

ÇANKAYA UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
COMPUTER ENGINEERING

MASTER THESIS

TEXT COHERENCE IN TURKISH
VIA
LATENT SEMANTIC ANALYSIS

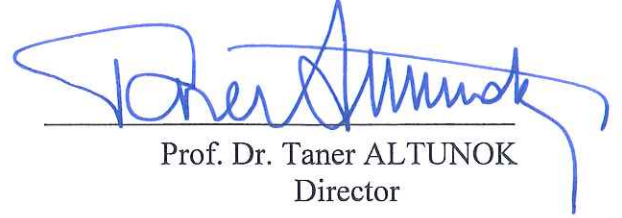
İBRAHİM KIŞLACIK

JUNE 2013

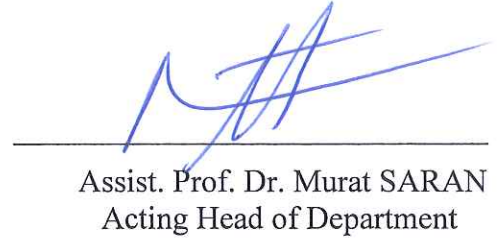
Title of the Thesis : **Text Coherence in Turkish via Latent Semantic Anlysis**

Submitted by : **İbrahim Kışlacık**

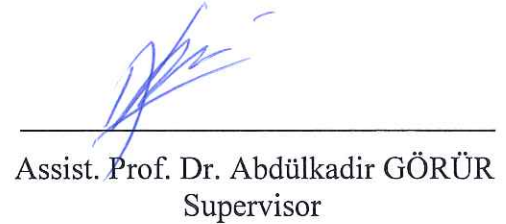
Approval of the Graduate School of Natural and Applied Sciences, Çankaya University.


Prof. Dr. Taner ALTUNOK
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.


Assist. Prof. Dr. Murat SARAN
Acting Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.


Assist. Prof. Dr. Abdülkadir GÖRÜR
Supervisor

Examination Date: 10.06.2013

Examining Committee Members:

Assist. Prof. Dr. Abdülkadir GÖRÜR

(Çankaya Univ.)

Assist. Prof. Dr. Bülent Gürsel EMİROĞLU

(Başkent Univ.)

Assist. Prof. Dr. Reza ZARE HASSANPOUR

(Çankaya Univ.)

STATEMENT OF NON PLAGIARISM

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : İbrahim Kışlacık

Signature :



Date

: 10.06.2013

ABSTRACT

TEXT COHERENCE IN TURKISH VIA LATENT SEMANTIC ANALYSIS

KIŞLACIK, İbrahim

M.Sc., Department of Computer Engineering

Supervisor: Assist. Prof. Dr. Abdülkadir GÖRÜR

June 2013, 47 pages

This study describes the use of Latent Semantic Analysis (LSA) for semantic similarity calculation to measure textual coherence and using this on Turkish documents to identify gender and the name of the authors. This study also provides an example application on how dimension reduction method and sliding window procedure affect the result of coherence measurement.

Keywords: Text Coherence Measurement, Latent Semantic Analysis, Gender Identification, Author Identification

ÖZ

**GİZİL ANLAMSAK ANALİZ YÖNTEMİ
İLE
METİN TUTARLILIĞI ÖLÇME**

KIŞLACIK, İbrahim

Yükseklisans, Bilgisayar Mühendisliđi Anabilim Dalı

Tez Yöneticisi : Yrd.Doç.Dr. Abdül Kadir GÖRÜR

Haziran 2013, 47 sayfa

Bu çalışmada, Gizil Anlamsal Analiz yöntemi kullanılarak Türkçe dökümanlar için metin tutarlılığı değeri ölçme ve ölçülen tutarlılık değeri kullanılarak yazar ve yazar cinsiyeti tanıma amaçlanmıştır. Aynı zamanda bu çalışma, metin tutarlılığı ölçümünde kullanılan anlamsal uzay boyutu azaltımı ve kayan pencere yöntemlerinin, metin tutarlılık değerleri üzerindeki etkisine örnek teşkil etmektedir.

Anahtar Kelimeler: Metin Tutarlılığı Ölçme, Gizil Anlamsal Analiz, Cinsiyet Tanıma, Yazar Tanıma

ACKNOWLEDGMENTS

I wish to express my deepest gratitude to my supervisor Assist. Prof. Dr. Abdül Kadir Görür for his patience, motivation, and significant comments. His guidance helped me in all the time of research and writing of this thesis. Besides my supervisor, I would like to thank the rest of my thesis committee: Assist. Prof. Dr. Reza Zare Hassanpour, and Assist. Prof. Dr. Bülent Gürsel Emirođlu for their insightful comments.

I would also like to thank my family: my parents, for supporting me spiritually throughout my life, and my engaged Şeyma, for her motivation and patience, and all of my friends, especially Yılmaz Atalar, Hasan Kavlak and Seçkin Dikbayır for their supports, and Dirisoft Family for their helps.

TABLE OF CONTENTS

STATEMENT OF NON PLAGIARISM.....	iii
ABSTRACT.....	iv
ÖZ.....	v
ACKNOWLEDGMENTS	vi
TABLE OF CONTENTS.....	vii
LIST OF TABLES	ix
LIST OF FIGURES	x
LIST OF SYMBOLS/ABBREVIATIONS	xii
CHAPTERS:	
1 INTRODUCTION	1
1.1 MOTIVATION.....	1
1.2 THESIS OUTLINE.....	3
2 COHERENCE MODELS AND RELATED WORKS	4
2.1 WHAT IS COHERENCE?	4
2.2 COHERENCE AND COHESION.....	5
2.3 APPLICATIONS OF COHERENCE	9
2.4 COHERENCE MODELS	11
2.5 COHERENCE MEASUREMENT APPROACHES AND RELATED WORKS	12
2.5.1 Vector Space Models	12
2.5.2 Coh-Metrix Tool	15
2.5.3 Centering Theory Approach.....	17
2.5.4 Other Approaches	19
2.5.5 Summary	19
3 LATENT SEMANTIC ANALYSIS	20
3.1 LSA SEMANTIC SPACE CONSTRUCTION	21
3.1.1 Segmentation.....	21
3.1.2 Stop-word Filtering.....	21
3.1.3 Stemming	22

3.1.4	Term Reduction.....	22
3.1.5	Term Weighting and Term-Document Matrix Creation	23
3.2	SVD COMPUTATION AND DIMENSION REDUCTION	24
3.3	QUERY MATCHING	26
3.4	SIMILARITY MEASUREMENT	27
3.5	LSA EXAMPLE	28
4	DATA AND EXPERIMENTAL METHODOLOGY	32
4.1	CONSTRUCTION OF LSA SEMANTIC SPACE	33
4.2	COHERENCE OF ARTICLES	35
4.3	RESULTS	39
5	CONCLUSION.....	47
	REFERENCES	R1
	APPENDICES:	
	APPENDIX A.....	A1
	APPENDIX B	A2
	APPENDIX C	A4

LIST OF TABLES

Table 1 - Example Sentences for The Four Conditions of Cohesion and Coherence	9
Table 2 - Small Term-By-Document Matrix	13
Table 3 - Centering Theory Transitions	18
Table 4 - Gender Distribution for 400 Newspaper Articles	35
Table 5 - Category Distribution for 400 Newspaper Articles	35
Table 6 - Example of Window Size 1	36
Table 7 - Example of Window Size 5	37
Table 8 - Correctly Classify Articles for Author Identification	40
Table 9 - Correctly Classify Articles for Gender Identification.....	40

LIST OF FIGURES

Figure 1 – Representation of Documents in a 3-Dimensional Vector Space.....	13
Figure 2 – Illustration of SVD	25
Figure 3 – Singular Value Decomposition Interpretation for Rank	26
Figure 4 – Cosine Similarity in Semantic Space.....	28
Figure 5 - Documents for Using In LSA.....	29
Figure 6 - Term-By-Document Matrix After Stop-Word Removing Process	29
Figure 7 – Term-By-Document Matrix After Weighting Process	29
Figure 8 – Singular Values Decomposition Matrices	30
Figure 9 – Meaning of Terms and Documents After Similarity Measure In LSA Semantic Space.....	31
Figure 10 – Semantic Space Creation Processes	34
Figure 11 – Similarity Measure Processes for an Article	38
Figure 12 - Coherence Measures for Appendix B Article Sentences Window to Sentences Window with Three SVD Ranks, Sliding Window Size 1	40
Figure 13 - Coherence Measures for Appendix B Article Sentences Window to Sentences Window with Three SVD Ranks, Sliding Window Size 2	41
Figure 14 - Coherence Measures for Appendix B Article Sentences Window to Sentences Window with Three SVD Ranks, Sliding Window Size 3	41
Figure 15 - Coherence Measures for Appendix B Article Sentences Window to Sentences Window with Three SVD Ranks, Sliding Window Size 4	42
Figure 16 - Coherence Measures for Appendix B Article Sentences Window to Sentences Window with Three SVD Ranks, Sliding Window Size 5	42
Figure 17 - Coherence Measures for All Articles with SVD Rank 200, Sliding Window Size 1 (1-200 Female / 201-400 Male).....	43
Figure 18 - Coherence Measures for All Articles with SVD Rank 200, Sliding Window Size 3 (1-200 Female / 201-400 Male).....	43
Figure 19 - Coherence Measures for All Articles with SVD Rank 200, Sliding Window Size 5 (1-200 Female / 201-400 Male).....	44
Figure 20 - Coherence Measures with SVD Rank 100 for 400 Newspaper Articles.....	44

Figure 21 - Coherence Measures with SVD Rank 150 for 400 Newspaper Articles.....	45
Figure 22 - Coherence Measures with SVD Rank 200 for 400 Newspaper Articles.....	45
Figure 23 - Coherence Measures with SVD Rank 250 for 400 Newspaper Articles.....	46
Figure 24 - Coherence Measures with SVD Rank 300 for 400 Newspaper Articles.....	46

LIST OF SYMBOLS/ABBREVIATIONS

- LSA** : Latent Semantic Analysis
- SVD** : Singular Value Decomposition
- HMM** : Hidden Markov Model
- IR** : Information Retrieval

CHAPTER I

INTRODUCTION

The aim of this thesis is to research textual coherence measurement approaches in literature and to measure textual coherence in Turkish documents using Latent Semantic Analysis (LSA) algorithm.

In this chapter, the motivation and the aims of the thesis are given, and then the outline of the work is presented.

1.1 MOTIVATION

Technology has obtained large space in our lives as a result of easy and simple access to information. At the same time, efficient data management and intelligent data analysis methods became more important in electronic world. Human resources are inadequate and costly for these works. In order to solve this issue automatic data analysis methods are used.

Automatic textual coherence evaluation is one of the data analysis methods. Readability and understandability of texts is important for readers. The document must be coherent for having high quality. Coherence is a key concept of text linguistics. It is especially relevant to the research on text comprehension and text clarity. It is about the semantic level structure of the document. Textual coherence assessment is essential for discourse analysis, machine translation, document summarization and etc. Textual coherence can be evaluating using coherence measurement methods simpler than using human judgments. Automatic coherence measure systems bring in time, cost and efficiency.

By using text coherence measure systems; a teacher can evaluate the student's essay quality or an editor can decide and maintain the article's coherence without reading the whole document or a translator can measure the translated article's coherence. Also other systems, such as document summarization systems, search engines, machine translations can use text coherence measure tools to perform their processes more efficiently.

The automatic evaluation of text coherence is similar to the way of a person evaluates text coherence. While evaluating text coherence, a person uses own prior knowledge, but this is a challenging task for a computer system. Measuring coherence remains as a difficult task in text linguistics in natural language processing.

LSA is statistical approach technique which uses semantic space oriented analysis for measuring textual coherence. This method is used in data mining for information retrieval, data clustering, document classification, cross language retrieval and etc. LSA compares the vectors for two document or term of text in a high dimensional semantic space and provides the degree of semantic relatedness between documents or terms. In this thesis LSA uses for predicting coherence of texts.

For performing the LSA based text coherence measurement system, term-by-document matrix creation, dimension reduction and term weighting methods are used. Our assessment is based on the effectiveness of dimension reduction and sliding window method on coherence values in Turkish documents set. After measurement, we want to try to analysis gender identification and author identification using coherence values on Turkish articles.

1.2 THESIS OUTLINE

The rest of this document is organized as;

Chapter 2 presents the coherence models and related works for coherence measurement. Coherence definition and coherence measurement approaches and tools in literature are explained.

Chapter 3 presents the LSA method details.

Chapter 4 presents the data contents, text coherence measurement method and gender and author identification using LSA.

Chapter 5 presents the conclusion for this study.

CHAPTER II

COHERENCE MODELS AND RELATED WORKS

2.1 WHAT IS COHERENCE?

This thesis examines coherence of a discourse. Discourse is a written text, a spoken conversation or anything that carries information either clearly expressed or indirectly stated. Coherence depicts the process of how elements of the text combine to form a unified whole. It can be thought of as how meanings and sequences of ideas relate to each other.

Coherence in linguistics is what makes a text semantically meaningful. According to the definition given in Oxford advance learner's dictionary, "coherence is a situation in which all the parts of something fit together well". De Beaugrande and Dressler (1996) define coherence as a "continuity of senses "and "the mutual access and relevance within a configuration of concepts and relations." [9]

Givon (1993) defines coherence "Coherence is fundamentally not an objective property of the produced text. Rather, that text is a by-product of the mental processes of discourse production and discourse comprehension, which are the real loci of coherence." [16]

Weigand (2009) defines "The *Coherence Principle* accounts for the fact that we do not communicate by verbal means only. The traditional concept of coherence, which is solely based relationships between verbal textual elements, is too narrow to account for coherence in interaction. Ultimately, coherence in interaction is not established in the text but created in the minds of the interlocutors in their attempt to make sense of the different verbal, perceptual, and cognitive means at their disposal ..." [42]

The term 'cohesion' is often used synonymously with coherence as well as being used to describe a kind of coherence. Renkema (2004) refers to cohesion as something that is *discourse internal*, whereas coherence is *discourse external*. [36] In other words, cohesion refers to connections that can be made within the text, and coherence refers to connections that are made to outside the text in the 'real world'. To further complicate terminology, both ideas, cohesion that discourse internal and coherence that discourse external, are included under Coherence (capitalized C), which is the overall generalization that captures the discourse external and discourse internal under one category.

Coherence is not something that can be directly observed or measured, but something that exists behind the scenes. Coherence should not to be confused with the text's meaning; it is how the individual meanings combine and form a single unified meaning.

Hovy (1988) defines coherent text as "text in which the hearer knows how each part of the text relates to the whole; i.e., (a) the hearer knows why it is said, and (b) the hearer can relate the semantics of each part to a single overarching framework." [22] Determining coherence is not a black and white assessment. Asher and Lascarides (2003) note that "coherence is not a yes/no matter, but rather the quality of coherence can vary". [1] There is thus a degree of coherence to a discourse and not merely a simple classification of being either coherent or incoherent.

2.2 COHERENCE AND COHESION

Although coherence and cohesion have different concepts, separating them is difficult process. Coherence can be thought that the text making sense at ideas level. Cohesion can be thought that the text has more mechanical links at a language level. They can be imaged that it is likely for a text to contain sufficiently of cohesion yet little coherence. Cohesion is the grammatical and lexical links within a text or sentence that holds a text composed and gives it meaning. Cohesion is associated to the larger model of coherence. Cohesion occurs between elements within a

discourse, whereas coherence occurs between an element external to the discourse and an element within the discourse.

