)

1. Basic measures:	number of sentences number of paragraphs number of links number of cells
2. Measures related to repetition type:	frequency of simple repetition derived repetition same unit repetition simple synonymy derived synonymy synonymy (simple and derived) simple opposites derived opposites hyponymy meronymy instantial relations different unit repetition ratio of same unit repetition to different unit repetition ratio of simple repetition to derived repetition
3. Measures related to the combination of links and bonds:	frequency of links frequency of bonds density of bonds frequency of adjacent bonds frequency of non-adjacent bonds cumulative bond span frequency of central sentences frequency of marginal sentences relative use of bonds at paragraph boundary strength of connection (1-8 links) bonds between: title & essay title & thesis statement title & topic sentences thesis statement & topic sentences thesis statement & essay