According to Halliday and Hasan (1976) cohesion is a semantic concept "occurs where the interpretation of some element in the discourse is dependent on that of another". [20] Cohesion deals with parallelism, narration, lexical relatedness between words, etc. Coherence deals with world-knowledge, logic and logical relations (reasoning, consequence, result, inference, induction, causation, etc.).

Weiser (1996) describes a summary about coherence and cohesion history, "Until the mid-1970s, cohesion and coherence were often used interchangeably, both referring either to a kind of vague sense of wholeness or to a more specific set of relationships definable grammatically and lexically. The work of Halliday and Hasan (1976) influenced scholars and researchers in rhetoric and composition so that, by the early 1980s, the two terms were distinguished. Cohesion is now understood to be a textual quality, attained through the use of grammatical and lexical elements that enable readers to perceive semantic relationships within and between sentences. Coherence refers to the overall consistency of a discourse its purpose, voice, content, style, form, and so on and is in part determined by readers' perceptions of texts, dependent not only on linguistic and contextual information in the texts but also on readers' abilities to draw upon other kinds of knowledge, such as cultural and intertextual knowledge." [43]

Coherence can be understood of as how meanings and orders of ideas relate to each other. Characteristic samples are problem-solution, claim-counter-claim statement-example, general-particular, and question-answer.

Cohesion can be understood of as how all grammatical links which referring to the structural content and lexical links which referring to the language contents of the part that connection one part of a discourse to another. This contains use of synonyms, grammatical references, lexical sets, time references, verb tenses, pronouns, etc. In English; for instance, 'it', 'either', and 'those' all refer to knowledge earlier stated. 'Last of all', 'after that' and 'therefore uses for help to order in a discourse. 'Although', 'also' and 'for sample' links ideas and opinions in a discourse.

Coherence is basically concerned the ways in which concepts and relations, which underlie the surface text, are linked, relevant and used, to achieve efficient communication. A concept is perspective content which can be fetched and initiated with a high degree of consistency in the mind. Relations are the links between concepts within a text, which each link identified with the concept that it connects to.

A cohesive text is created in many different ways. Halliday and Hasan (1976) classify five general classes of cohesive devices that make coherence in texts. The five types of cohesion according to Halliday and Hasan (1976) are substitution, ellipses, reference, conjunction and lexical cohesion. [20]

Substitution:

Substitution is the replacement of a word group or sentence segment by a dummy word. The reader or listener can fill in the correct element based on the preceding. It can be occurred by a noun, a verb and a clause. Example for noun “These *biscuits* are stale. Get some fresh *ones*.” [36]

Ellipsis:

Ellipses occur when words or parts of a sentence are omitted. “John went to the movies, but Steve didn't.” Part of the subordinate clause has been omitted (*go to the movies*), where the omitted part can be found in the main clause. [36]

Reference:

Reference is semantic substitution, where the dummy word is typically a pronoun. “I see John is here. *He* has not changed a bit.” The word *He* refers to *John*. [36]

Conjunction:

Conjunction is the relationship which indicates how the subsequent sentence or clause should be linked to the preceding or the following parts of the sentence. It can be defined best in the words of Cook (Cook 1989) as, the words which draw attention towards the relationships between sentences, clauses and words. It can be

occurred additive, temporal, and causal. Example for addition “*Besides* being mean, he is also hateful”. [36]

Lexical cohesion:

Lexical cohesion refers to the links between the content words such as nouns, verbs, adjectives and adverbs which are used in subsequent segments of discourse. There are two types of lexical cohesion; reiteration and collocation. The links are manifested through repetition, synonymy, hyponymy, hyperonymy, meronymy, antonymy. Example for synonymy “A conference will be held on national environmental policy. This environmental symposium will be primarily a conference dealing with water.” [36]

Basically can say, coherence means the connection of thoughts at the idea level, and cohesion means the connection of thoughts at the sentence level. A cohesive text can be incoherent or a coherent text can be incohesive. Table 1 shows examples for those conditions. This table is taken Ferstl and Cramon (2001). [12]

[1] Coherent/Incohesive	<ul style="list-style-type: none"> • Mary’s exam was about to start. The palms were sweaty. • Laura got a lot of mail today. Some friends had remembered the birthday. • Sometimes a big truck drives by the house. The dishes start to rattle. • The lights have been on since last night. The car doesn’t start.
[2] Coherent/Cohesive	<ul style="list-style-type: none"> • Mary’s exam was about to start. Therefore, her palms were sweaty. • Laura got a lot of mail today. Her friends had remembered her birthday. • Sometimes a truck drives by the house. That’s when the dishes start to rattle. • The lights have been on since last night. That’s why the car doesn’t start.
[3] Incoherent/Incohesive	<ul style="list-style-type: none"> • Laura got a lot of mail today. The palms were sweaty. • Mary’s exam was about to start. Some friends had remembered the birthday. • The lights have been on since last night. The dishes start to rattle. • Sometimes a big truck drives by the house. The car doesn’t start.
[4] Incoherent/Cohesive	<ul style="list-style-type: none"> • Laura got a lot of mail today. Therefore, her palms were sweaty. • Mary’s exam was about to start. Her friends had remembered her birthday. • The lights have been on since last night. That’s when the dishes start to rattle. • Sometimes a big truck drives by the house. That’s why the car doesn’t start

Table 1 - Example Sentences for The Four Conditions of Cohesion and Coherence

2.3 APPLICATIONS OF COHERENCE

Coherence is used in Natural Language Processing (NLP) systems which are a field of computer science, artificial intelligence and linguistics concerned with the relations between computers and natural human languages. NLP contains as systems discourse analysis, natural language generation, machine translation, automatic

summarization and etc. Coherence and cohesion is a branch of text linguistics and text linguistics is an application of NLP.

Coherence is used for document summarization systems. [2, 30, 31] In summarization systems, after the sentences have been generated or extracted, the system uses to coherence for determining and selecting the most appropriate sentences. The most coherence sentences are appropriate to selection in summarization systems.

Coherence is used to automatic evaluation of student essays by Miltsakaki and Kukich (2004). [34] They inspect “whether local discourse coherence, as defined by a measure of Centering Theory's Rough-Shift transitions, might be a significant contributor to the evaluation of essays. Rough-Shifts within students' paragraphs often occur when topics are short-lived and unconnected, and are therefore indicative of poor topic development”. Essentially, Miltsakaki and Kukich (2004) are correlating the type of topic transitions with the coherence of a text under the assumption that a text exhibiting a large number of drastic topic transitions corresponds to an incoherent text. [34] The essays “are scored on a scale of 1-6 points, where a score of 1 indicates an extremely poor essay and a score of 6 indicates an excellent essay”. In their experiment, they manually tagged coreferring expressions rather than use coreference software. Automatically determining student essay quality saves human resources, time and money.

Coherence models and machine translation evaluation metrics are combined for summarization evaluation Lin et al (2012). [29] They says “we adapt a machine translation metric to measure content coverage, apply an enhanced discourse coherence model to evaluate summary readability, and combine both in a trained regression model to evaluate overall responsiveness”.

2.4 COHERENCE MODELS

Coherence is a way of determining the quality of the discourse in terms of its effort to at clarity and ability to carry purposeful. There are two different coherence scopes. These are *global* coherence and *local* coherence. Coherence between small parts of texts typically sentences or no longer than a paragraph is mostly known as local coherence, whereas coherence between larger parts of texts is mostly identified as global coherence.

Grosz et al (1995) defines global coherence as coherence occurring between discourse segments. A discourse segment is a smaller unit of text which comprises the discourse. [19] The effectiveness of the individual discourse segments on the overall discourse purpose would be a measure of the text's global coherence. Taboada and Zabala (2008) say that “A discourse segment is recognizable because it always has an underlying intention associated with it. Discourse segments can also be embedded. They exhibit local coherence (among the utterances in the segment), and global coherence (with other segments in the discourse).” [40]

Grosz et al (1995) defines local coherence as coherence occurring between utterances within a discourse segment. [19] A discourse segment is comprised of utterances. An utterance is an elementary discourse unit, which is defined to be the simplest unit in a discourse that cannot be decomposed into a simpler unit. An utterance can take the form of a clause, simple sentence, complex sentence, paragraphs, etc.

Grosz and Sidner (1986) describe the relationship between utterance and discourse segment as such: “The utterances in a segment, like the words in a phrase, serve particular roles with respect to that segment. In addition, the discourse segments, like the phrases, fulfill certain functions with respect to the overall discourse. Although two consecutive utterances may be in the same discourse segment, it is also common for two consecutive utterances to be in different segments. It is also possible for two utterances that are nonconsecutive to be in the same segment.” [18]

In this thesis local coherence can be handled between adjacent sentences. The overall coherence of the text is thus determined by the sum of all the local coherence values. This is not to be bewildered with the global coherence of the text, which is coherence between discourse segments and their discourse intentions.

2.5 COHERENCE MEASUREMENT APPROACHES AND RELATED WORKS

A number of different metrics for evaluating the coherence of texts have been devised. In this section we provide an overview of some coherence measurement models, theories and methods in literature.

2.5.1 Vector Space Models

The vector space models for information retrieval are one class of retrieval techniques that have been studied in latest years. SMART (system for the mechanical analysis and retrieval of text) Salton and McGill (1983) developed by Gerald Salton and his colleague at Cornell University was one of the first examples of a vector space information retrieval model. [37] The major application in vector based methods to semantics for usage in assessment textual coherence is the normal model of content vector analysis. In vector based model, each document is related with a vector of the terms in the document, and each term is represented as a vector list the documents in which the term occurs.

Sample figures from Berry and Browne (2005); Table 2 shows how a simple vector space model can be represented as a term by document matrix. Each column defines a document and each row matches to a single term in the corpus. The value stored in each matrix cell explains the frequency that a term occurs in a document. For instance, *Term 3* appears once in *Document 2* and *Document 3*, but not in other documents. [5]

	Document 1	Document 2	Document 3	Document 4
Term 1	1	0	1	0
Term 2	0	0	1	1
Term 3	0	1	1	0

Table 2 - Small Term-By-Document Matrix

Figure 1 Berry and Browne (2005) shows how each column of the 3x4 matrix in Table 2 can be represented as a vector in 3-dimensional vector space. [5] Measures such as cosine similarity and Euclidean distance between document or term vectors provide the similarity values for coherence measurement or other rankings.

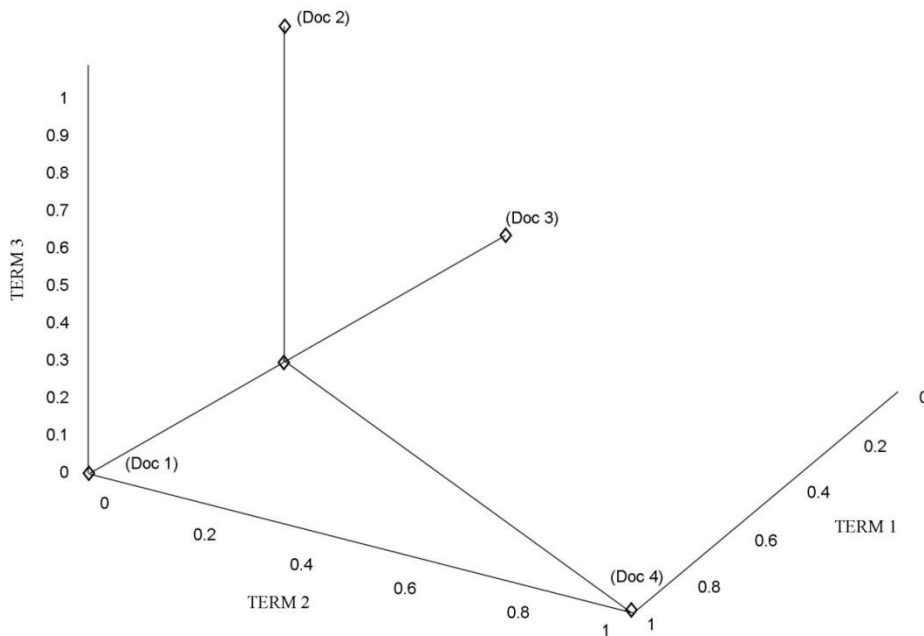


Figure 1 – Representation of Documents in a 3-Dimensional Vector Space

Standard information retrieval vector space model approach is not an effective method of calculating similarity scores. Its problem is that the components of a document vector match directly to terms and a sentence of 20 terms or so is not a large enough sample for the frequencies of terms within the sentence to regularly be found near their expectations. It is barely surprising to find a sentence about *computer* which nevertheless does not contain the term '*computer*', but it is more noteworthy to find a *computer* article of multiple paragraphs which does not contain the term '*computer*'. For example, we have two sentences about computer hardware,

one contain the terms '*Motherboard*', '*CPU*', and '*Memory*', other contain the terms '*DDR*', '*Kernel*', and '*Processor*'. These two sentences have no terms in common, in order that their vectors are orthogonal in semantic space. This means that these sentences have no semantic association. This is exactly intolerable for sentence similarity which we want to be suitable in evaluate textual coherence, and to accomplish this problem, there are several models which use dimension reduction.

LSA is the best known and most commonly used of the vector space methods semantic similarity using dimension reduction Landauer et al (2007). [26] LSA uses SVD method to rank reducing in the term by document matrix. Due to the fact that information is lost in this dimension reduction process, some related vectors are fused or moved from the space. It provides us automatically relations extraction such as the previous problematic example which a sentence contains computer hardware terms, and when different terms from this subject occur in different documents. LSA method deals with the synonymy problem and it offers a partial solution to the polysemy problem.

LSA has proved to be a great improvement over the standard vector space model for some domains. The term-term similarity scores it produces are more healthy Landauer et al (1998), as evidenced by the much cited results that LSA has been used to achieve a 64% score on a test of 80 synonym items from the Test of English as a Foreign Language. [25]

Deerwester et al (1990) used LSA method for IR and compare it with the standard vector space model methods like SMART and their results are modestly encouraging to them. The results showed that LSA to be superior to simple term to simple term matching in one standard case and equal in another. [10]

Foltz (1996) illustrate how LSA may be used in text based research. He researched three concepts. His two experimentations define methods for evaluating a subject's essay for determining from what text a subject learned the information and for grading the quality of information cited in the essay. His third experiment describes using LSA to measure the coherence and understandability of texts. After research,

LSA appears to be a talented application for these researchers. The method is automatic and fast, allowing fast measurements of the semantic similarity between parts of textual information. [13]

Foltz et al (1998) used LSA for measuring the coherence of texts. They describe their method:

“By comparing the vectors for two adjoining segments of text in a high-dimensional semantic space, the method provides a characterization of the degree of semantic relatedness between the segments. We illustrate the approach for predicting coherence through re-analyzing sets of texts from two studies that manipulated the coherence of texts and assessed readers' comprehension. The results indicate that the method is able to predict the effect of text coherence on comprehension and is more effective than simple term-term overlap measures.” [14]

At the next chapters will be mentioned about LSA techniques and works which use it. In this thesis, LSA uses for measurement of Turkish textual coherence because of LSA provides a automatic method for comparing parts of textual information to each other in order to evaluate their semantic relatedness and its results seems coherent when examine previous works.

2.5.2 Coh-Metrix Tool

The Coh-Metrix is a tool that evaluates the cohesion and coherence of texts. The Coh-Metrix Project is developed at the University of Memphis by the Institute for Intelligent Systems, and funded by a grant from the Institute of Education Sciences. The Coh-Metrix tool generally has three aims. These are: to measure textual coherence, to study the effects of textual cohesion on readers, and to fine-tune the cohesion metrics developed for the project.

This tool uses lexical, syntactic and semantic features present in a text for readability evaluation. It measures coherence by performing an empirical analysis of the linguistic aspects of the text. The approach that Graesser et al (2004) take is

commendable in that they leverage as much linguistic information about the text that they possibly can.

“Coh-Metrix is a computer program that analyzes various text features relevant to text comprehension by incorporating techniques informed by theories of text processing, cognitive psychology, and computational linguistics. Three key classes of cohesion indices (i.e., coreference, conceptual relations, connectivity) measured by Coh-Metrix are evaluated with texts used in published studies of cohesion effects on reading comprehension. The results confirmed that Coh-Metrix successfully detects levels of cohesion in texts.” [17]

The difficulty with the Coh-Metrix approach of making as many measurements of the text as possible is that there is no practical way to decrease all such measures to a single score. What has been brought together is a set of empirical data, with the objective being to determine whether the data represents a text that is coherent or not. As an example, it is as if one has made many measurements about a person's features, such as their height, foot size, hair color, eye color, etc., but the aim is to decide whether the sum of these characterizes essentially give an account of an real person being which coherent data or a non-existent person which incoherent data.

Without thinking about the defects connected to a purely empirical analysis of discourse, the contributions of Coh-Metrix are nevertheless important. The attributes of a text that Graesser et al (2004) measure are all valid in terms of evaluation coherence, with closely every single measure motivated as the result of a psychological study of coherence in text. [17]

The indices that Coh-Metrix version 3.0 uses are categorized into eleven groups:

- 1- Descriptive indices
- 2- Text easability principal component scores
- 3- Referential cohesion
- 4- Latent Semantic Analyses
- 5- Lexical diversity
- 6- Connectives
- 7- Situation Model
- 8- Syntactic Complexity
- 9- Syntactic Pattern Density
- 10- Word Information
- 11- Readability

2.5.3 Centering Theory Approach

Centering Theory is a theory that evaluates local coherence by following entity movement between sentences. [19, 41] It is one of the most powerful method for local coherence measurement. Grosz and Sidner (1986) have ideas which focus spaces and stacks continue in Centering Theory. [18] Miltsakaki (2003) says:

“The notion of focus space is, also, elusive. What is a focus space and how is it identified? Is the focus space equivalent to an abstract segment associated with a discourse purpose or is it an attentional update unit? A first attempt to model aspects of attentional structure yielded a reformulation of Centering as a model of local discourse coherence... Centering was developed as a model of the center of attention between speakers in natural language discourse. The model aimed at modeling the interaction between ‘attentional state’, inferential complexity and the form of referring expressions.” [33]

Discourse is composed of an order of textual segments and each segment is composed of an order of utterances. Utterances are intended by $U_i - U_N$. Each utterance U_i suggests a set of discourse entities, the *forward-looking centers*,

intended by $C_f(U_i)$. The members of the C_f set are ranked Brennan et al (1987) according to discourse salience. The highest ranked member of the C_f set is the preferred center, C_p . A *backward-looking center*, C_b , is recognized for utterance U_i . The highest ranked entity in the previous utterance, $C_f(U_{i-1})$, that is understood in the current utterance, U_i , is its intended *backward-looking center*, C_b . The *backward-looking center* is a special member of the C_f set because it represents the discourse entity that U_i is about, what in the literature is often called the “topic”. [35, 21] The C_p for a given utterance may be identical with its C_b , in addition to this is not necessarily. It is exactly this difference between looking back in the discourse with the C_b and planning preferences for clarification in following discourse with the C_p that provides the key element in calculating local coherence. [34]

Centering Theory groups the degree of coherence based on orders of utterance transitions. The degree of coherence groups the utterance transitions into four types:

- 1- Continue
- 2- Retain
- 3- Smooth - shift
- 4- Rough – shift

They are shown in transition ordering rule. It is *Continue* is preferred to *Retain*, which is preferred to *Smooth-Shift*, which is preferred to *Rough - Shift*. Table of transitions:

	$C_b(U_i) = C_b(U_{i-1})$	$C_b(U_i) \neq C_b(U_{i-1})$
$C_b(U_i) = C_p$	Continue	Smooth-Shift
$C_b(U_i) \neq C_p$	Retain	Rough-Shift

Table 3 - Centering Theory Transitions

Barzilay and Lapata (2005) and Barzilay and Lapata (2008) planned an entity founded model to present and estimate local textual coherence. [3, 4] The model is interested by Centering Theory. They operationalized Centering Theory by creating an entity grid model to capture discourse entity transitions at the sentence to

sentence level. Then, they demonstrated their model's skill to distinguish coherent texts from incoherent ones. Miltsakaki and Kukich (2004) used Centering Theory's *Rough – Shift* transitions for evaluation of text coherence for electronic essay scoring systems. [34]

2.5.4 Other Approaches

Barzilay and Lee (2004) suggested a domain dependent Hidden Markov Model (HMM) to capture topic shift in a discourse, where topics are presented by hidden sentences are observations. [2] The global coherence of a discourse can be presented by the overall likelihood of topic shift from the first sentence to the last sentence. Following Barzilay and Lee (2004); Barzilay and Lapata (2005); Soricut and Marcu (2006); Elsner et al (2007)) merger the entity and HMM based models and showed that these two models are perfecting to each other in coherence evaluation. [2, 3, 38, 11]

Lin et al (2011) present a new model to represent and evaluate the coherence of a discourse. Their project supposes that coherent text absolutely prefers certain types of discourse relation transitions. Their purpose is automatically evaluating text coherence using discourse relations. [28]

2.5.5 Summary

This chapter has outlined the terminology, ideas and motivations used in the rest of this thesis. A discussion of coherence and cohesion models, how they are related. The different approaches have been inspected in order to provide a basis for the idea that how can textual coherence measure. A computationally implementable approach of measuring Turkish textual coherence, LSA, leads us to the process presented in this thesis.

CHAPTER III

LATENT SEMANTIC ANALYSIS

Latent Semantic Analysis was developed late 1980s at Bell Core/Bell Laboratories by Landauer and his team of Cognitive Science Research. "LSA is a theory and method for extracting and representing the contextual- usage meaning of words by statistical computations applied to a large corpus of text. The corpus embodies a set of mutual constraints that largely determine the semantic similarity of words and sets of words. These constraints can be solved using linear algebra methods, in particular, Singular Value Decomposition." [26]

LSA is a mathematical and statistical approach. It is based on vector space model, an algebraic representation of text documents used in IR. LSA is closely related to neural networks models, but is based on SVD which is mathematical matrix decomposition technique.

LSA analyses relationships between a set of documents and the terms and extracts information such as which terms are common or uncommon terms and which terms are used together and which documents are semantically related. LSA uses a semantic space for extracts information. Semantic space is derived from a term-by-document co-occurrence matrix and terms and documents can be represented as vectors. Essential part of this derivation is dimensionality reduction using SVD.

Acceptances to LSA focused on following labels: latent dimensions, synonymy, polysemy and term dependence. [10]

LSA has challenges that focused on scalability and performance. LSA needs high memory and computational performance in comparison to other information retrieval techniques. [23] In addition to these, determining the optimal number of dimensions for SVD calculation is another challenge to LSA.

LSA method usually contains four main steps for measure textual coherence:

1. LSA Semantic Space Construction
2. SVD Computation and Dimension Reduction
3. Query Matching
4. Similarity Measurement

3.1 LSA SEMANTIC SPACE CONSTRUCTION

3.1.1 Segmentation

Segmentation is important to comprehend input documents for improving performance. Segmentation is an important preprocessor because of each language has own building and own segmentation procedures. The little part is segment that is extracted from document. It can be a sentence, a paragraph or a phrase. In order to measure textual coherence sentence segmentation uses commonly. [14]

3.1.2 Stop-word Filtering

Stop-word filtering is important to extract meaningless terms which less useful and less informative. If mentioned about the English Language “a”, “the”, “is” and etc. are stop-words. If these terms uses, they cause extra noise in term-by-document matrix.

There are two approaches for remove stop words. One of them is using predefined human-made words lists. This approach depends on language. Another approach is used a frequency threshold. Terms which are observed more and less frequently can be considered as stop-word. Decision of frequency limits are another issue to be considered. In this thesis, human-made words lists are used for stop-word removing method. (See Appendix A).

3.1.3 Stemming

Stemming technique is important for improving performance and accuracy. Stemming is the process for reducing inflected forms and sometimes derivationally related forms of a term to a common base form. The purpose of the stemming is to improve the ability to discover similarity of the use of term differs like decreases the number of synonyms which multiple terms using the same stem are mapped on the same stem, but sometimes because of stemming errors, it produce new homonyms. At the same time, stemming provides less noisy and denser term-by-document matrix. So that calculation performance is higher.

Stemming algorithms are language dependent. If the morphology, orthography, and character encoding of the target language becomes more complex, stemmers design become harder.

In this thesis, Zemberek Morphological Analyzer is used for stemming process. [45] Zemberek is an open source, platform independent, general purpose NLP library and toolset designed for Turkic languages.

3.1.4 Term Reduction

Term reduction method provides noise extraction from semantic space. According to Zip's law a large number of terms only appear in one document can be extracted from the terms because they have little information for finding associations between documents.

3.1.5 Term Weighting and Term-Document Matrix Creation

Representing documents founded on the occurrence of terms can be filtered by introducing a weighting scheme to better identify the characteristic terms. Term weighting hence tend to filter out common terms. It has a similar effect as stop-word removal. Terms commonly used across all documents in the corpus is down weighted compared to medium frequency terms, which carry the most important information as can be expected according to Zip's law. From the other point of view, TF-IDF weighting process can establish extreme weights to words with very low frequencies. In addition to this, it is not hold tight synonyms, therefore weights of commonly used synonyms are overrated, and as the weights of the synonym terms are higher than the weight of the underlying common concept.

There are several term weighting calculation approaches in literature. These approaches are as follows: frequency of word, binary representation, log entropy, root type and TF - IDF. In this thesis commonly used weighting scheme is TF - IDF (Term Frequency – Inverse Document Frequency) weighting scheme Salton and McGill (1983) are used. [37] For creation of term-by-document matrix, its cells filled out with TF - IDF value of the terms. TF – IDF is the product of two statistics, term frequency and inverse document frequency.

For calculating term frequency Formula (1) uses. The $f(t,s)$ is the number of times that term t occurs in sentence s , and $f(w,s)$ is the total number of occurrences of all terms in the sentence s .

$$tf(t,s) = \frac{f(t,s)}{\max\{f(w,s) : w \in s\}} \quad (1)$$

The inverse document frequency value is calculated using Formula (2). It is come by dividing the total number of sentences in the corpus by the number of sentences containing the term t , and then taking the logarithm of that quotient.

$$idf(t, S) = \log \frac{|S|}{|\{s \in S : t \in s\}|} \quad (2)$$

In order to calculate TF - IDF value Formula (3) uses. Term frequency (TF) is multiplied by the inverse document frequency (IDF). Terms which are commonly used across all documents in the document set that its IDF values will closer to zero. The higher TF - IDF value indicates that the term is much more characteristic for that document than others.

$$tfidf(t, s, S) = tf(t, s) \times idf(t, S) \quad (3)$$

3.2 SVD COMPUTATION AND DIMENSION REDUCTION

Singular Value Decomposition is a factorization of a real or complex matrix. LSA applies SVD to the term-by-document matrix. SVD design a mapping such that the low-dimensional space reflects semantic associations. SVD is used as a rank reducing method to truncate the original vector space to reveal the underlying or latent semantic structure in the pattern of word usage to define documents in a collection. This truncation allows dealing with typical language issues like synonymy as different words expressing the same idea are supposed to be close to each other in the reduced k-dimensional vector space. The magnitude of singular vectors gives information about the importance of the concept.

In full SVD, original term-by-document matrix A is decomposed into three new matrices is defined as equation (4).

$$A = USV^T \quad (4)$$

A : Original term-by-document matrix ($m \times n$)

U : The left singular vectors of A ($m \times n$)

S : The singular (scaling) values of A ($n \times n$)

V^T : The right singular values of A ($n \times n$)

In reduced SVD or truncated SVD, original term-by-document matrix A is decomposed into three new matrices is defined as equation (5). These matrices are given in Figure 2. There are two cases in figure; first half of the figure presents matrix A for terms count bigger than documents count and second half of the figure presents matrix A for documents count bigger than terms count.

$$A_k = U_k S_k V_k^T \quad (5)$$

A_k : Rank k approximation of the original term-by-document matrix ($m \times n$)

U_k : Rank k approximation of the left singular vectors of A (terms profile) ($m \times k$)

S_k : Rank k approximation of the singular (scaling) values of A ($k \times k$)

V_k^T : Rank k approximation of the right singular values of A (documents profile) ($k \times n$)

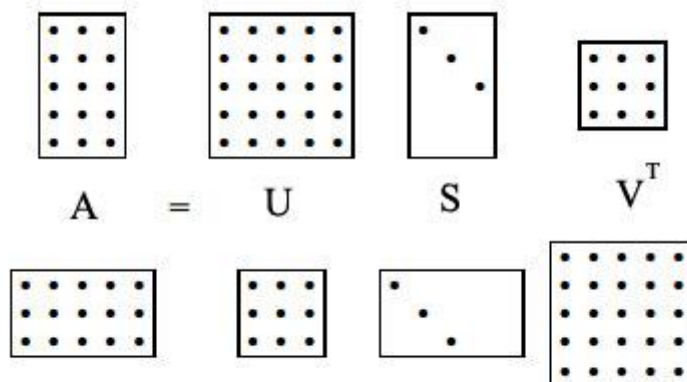


Figure 2 – Illustration of SVD

In this thesis we use reduced SVD and the reduction in dimension, k value, allows eliminating noise and capturing the underlying latent structure. By using SVD the major associative patterns are extracted from the document space and the small patterns are ignored. The choice of the k value to be retained is not straightforward and it is an open argument subject. Letsche and Berry (1997) say that, for very large databases, the number of dimensions used usually ranges between 100 and 300. [27] And also Landauer and Dumains (1997) say that, for some applications it might be better to use a subset of the first 100 or 300 dimensions. [24]

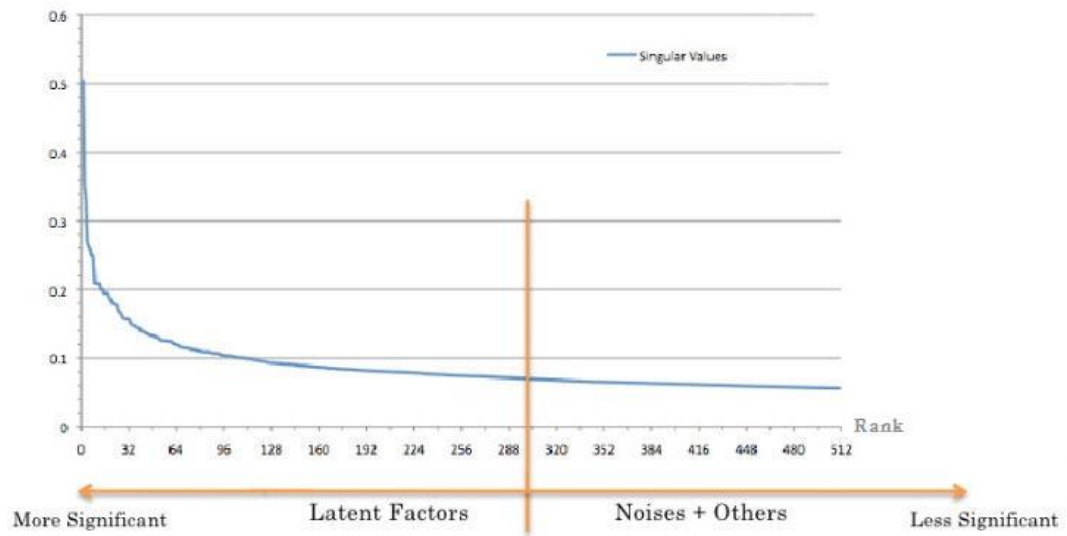


Figure 3 – Singular Value Decomposition Interpretation for Rank

3.3 QUERY MATCHING

For comparing the query vector to the vectors of the reduced rank term-by-document matrix or another query vector that requires the query vector reduction in same semantic space rank value. A query vector is mapped into reduced LSA semantic space by the transformation. (Formula 6)

$$q = q^T U_k S_k^{-1} \quad (6)$$

The cosines of the angles between the transformed query vector and approximated document vector give similarity. (Formula 7)

$$\text{sim}(q, d) = \text{sim}(q^T U_k S_k^{-1}, d^T U_k S_k^{-1}) \quad (7)$$

The cosines of the angles between the transformed query vector and another transformed query vector give similarity. (Formula 8)

$$\text{sim}(q_1, q_2) = \text{sim}(q_1^T U_k S_k^{-1}, q_2^T U_k S_k^{-1}) \quad (8)$$

3.4 SIMILARITY MEASUREMENT

In this thesis, cosine similarity measure Berry and Browne (1999) are used. [5] It is typically used in IR applications Yates and Neto (1999). [44] It is an expression for the angle between vectors, formulated as an inner product of two vectors, divided by the product of their Euclidean norms.

While other similarity measures are possible, the cosine measure is amongst the most commonly used when using LSA and looks greater as a similarity measure in LSA applications.

$$\text{sim}(q, d) = \frac{\sum_{i=1}^n q_i \times d_i}{\sqrt{\sum_{i=1}^n q_i^2} \times \sqrt{\sum_{i=1}^n d_i^2}} \quad (9)$$

In the original vector space, all vector elements are positive. The results are values between 1 and 0 after application of a weighting scheme. If vectors are similar to each other, similarity values will be closer to 1 and otherwise closer to 0.

In the reduce concept space after SVD, vector elements may become negative because of SVD calculation, so the measures of similarity can range from -1 to 1. If

the value of the measure is near 1 then it means that the vectors are similar, otherwise if the value of the measure is near -1 then it means that the vectors are dissimilar.

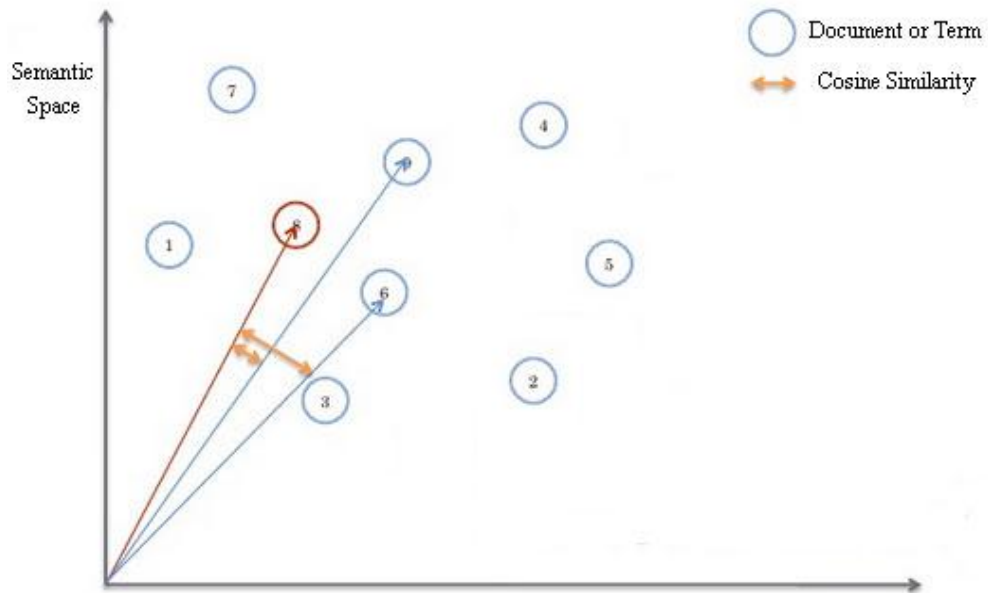


Figure 4 – Cosine Similarity in Semantic Space

3.5 LSA EXAMPLE

In order to see how LSA can represent the meaning of terms and documents an example is presented in Figure 5, Figure 6, Figure 7, Figure 8 and Figure 9. The example is taken in Martin and Berry (2007)'s study: [32]

<i>Label</i>	<i>Titles</i>
M1	<i>Rock and Roll Music in the 1960's</i>
M2	<i>Different Drum Rolls, a Demonstration of Techniques</i>
M3	<i>Drum and Bass Composition</i>
M4	<i>A Perspective of Rock Music in the 90's</i>
M5	<i>Music and Composition of Popular Bands</i>
B1	<i>How to Make Bread and Rolls, a Demonstration</i>
B2	<i>Ingredients for Crescent Rolls</i>
B3	<i>A Recipe for Sourdough Bread</i>
B4	<i>A Quick Recipe for Pizza Dough using Organic Ingredients</i>

Figure 5 - Documents for Using In LSA

<i>Types</i>	<i>Documents</i>								
	M1	M2	M3	M4	M5	B1	B2	B3	B4
Bread	0	0	0	0	0	1	0	1	0
Composition	0	0	1	0	1	0	0	0	0
Demonstration	0	1	0	0	0	1	0	0	0
Dough	0	0	0	0	0	0	0	1	1
Drum	0	1	1	0	0	0	0	0	0
Ingredients	0	0	0	0	0	0	1	0	1
Music	1	0	0	1	1	0	0	0	0
Recipe	0	0	0	0	0	0	0	1	1
Rock	1	0	0	1	0	0	0	0	0
Roll	1	1	0	0	0	1	1	0	0

Figure 6 - Term-By-Document Matrix After Stop-Word Removing Process

<i>Types</i>	<i>Documents</i>								
	M1	M2	M3	M4	M5	B1	B2	B3	B4
Bread	0	0	0	0	0	.474	0	.474	0
Composition	0	0	.474	0	.474	0	0	0	0
Demonstration	0	.474	0	0	0	.474	0	0	0
Dough	0	0	0	0	0	0	0	.474	.474
Drum	0	.474	.474	0	0	0	0	0	0
Ingredients	0	0	0	0	0	0	.474	0	.474
Music	.347	0	0	.347	.347	0	0	0	0
Recipe	0	0	0	0	0	0	0	.474	.474
Rock	.474	0	0	.474	0	0	0	0	0
Roll	.256	.256	0	0	0	.256	.256	0	0

Figure 7 – Term-By-Document Matrix After Weighting Process

Matrix U-Type Vectors

Bread	.42	-.09	-.20	.33	-.48	-.33	.46	-.21	-.28
Composition	.04	-.34	.09	-.67	-.28	-.43	.02	-.06	.40
Demonstration	.21	-.44	-.42	.29	.09	-.02	-.60	-.29	.21
Dough	.55	.22	.10	-.11	-.12	.23	-.15	.15	.11
Drum	.10	-.46	-.29	-.41	.11	.55	.26	-.02	-.37
Ingredients	.35	.12	.13	-.17	.72	-.35	.10	-.37	-.17
Music	.04	-.35	.54	.03	-.12	-.16	-.41	.18	-.58
Recipe	.55	.22	.10	-.11	-.12	.23	-.15	.15	.11
Rock	.05	-.33	.60	.29	.02	.33	.28	-.35	.37
Roll	.17	-.35	-.05	.24	.33	-.19	.25	.73	.22

Matrix Σ -Singular Values

	1.10	0	0	0	0	0	0	0	0
	0	.96	0	0	0	0	0	0	0
	0	0	.86	0	0	0	0	0	0
	0	0	0	.76	0	0	0	0	0
	0	0	0	0	.66	0	0	0	0
	0	0	0	0	0	.47	0	0	0
	0	0	0	0	0	0	.27	0	0
	0	0	0	0	0	0	0	.17	0
	0	0	0	0	0	0	0	0	.07
	0	0	0	0	0	0	0	0	0

Matrix V-Document Vectors

M1	.07	-.38	.53	.27	.08	.12	.20	.50	.42
M2	.17	-.54	-.41	.00	.28	.43	-.34	.22	-.28
M3	.06	-.40	-.11	-.67	-.12	.12	.49	-.23	.23
M4	.03	-.29	.55	.19	-.05	.22	-.04	-.62	-.37
M5	.03	-.29	.27	-.40	-.27	-.55	-.48	.21	-.17
B1	.31	-.36	-.36	.46	-.15	-.45	.00	-.32	.31
B2	.19	-.04	.06	-.02	.65	-.45	.41	.07	-.40
B3	.66	.17	.00	.06	-.51	.12	.27	.25	-.35
B4	.63	.27	.18	-.24	.35	.10	-.35	-.20	.37

Figure 8 – Singular Values Decomposition Matrices

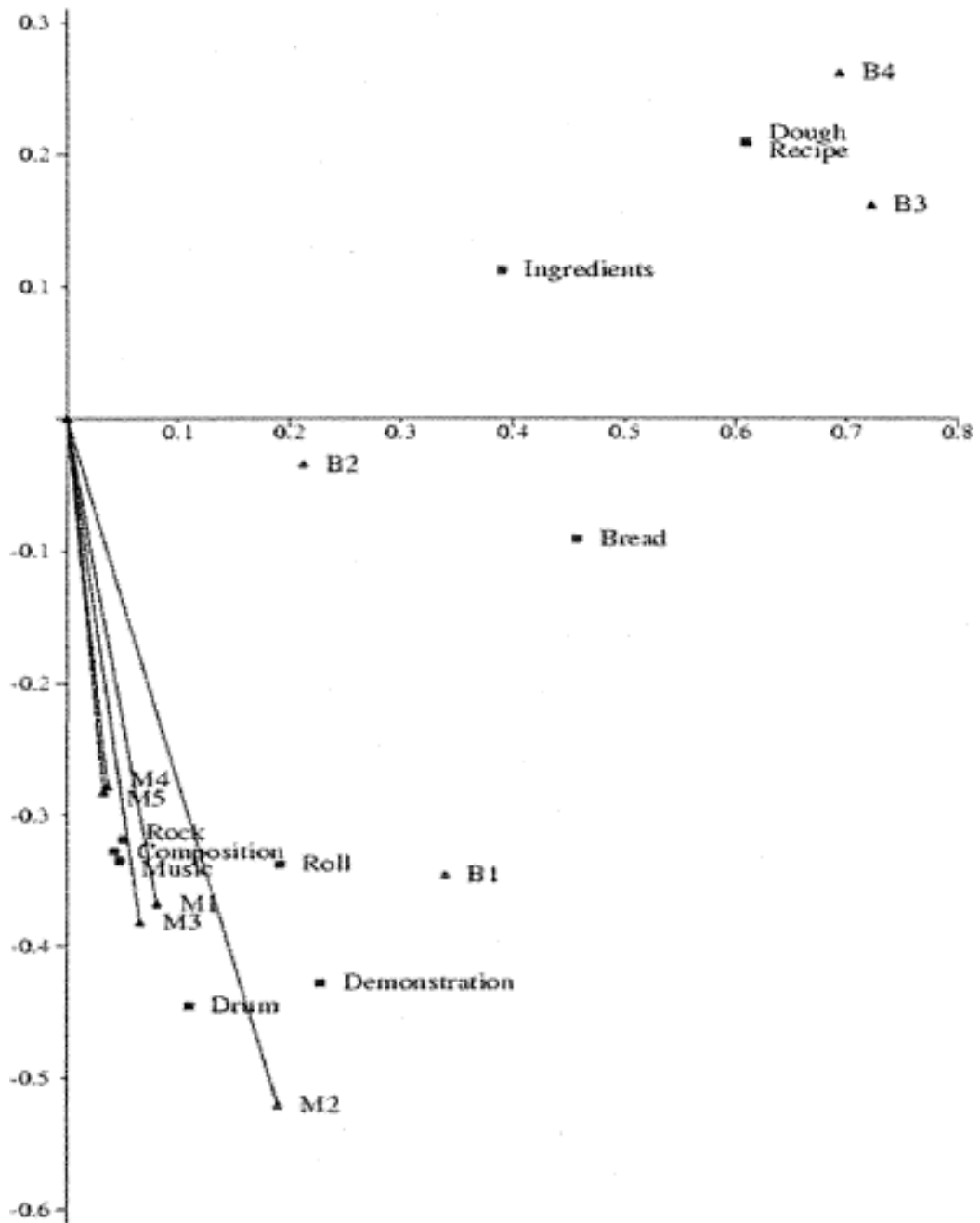


Figure 9 – Meaning of Terms and Documents After Similarity Measure In LSA Semantic Space

From the Figure 9, we can see that document M3 is more related to M1 than B3 and the term “Rock” is related to the term “Music” than “Bread”. This type of analysis can represent term and documents relationships by using Latent Semantic Analysis.

CHAPTER IV

DATA AND EXPERIMENTAL METHODOLOGY

LSA for text measurement systems perform the four steps in Chapter 3. These are LSA semantic space construction, SVD computation and dimension reduction, query matching, similarity measurement. In this thesis, different methods are used to measure similarity. Methods which cosine similarity and sliding window procedure that compares one group of sentences to the next group of sentences repeatedly across the text are used.

For example, a sliding window of size 4 sentences, computing the cosine between the vectors based on the first four sentences of a document and the vector for the next four sentences and then sliding both segments one sentence further to compute the next cosine. It is certain that the classical sentence to adjacent sentence mean cosine is a special case of this procedure with a window size of 1.

Sliding window procedure suggested by Foltz (2007) and he describes its advantages;

“The advantage of the sliding window is that it tends to smooth the coherence predictions, although a large drop in coherence still would indicate that there is a marked change in the general semantic content of the text at a particular point. A second advantage of the sliding window technique is that it captures, to some degree, the fact that some propositions are held over in working memory for several sentences.” [15]

4.1 CONSTRUCTION OF LSA SEMANTIC SPACE

In this study, Turkish newspaper columnist's 400 articles are used to measure textual coherence based on content similarity and Turkish Wikipedia documents are used to LSA to derive content similarity. Construction of LSA semantic space from Turkish Wikipedia corpus has multiple preprocessing steps such as segmentation, stop-word filtering, stemming, term reduction and weighting.

For creation LSA semantic space, each Wikipedia subject page are used one segment for indexing documents and terms. We use own software for the index documents into token or terms and collecting a list how many times a given term in a given document. We used human-made words lists for stop-word filtering (See Appendix A). Punctuations and numbers are removed from documents. Zemberek is applied for stemming.

We use own software for construction term-by-document matrix by converting our indexes. This results in a matrix with 279,396 stemmed terms (rows) and 50,000 documents (columns). We eliminate terms the only appearing one, two and three document and top 100 terms. After removal of terms, we have a matrix with 53,109 stemmed terms and 50,000 documents. This term-by-document matrix contains the raw term frequencies. To improve retrieval and matching performance, matrix cell values which raw term frequencies are weighted by TF-IDF weighting method. Hereby, terms gain the relative importance in the corpus.

After weighting process, dimension reduction process is applied. Dimension reduction is fundamental part of the LSA method. Vector space is reduced by applying SVD. There are many libraries for calculating SVD in various programming languages. We use SVDLIBC which is a C library written by Doug Rohde. It was based on the SVDPACKC library, which was written by Michael Berry, Theresa Do, Gavin O'Brien, Vijay Krishna and Sowmini Varadhan at the University of Tennessee. [39]

Reduction is done to get a rank k approximation matrix from original matrix. Approximation matrix is intended to remove noise because of synonymy and

polysemy give in documents. In Turkish there are many synonymy and polysemy words.

Dimension reduction level and best selection of the rank k is an open question. As mentioned Chapter 3, observations show the better rank value between 100 and 300 for large datasets. In this thesis, we select 100, 150, 200, 250, and 300 rank approximation values. We compare these five levels of dimension reduction LSA space and sliding window sizes for cosine similarity measures.

Figure 10 presents processes for creating LSA space.

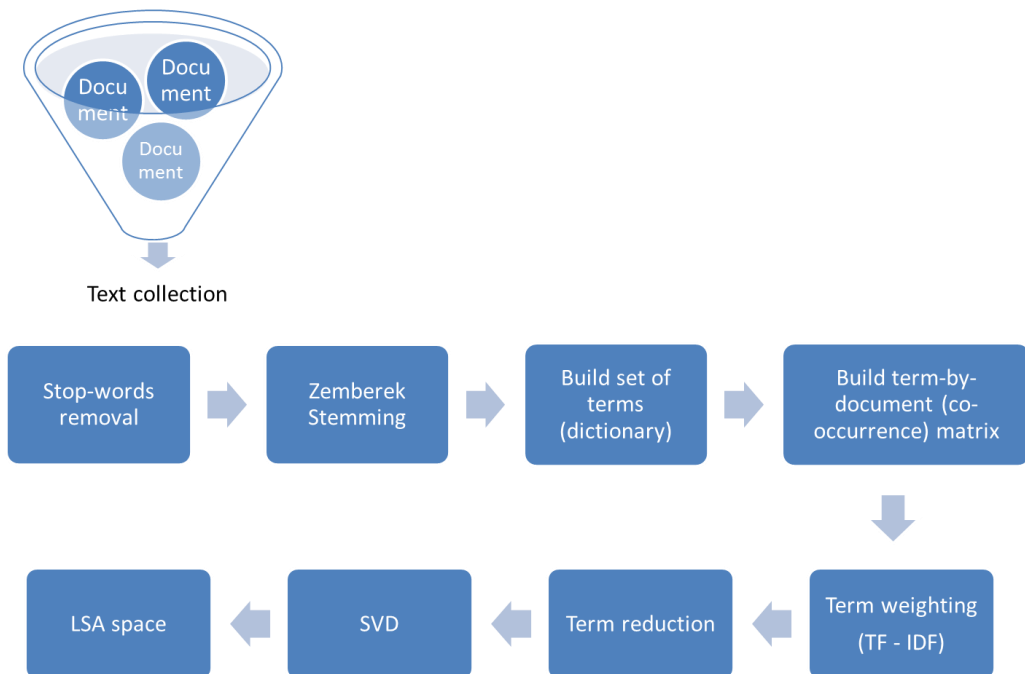


Figure 10 – Semantic Space Creation Processes

4.2 COHERENCE OF ARTICLES

In this thesis, 400 newspaper articles' coherence scores are calculated by using cosine similarity. Aim of the study is calculation coherence values and gender and the name of the authors' identifications for articles using the coherence values. Table 4 and Table 5 present the gender and author distributions for articles.

Gender	Count
Male	200
Female	200

Table 4 - Gender Distribution for 400 Newspaper Articles

Author	Count	Author	Count
Abbas Güçlü	4	Ahmet Hakan	6
Ali Esad Göksel	1	Aziz Üstel	10
Ayşe Önal	20	Can Dünder	6
Cüneyt Arcayürek	10	Çetin Altan	6
Doğan Hızlan	17	Ebru Çapa	20
Ertuğrul Özkök	7	Gülse Bırsel	20
Hasan Cemal	5	Hasan Pulur	6
İdil Çeliker	10	Mehmet Barlas	7
Mehmet Y. Yılmaz	7	Melda Narmanlı Çimen	10
Meliha Okur	20	Meral Tamer	20
Necati Doğru	10	Necef Uğurlu	20
Nuri Soysal	10	Osman Müftüoğlu	10
Özlem Yüzak	10	Pakize Suda	20
Ruşen Çakır	10	Saadet Oruç	10
Sedat Ergin	5	Şükrü Kızılot	6
Taha Akyol	8	Vahap Munyar	6
Vehbi Tülek	10	Yalçın Doğan	6
Yıldırım Türker	10	Yılmaz Özdil	17
Zuhal Kızılot	20		

Table 5 - Category Distribution for 400 Newspaper Articles

For prepare newspaper articles to measure coherence; segmentation, stop-word filtering and stemming methods are applied to newspaper articles. Sentences groups are used for segmentation. We use sliding window procedure for measure coherence scores. For using this method, each of the articles is separated individual sentences

and then sentences combine according as window sizes. We have five window sizes; 1 sentence window, 2 sentences window, 3 sentences window, 4 sentences window and 5 sentences window. The example of sliding window procedure window size 1 and window size 5 are presented in Table 6 and Table 7 with using a sample article (See Appendix B).

Groups	Sentences	Sentences
1.sentence – 2.sentence	çocuk üniversiteleridünyada bilim toplumu yaratmak için uygulanan çok farklı projeler var	bunlardan birisi de çocuk üniversiteleri
2.sentence – 3.sentence	bunlardan birisi de çocuk üniversiteleri	amaç olabildiğince küçük yaşlarda çocukları bilimle tanıştırmak ve üniversiteye yönlendirmek
3.sentence – 4.sentence	amaç olabildiğince küçük yaşlarda çocukları bilimle tanıştırmak ve üniversiteye yönlendirmek	çocuk üniversitelerinin ilköğretim öğrencilerine yönelik olanı da var orta öğretime yönelik olanları da
4.sentence – 5.sentence	çocuk üniversitelerinin ilköğretim öğrencilerine yönelik olanı da var orta öğretime yönelik olanları da	eskiden başka ülkelerde olduğunu duyar neden bizde de yok diye iç geçirirdik
5.sentence – 6.sentence	eskiden başka ülkelerde olduğunu duyar neden bizde de yok diye iç geçirirdik	ama son yıllarda bizde de çok güzel örneklerini görmeye başladık
continues like above	continues like above	continues like above

Table 6 - Example of Window Size 1

Groups	Sentences	Sentences
1.2.3.4.5.sentences – 6.7.8.9.10 sentences	çocuk üniversiteleridünyada bilim toplumu yaratmak için uygulanan çok farklı projeler var. bunlardan birisi de çocuk üniversiteleri.amaç olabildiğince küçük yaşlarda çocukları bilimle tanıştırmak ve üniversiteye yönlendirmek. çocuk üniversitelerinin ilköğretim öğrencilerine yönelik olanı da var orta öğretime yönelik olanları da.eskiden başka ülkelerde olduğunu duyar neden bizde de yok diye iç geçirirdik.	ama son yıllarda bizde de çok güzel örneklerini görmeye başladık.çocuk üniversiteleri abd gibi dünya bilimine en fazla katkıyı sağlayan ülkelerin olmazsa olmazlarının başında geliyor.sadece ciddi finansal destek sağlamakla kalmıyor yaygınlaştırılması için her türlü çabayı gösteriyorlar.yani dayatmaya dayalı yönlendirme yerinebilime yönelik bilgilendirme sevdirm ve özendirmesöz konusu.başka türlü de zaten bilim toplumu olunmuyor.
2.3.4.5.6.sentences – 7.8.9.10.11.sentences	bunlardan birisi de çocuk üniversiteleri.amaç olabildiğince küçük yaşlarda çocukları bilimle tanıştırmak ve üniversiteye yönlendirmek. çocuk üniversitelerinin ilköğretim öğrencilerine yönelik olanı da var orta öğretime yönelik olanları da.eskiden başka ülkelerde olduğunu duyar neden bizde de yok diye iç geçirirdik. ama son yıllarda bizde de çok güzel örneklerini görmeye başladık.	çocuk üniversiteleri abd gibi dünya bilimine en fazla katkıyı sağlayan ülkelerin olmazsa olmazlarının başında geliyor.sadece ciddi finansal destek sağlamakla kalmıyor yaygınlaştırılması için her türlü çabayı gösteriyorlar.yani dayatmaya dayalı yönlendirme yerinebilime yönelik bilgilendirme sevdirm ve özendirmesöz konusu.başka türlü de zaten bilim toplumu olunmuyor.üniversitelerimizin bu konudaki çabaları takdire şayan.
3.4.5.6.7.sentences – 8.9.10.11.12.sentences	amaç olabildiğince küçük yaşlarda çocukları bilimle tanıştırmak ve üniversiteye yönlendirmek. çocuk üniversitelerinin ilköğretim öğrencilerine yönelik olanı da var orta öğretime yönelik olanları da.eskiden başka ülkelerde olduğunu duyar neden bizde de yok diye iç geçirirdik. ama son yıllarda bizde de çok güzel örneklerini görmeye başladık.çocuk üniversiteleri abd gibi dünya bilimine en fazla katkıyı sağlayan ülkelerin olmazsa olmazlarının başında geliyor.	sadece ciddi finansal destek sağlamakla kalmıyor yaygınlaştırılması için her türlü çabayı gösteriyorlar.yani dayatmaya dayalı yönlendirme yerinebilime yönelik bilgilendirme sevdirm ve özendirmesöz konusu.başka türlü de zaten bilim toplumu olunmuyor.üniversitelerimizin bu konudaki çabaları takdire şayan.şu anda istanbul ankara ve inönü üniversiteleri bu konuda öncü durumunda.
continues like above	continues like above	continues like above

Table 7 - Example of Window Size 5

After these methods, the vector for each group of sentences is computed. We calculate similarity scores between these groups of sentences' vectors. Each group of sentences' vector compares the next group of sentences' vector. The cosine between these two vectors showed their similarity or semantic relatedness or coherence. After the measure similarity the both segments are shifted one sentence further. An overall local coherence of article A is calculated for each article by averaging the all similarity measures. The overall coherence result is closer to 1; we can say article's coherence is high. On the other hand, the overall coherence result is closer to -1; we can say article's coherence is low.

$$coherence(A) = \frac{\sum_{i=1}^{n-1} sim(s_i, s_{i+1})}{n-1} \quad (10)$$

Figure 11 presents processes for calculation coherence for an article.

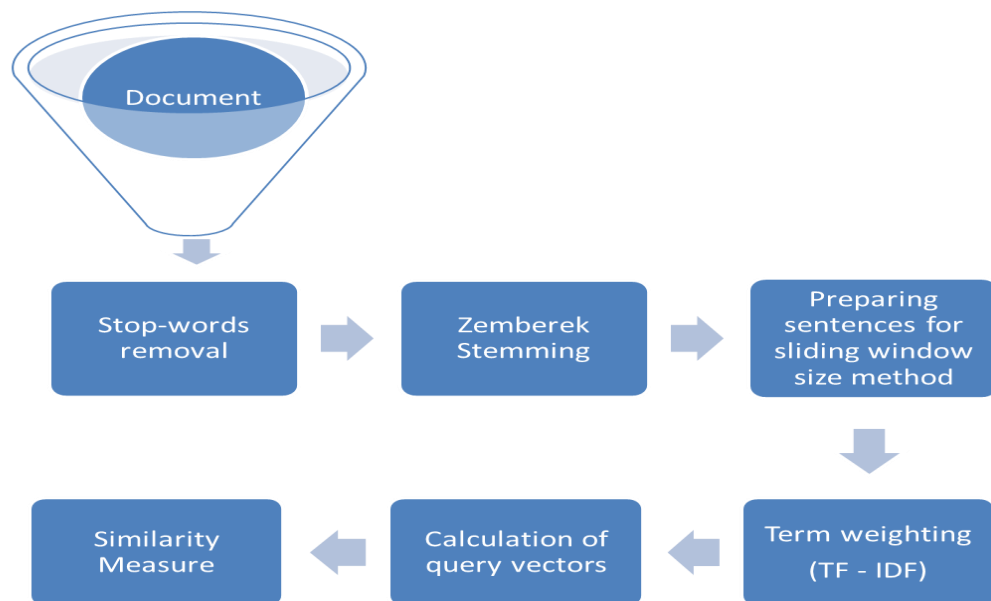


Figure 11 – Similarity Measure Processes for an Article

4.3 RESULTS

Our LSA based text coherence measurement system is performed on 400 newspaper article in Turkish. We can successfully measure sentences to sentences coherence values and overall local coherence values for all articles. The system is performed for SVD ranks 100, 150, 200, 250, 300 and for each rank approximation window sizes 1, 2, 3, 4 and 5 are applied. In total for each article, the system presents 25 different overall coherence values and 25 different sentence to sentence transition coherence values. The example of sliding window procedure coherence measurement method for window sizes with 100, 200 and 300 SVD rank approximations are presented in Figure 12, Figure 13, Figure 14, Figure 15 and Figure 16. We can say coherence breaks become smoother with increasing window size.

Figure 20, Figure 21, Figure 22, Figure 23 and Figure 24 presents coherence values for all articles. We can say coherence values increase and became more consistent with increasing window size. In the other hand, we can say our five SVD rank values do not provide significant difference on coherence values. So that, Turkish Wikipedia corpus is a large dataset and rank of SVD between 100 and 300 are optimal for that.

In Figure 17, Figure 18 and Figure 19 represents articles' coherence values for female and male distributions with 200 SVD ranks. We try to identify gender and author using Weka (Waikato Environment for Knowledge Analysis) software. It contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. In the study, naive bayes, j48 decision tree and multilayer perceptron algorithms are used to identify gender and the name of the authors. Identification results are presented in Table 8 and Table 9. With reference to identification results, we can say coherence values are not appropriate to identify gender and the name of the authors successfully.

Algorithm	Success
Multilayer Perceptron	%8.5
Naive Bayes	%9
Decision Tree	%6.25

Table 8 - Correctly Classify Articles for Author Identification

Algorithm	Success
Multilayer Perceptron	%56
Naive Bayes	%57.75
Decision Tree	%59.75

Table 9 - Correctly Classify Articles for Gender Identification

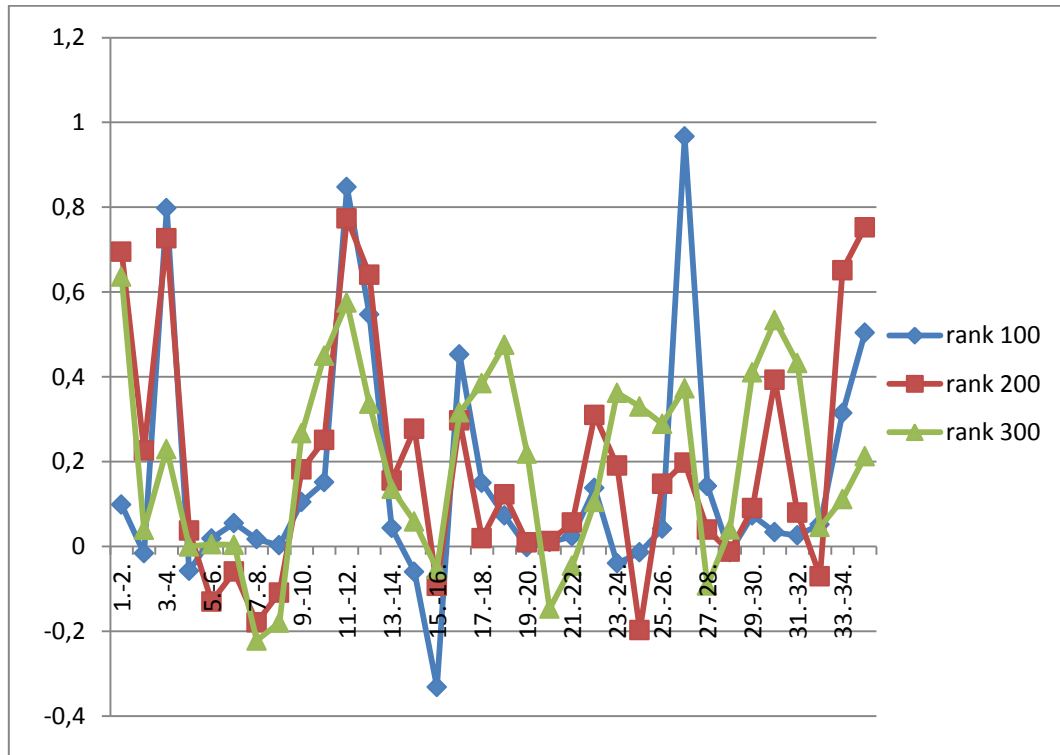


Figure 12 - Coherence Measures for Appendix B Article Sentences Window to Sentences Window with Three SVD Ranks, Sliding Window Size 1

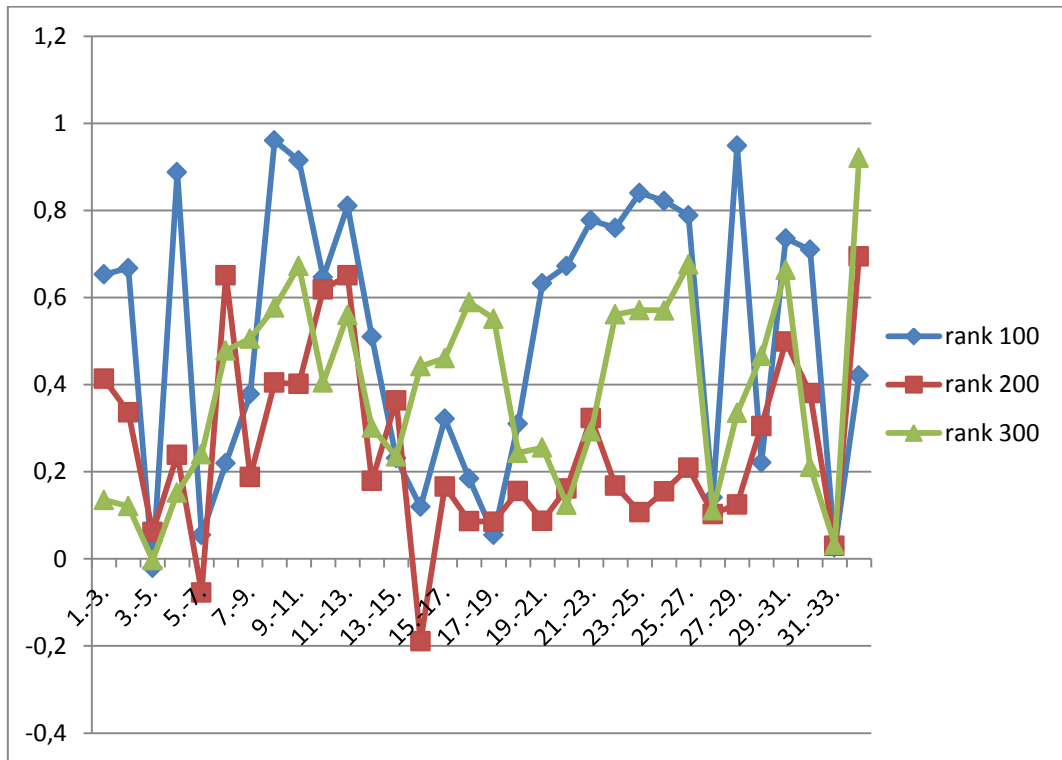


Figure 13 - Coherence Measures for Appendix B Article Sentences Window to Sentences Window with Three SVD Ranks, Sliding Window Size 2

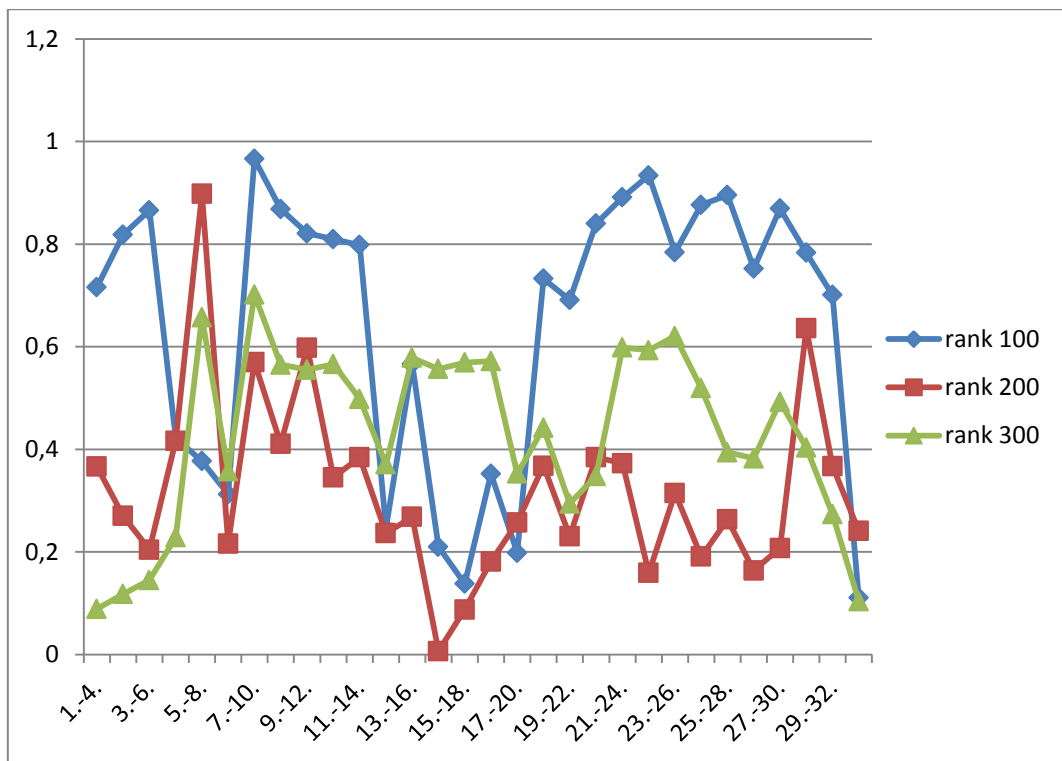


Figure 14 - Coherence Measures for Appendix B Article Sentences Window to Sentences Window with Three SVD Ranks, Sliding Window Size 3

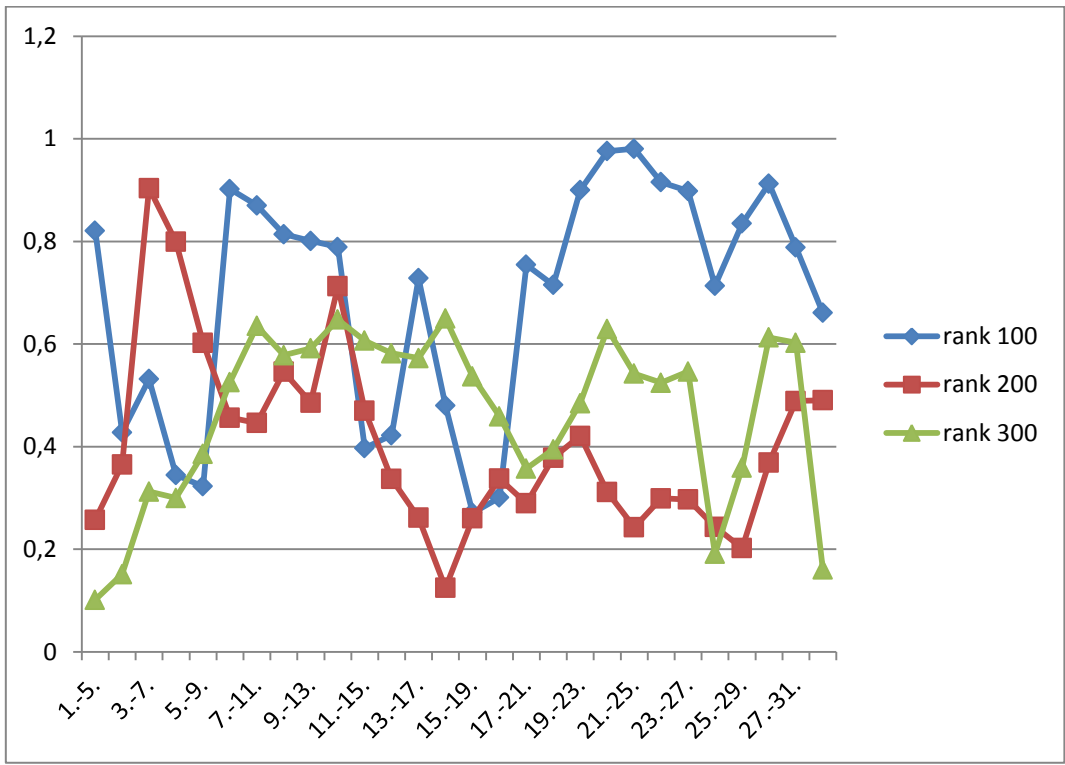


Figure 15 - Coherence Measures for Appendix B Article Sentences Window to Sentences Window with Three SVD Ranks, Sliding Window Size 4

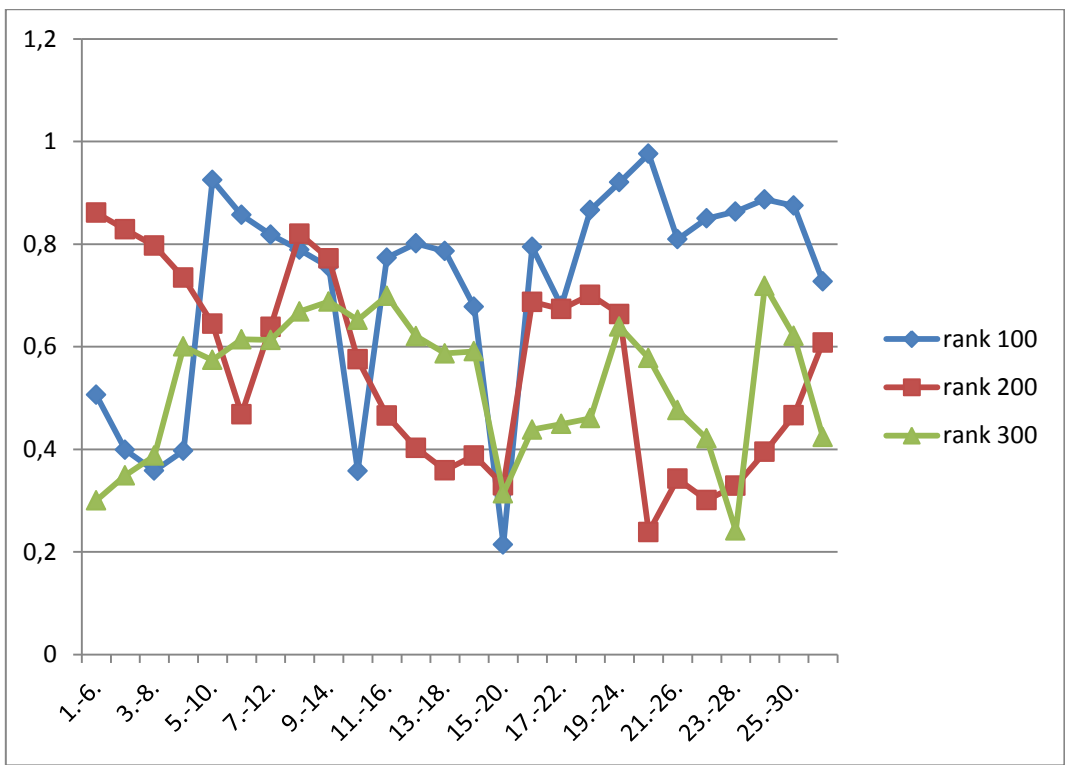


Figure 16 - Coherence Measures for Appendix B Article Sentences Window to Sentences Window with Three SVD Ranks, Sliding Window Size 5

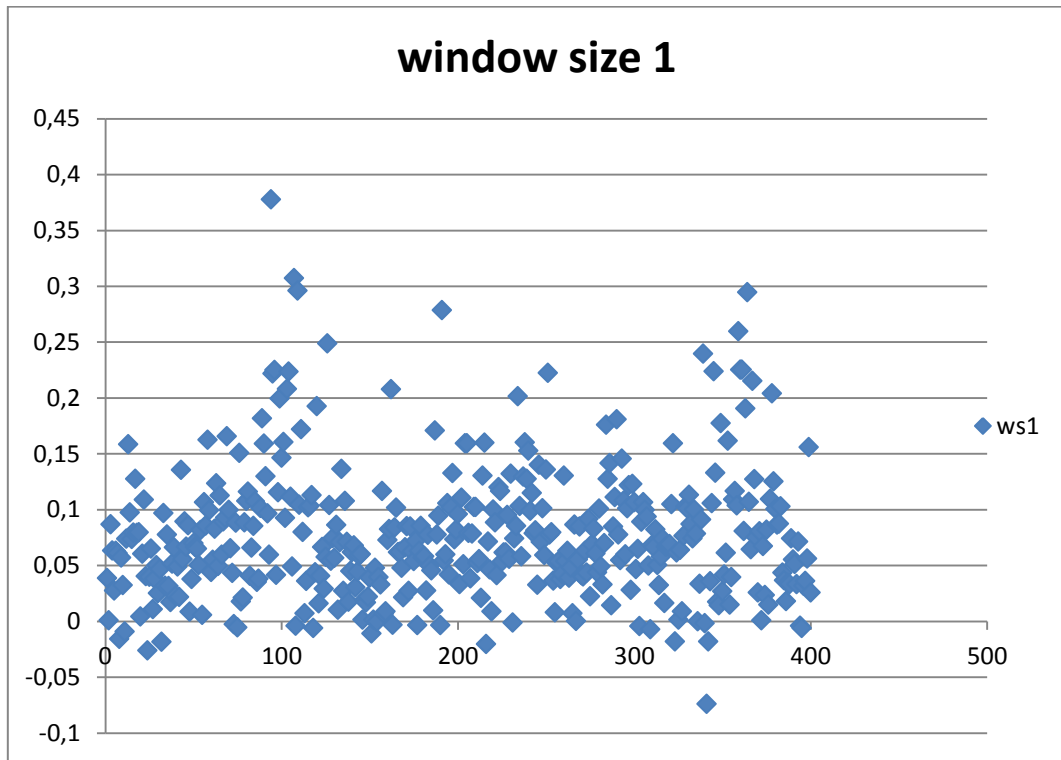


Figure 17 - Coherence Measures for All Articles with SVD Rank 200, Sliding Window Size 1 (1-200 Female / 201-400 Male)

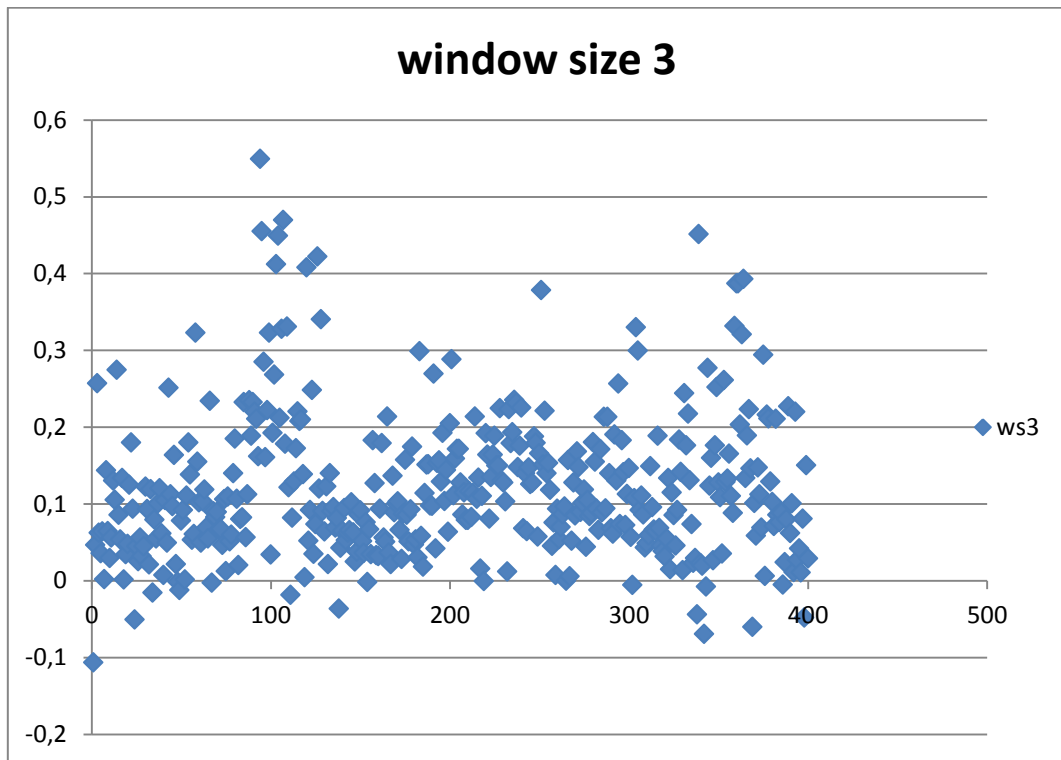


Figure 18 - Coherence Measures for All Articles with SVD Rank 200, Sliding Window Size 3 (1-200 Female / 201-400 Male)

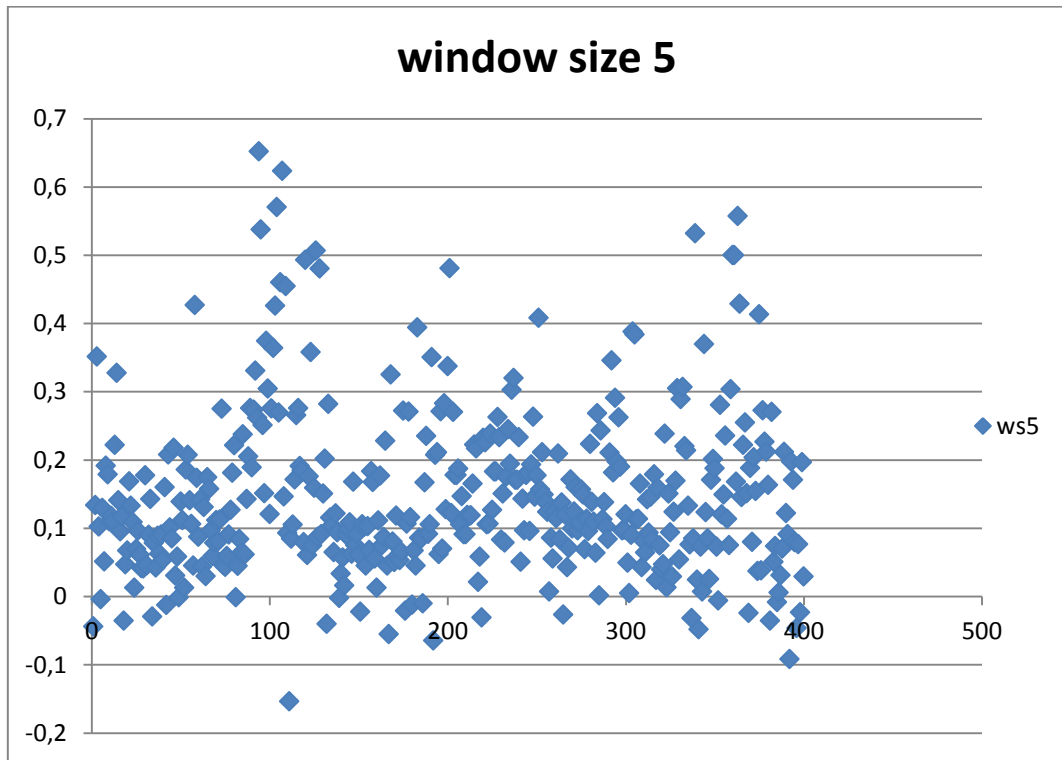


Figure 19 - Coherence Measures for All Articles with SVD Rank 200, Sliding Window Size 5 (1-200 Female / 201-400 Male)

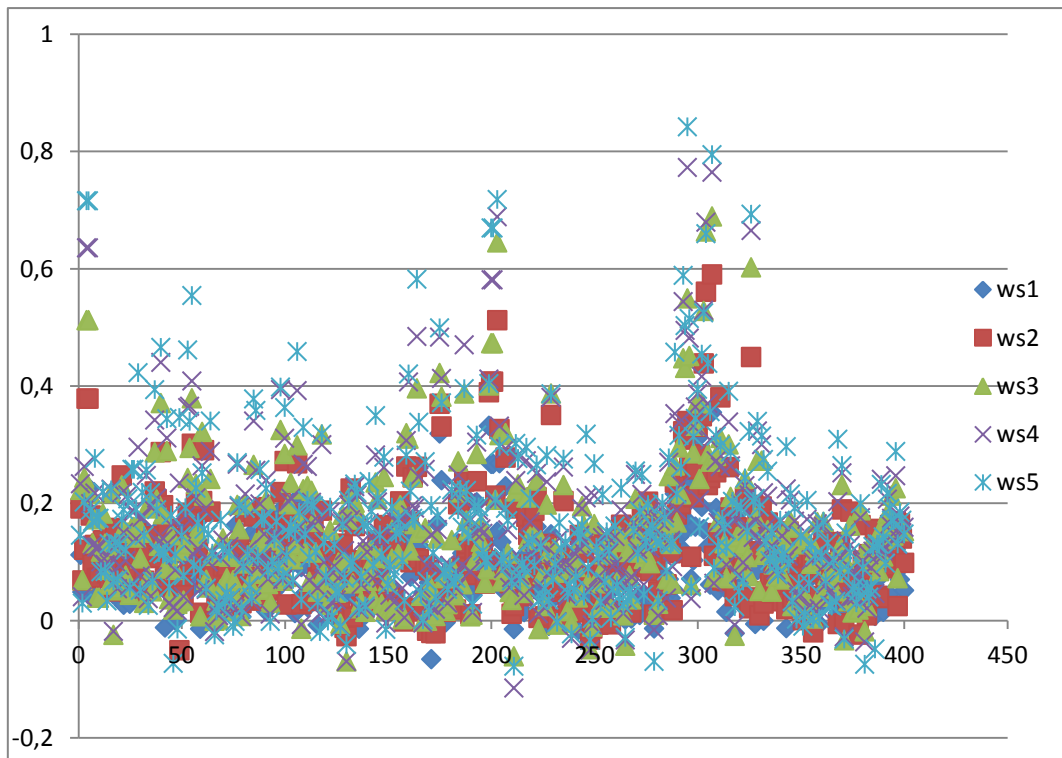


Figure 20 - Coherence Measures with SVD Rank 100 for 400 Newspaper Articles

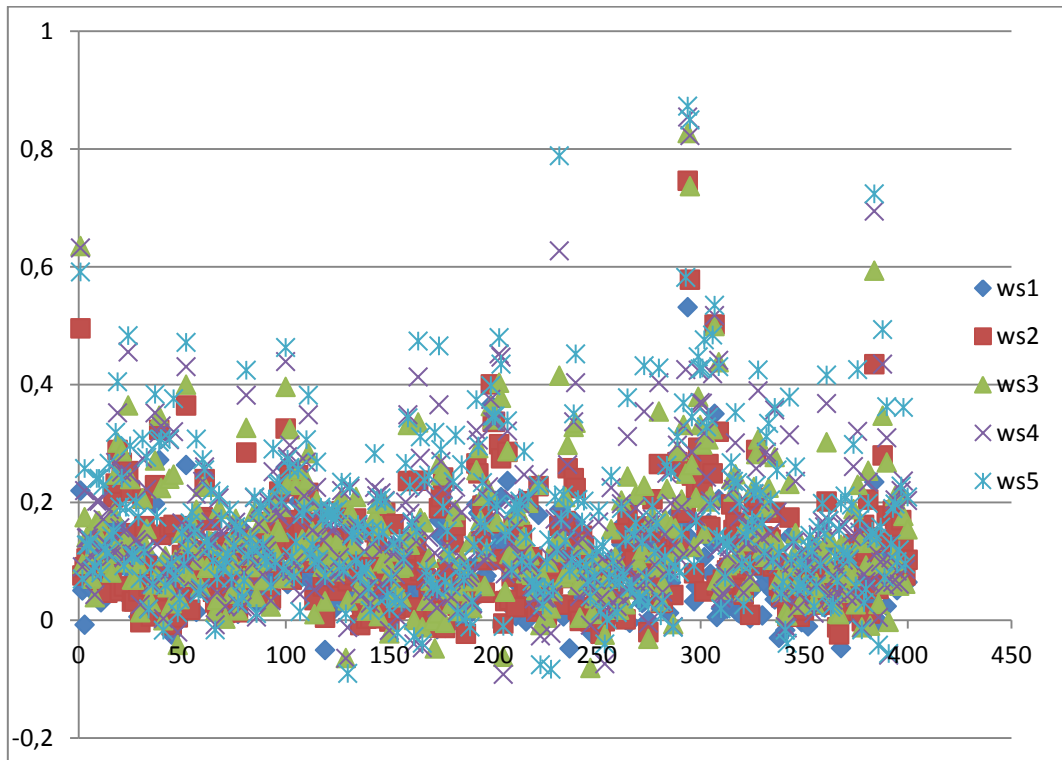


Figure 21 - Coherence Measures with SVD Rank 150 for 400 Newspaper Articles

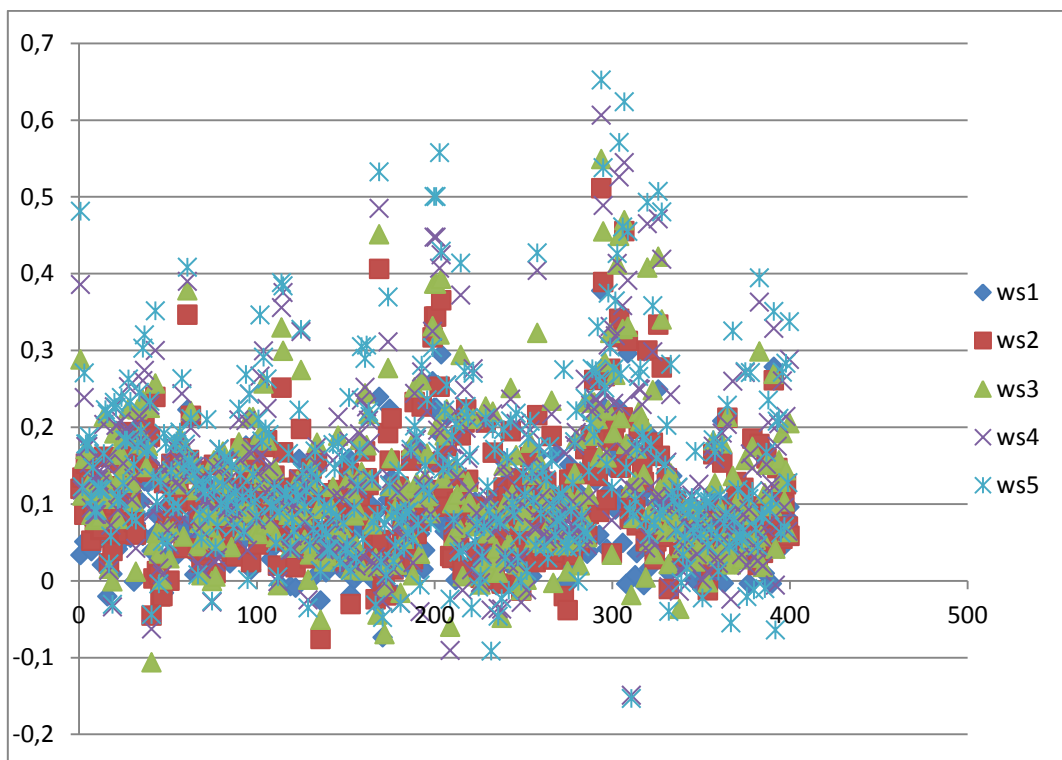


Figure 22 - Coherence Measures with SVD Rank 200 for 400 Newspaper Articles

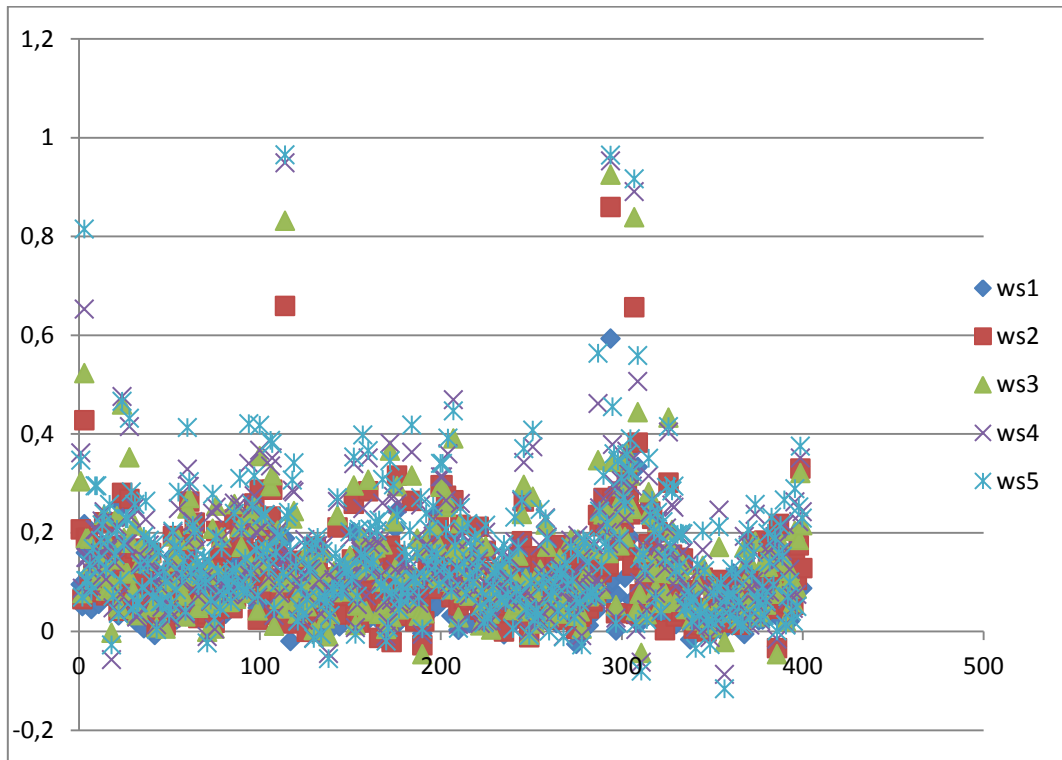


Figure 23 - Coherence Measures with SVD Rank 250 for 400 Newspaper Articles

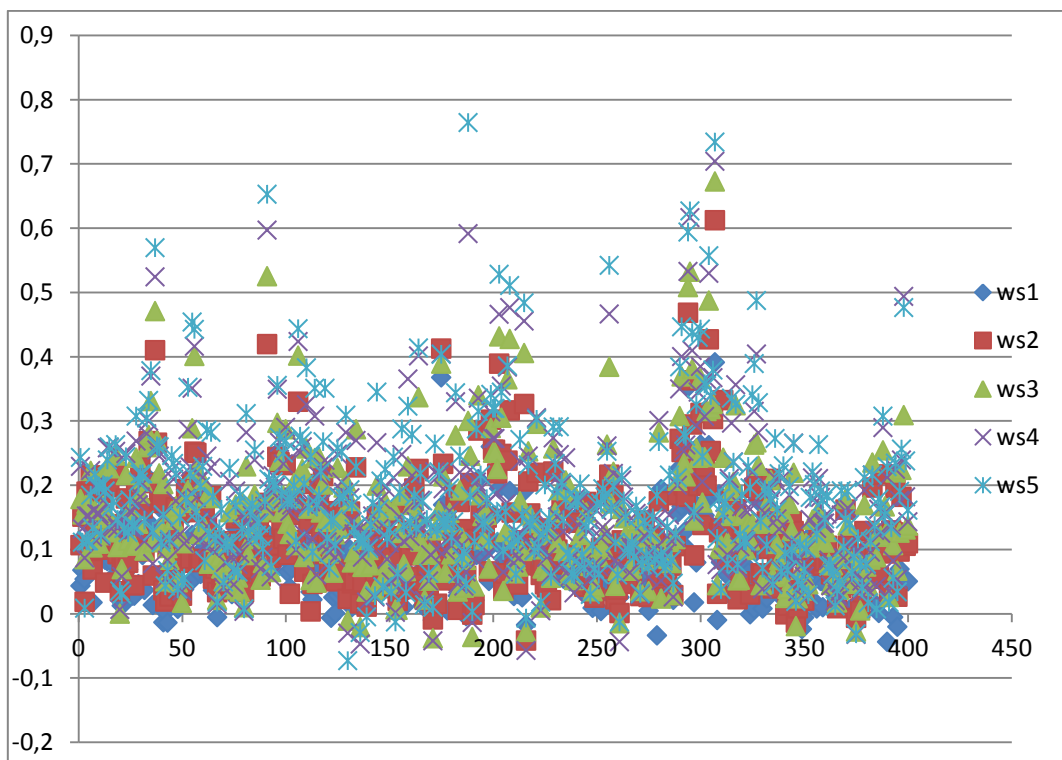


Figure 24 - Coherence Measures with SVD Rank 300 for 400 Newspaper Articles

CHAPTER V

CONCLUSION

Latent semantic analysis uses a reduced rank vector space model to identify the relation between the latent semantic structure of the term and the document. In this study, textual coherence and measurement approaches in literature are examined. Textual coherence measurement method based on the LSA is explained; then, a coherence measurement system is designed for LSA. The system provides coherence calculations for Turkish datasets. We experimented with a variety of SVD rank approximation values and sliding window procedure effects on Turkish articles. Also we have discussed the availability of gender and the name of the authors' identification using the article coherence values.

The results of our work shows that sliding window procedure and the suggested SVD rank approximation values in literature are suitable on Turkish documents like on English documents. Although one of the important results of this study is that the coherence values of documents are not solely enough to identify gender and the name of the authors, with a better formed corpus that is well formed for LSA may affect the results of these measurements in a different way. In addition to this suggestion, a specific domain based corpus may be used which the smallest space corpus which contained the most of the terms used in the target articles.

REFERENCES

- [1] Asher, N., Lascarides, A. (2003), *Logics of Conversation*, Cambridge University Press, Cambridge.
- [2] Barzilay, R., Lee, L. (2004), *Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization*, In Proceedings of HLT-NAACL, 113-120.
- [3] Barzilay, R., Lapata, M. (2005), *Modeling local coherence: an entity-based approach*, In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), 141–148.
- [4] Barzilay, R., Lapata, M. (2008), *Modeling local coherence: An entity-based approach*, Computational Linguistics, 34, 1–34.
- [5] Berry, M.W., Browne, M. (2005), *Understanding Search Engines: Mathematical Modeling and Text Retrieval (Software, Environments, Tools), Second Edition*, Society for Industrial and Applied Mathematics Philadelphia, PA, USA.
- [6] Brennan, S. et. al. (1987), *A Centering approach to pronouns*, In Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics, Stanford, California, 155-162.
- [7] **Cohmetrix Tool** (2013), Retrieved April 22, 2013, from <http://cohmetrix.memphis.edu/cohmetrixpr/cohmetrix3.html>
- [8] Cook, G. (1989), *Discourse*, Oxford University Press.
- [9] De Beaugrande, R., Dressler, W. (1996), *Introduction to Text Linguistics*, New York, 84-112.
- [10] Deerwester, S. et. al. (1990), *Indexing by latent semantic analysis*, Journal of the American Society for Information Science, 41, 391-407.
- [11] Elsner, M. et. al. (2007), *A unified local and global model for dis-course coherence*, In Proceedings of the Conference on Human Language Technology and North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2007).
- [12] Ferstl, E.C., von Cramon, D.Y. (2001), *The role of coherence and cohesion in text comprehension: an event-related fMRI study*, Cognitive Brain Research, 11(3), 325-340.

- [13] **Foltz, P. W.** (1996), *Latent Semantic Analysis for text-based research*, Behavior Research Methods, Instruments and Computers. 28(2), 197-202.
- [14] **Foltz, P.W.** et. al. (1998), The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, 25, 2&3, 285-307.
- [15] **Foltz, P. W.** (2007), Discourse Coherence and LSA, *LSA: A Road to Meaning*, ed. T. K Landauer, W. Kintsch, D. McNamara & S. Dennis, Lawrence Erlbaum Publishing.
- [16] **Givon, T.** (1993), English Grammar: A Function-Based Introduction, *John Benjamins Publishing*, Vol. 2.
- [17] **Graesser, A.** et. al. (2004), *Coh-Matrix: Analysis of Text on Cohesion and Language*, Behavior Research Methods Instruments and Computers, 36, 193-202.
- [18] **Grosz, B., Sidner, C.** (1986), *Attention, Intentions, And The Structure Of Discourse*, Computational Linguistics, 12(3), 175-204
- [19] **Grosz, B.** et. al. (1995), *Centering: A Framework for Modeling the Local Coherence of Discourse*, Computational Linguistics, 21(2), 203-225.
- [20] **Halliday, M.A.K., Hasan, R.** (1976), *Cohesion in English*, Longman Group Ltd, London.
- [21] **Horn, L.** (1986) Presupposition, theme and variations, *In Chicago Linguistics Society*, 168-192, Vol. 22.
- [22] **Hovy, E.** (1988), *Planning Coherent Multisentential Text*, In Proceedings of ACL, 163-169.
- [23] **Karypis, G., Han, E.** (2000), *Fast Supervised Dimensionality Reduction Algorithm with Applications to Document Categorization and Retrieval*, *Proceedings of CIKM-00*, 9th ACM Conference on Information and Knowledge Management.
- [24] **Landauer, T.K., Dumais, S.T.** (1997), *A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge*. *Psychological Review*, 104(2), 211-240.
- [25] **Landauer, T.K.** et. al. (1998), Learning human-like knowledge by singular value decomposition: A progress report, *Advances in Neural Information Processing Systems 10*, ed. M.I. Jordan, M.J. Kearns, and S.A. Solla, The MIT Press, Cambridge, MA, 45-51.
- [26] **Landauer T.K.** et.al. (2007), *Handbook of Latent Semantic Analysis*, Psychology Press.

- [27] **Letsche, T. A., Berry, M. W.** (1997), *Large-scale information retrieval with latent semantic indexing*, Information Sciences, 100, 105-137.
- [28] **Lin, Z.** et. al. (2011), *Automatically Evaluating Text Coherence Using Discourse Relations*, In The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA, 997-1006.
- [29] **Lin, Z.** et. al. (2012), *Combining Coherence Models and Machine Translation Evaluation Metrics for Summarization Evaluation*, In Proceedings of ACL (1), 1006-1014.
- [30] **Madnani, N.** et. al. (2007), *Measuring Variability in Sentence Ordering for News Summarization*, In Proceedings of the 11th European Workshop on Natural Language Generation.
- [31] **Mani, I.** (2001), *Automatic Summarization*, John Benjamins Publishing, Amsterdam, The Netherlands.
- [32] **Martin, D.I., Berry, M.W.** (2007), *Mathematical Foundations Behind Latent Semantic Analysis*, In *Handbook of Latent Semantic Analysis*, ed. T.K. Landauer, D.S. McNamara, S. Dennis, W. Kintsch, Psychology Press .
- [33] **Miltsakaki, E.** (2003), *The Syntax-Discourse Interface: Effects of the Main-Subordinate Distinction on Attention Structure*, Ph.D. Dissertation, University of Pennsylvania.
- [34] **Miltsakaki, E., Kukich, K.** (2004), *Evaluation of Text Coherence for Electronic Essay Scoring Systems*, Natural Language Engineering 10(1), 25-55.
- [35] **Reinhart, T.** (1981), *Pragmatics and linguistics: An analysis of sentence topics*, Philosophica, 27, 53-94.
- [36] **Renkema, J.** (2004), *Introduction to Discourse Studies*, John Benjamins Publishing.
- [37] **Salton, G., McGill, M.** (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.
- [38] **Soricut, R., Marcu, D.** (2006), *Discourse generation using utility trained coherence models*, In Proceedings of the COLING/ACL Main Conference Poster Sessions, 803-810.
- [39] **SVDLIBC** (2013), Retrieved April 22, 2013 from <http://tedlab.mit.edu/~dr/SVDLIBC/>
- [40] **Taboada, M., Zabala, L.H.** (2008), *Deciding on Units of Analysis within Centering Theory*, Corpus Linguistics and Linguistic Theory, 4(1), 63-108.

[41] **Walker, M.A.** et. al. (1998), *Centering Theory in Discourse*, Oxford University Press.

[42] **Weigand, E.** (2009), *Language as Dialogue: From Rules to Principles*, John Benjamins Publishing.

[43] **Weiser, I.** (1996), Linguistics, *Encyclopedia of Rhetoric and Composition*, ed. T. Enos, Taylor & Francis.

[44] **Yates, R. B., Neto, B. R.** (1999), *Modern Information Retrieval*. ADDISON-WESLEY, New York.

[45] **Zemberek** (2013), Retrieved April 22, 2013 from [http://tr.wikipedia.org/wiki/Zemberek_\(yazılım\)](http://tr.wikipedia.org/wiki/Zemberek_(yazılım)).

APPENDIX A

TURKISH STOPWORDS

a, acaba, altı, ama, ancak, artık, asla, aslında, az, b, bana, bazen, bazı, bazıları, bazısı, belki, ben, beni, benim, beş, bile, bir, birçoğu, birçok, birçokları, biri, birisi, birkaç, birkaçı, birşey, birşeyi, biz, bize, bizi, bizim, böyle, böylece, bu, buna, bunda, bundan, bunu, bunun, burada, bütün, c, ç, çoğu, çoğuna, çoğunu, çok, çünkü, d, da, daha, daki, de, deki, değil, de, ek, diğer, diğeri, diğerleri, diye, dokuz, dolay, dört, e, elbette, en, f, fakat, falan, felan, filan, g, gene, gibi, ğ, h, hal, hangisi, hani, hatta, hem, henüz, hep, hepsi, hepsine, hepsini, her, her biri, herkes, herkese, herkesi, hiç, hiç kimse, hiçbir, hiçbirine, hiçbirini, ı, i, için, içinde, iki, ile, ise, işte, j, k, kaç, kadar, kendi, kendine, kendini, ki, kim, kime, kimi, kimin, kimisi, l, m, madem, mı, mu, mü, n, nasıl, ne, ne kadar, ne zaman, neden, nerde, nerede, nereden, nereye, nesi, neyse, niçin, niye, o, on, ona, ondan, onlar, onlara, onlardan, onların, onların, onu, onun, orada, oysa, oysaki, ö, öbürü, ön, önce, ötürü, öyle, p, r, rağmen, s, sana, sekiz, sen, senden, seni, senin, siz, sizden, size, sizi, sizin, son, sonra, ş, şayet, şey, şeyden, şeye, şeyi, şeyler, şimdi, şöyle, şu, şuna, şunda, şundan, şunlar, şunu, şunun, t, ta, tabi, tamam, tüm, tümü, u, ü, üç, üzere, v, var, ve, veya, veyahut, y, ya, ya da, yani, ye, yedi, yerine, yine, yoksa, z, zaten, zira.

APPENDIX B

SAMPLE NEWSPAPER COLUMN

Çocuk üniversiteleri

Dünyada, bilim toplumu yaratmak için uygulanan çok farklı projeler var. Bunlardan birisi de çocuk üniversiteleri.

Amaç olabildiğince küçük yaşlarda, çocukları bilimle tanıştırmak ve üniversiteye yönlendirmek. Çocuk üniversitelerinin ilköğretim öğrencilerine yönelik olanı da var orta öğretime yönelik olanları da.

Eskiden başka ülkelerde olduğunu duyar, neden bizde de yok diye iç geçirirdik. Ama son yıllarda bizde de çok güzel örneklerini görmeye başladık.

Çocuk üniversiteleri, ABD gibi dünya bilimine en fazla katkıyı sağlayan ülkelerin, olmazsa olmazlarının başında geliyor.

Sadece ciddi finansal destek sağlamakla kalmıyor, yaygınlaştırılması için her türlü çabayı gösteriyorlar.

Yani dayatmaya dayalı yönlendirme yerine, bilime yönelik bilgilendirme, sevdirmeye ve özendirme söz konusu.

Başka türlü de zaten bilim toplumu olunmuyor.

Üniversitelerimizin bu konudaki çabaları takdire şayan.

Şu anda İstanbul, Ankara ve İnönü üniversiteleri bu konuda öncü durumunda. Adana, Bursa ve İzmir’de de bu yönde çalışmalar sürüyor.

Bilim Müzeleri de artık hayal olmaktan çıktı. Tek tük de olsa açılmaya başlandı. Ama çok yetersizler.

MEB VE TÜBİTAK'ın çocuklara yönelik bilimi sevdirmeye çabalarının eskiye göre daha iyi olduğunu söylemek ise abartılı olur. Çünkü akılları başka yerlerde.

Var olan cılız çabalar da yok olmaya başladı.

Çok yazık!..

APPENDIX C

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Kışlacık, İbrahim

Nationality: Turkish (TC)

Date and Place of Birth: 1 September 1987, Afyonkarahisar

Phone: +90 505 635 95 07

email: ibrahimkislacik@hotmail.com

EDUCATION

Degree	Institution	Year of Graduation
BS	Çankaya Univ. Computer Engineering	2010
High School	Afyonkarahisar Kocatepe Anadolu Lisesi	2005

WORK EXPERIENCE

Year	Place	Enrollment
2010 -	Dirisoft Bilgi ve İletişim Teknolojileri Ltd. Şti.	Computer Engineer

FOREIGN LANGUAGES

Advanced English, Beginner Spanish