MEASURING DIGITAL CONTENT: THE IMPACT OF CATA ON DIGITAL MEDIA ANALYSIS*

Assist.Prof.Dr. Aykut Arıkan
Yeditepe University,
Faculty of Communication,
Dept. of Journalism, Turkey.
aarikan@yeditepe.edu.tr

1. Introduction

In 2007, a group of researchers at Yeditepe University decided to develop a computer assisted content analysis system for Turkish media content. The project was proposed to TUBITAK (The Scientific and Technological Research Council of Turkey) in form of a scientific research project, and supported by TUBITAK under the project number 107K209.

The purpose of the project was to develop a Turkish based computer assisted content analysis system to analyze the content of the Turkish press of the Republican Era. The Turkish based computer assisted content analysis system provides a basic approach to meet the requirement of a computer assisted system for the analysis of Turkish content which has ever been an infrastructural problem for the analysis of Turkish newspapers using Latin based modern Turkish script after 1928.

The purpose of this paper is to deliver primary scientific information on the project and to discus methodological, scientific and applied problems and issues related to digital media analysis.

2. Methodology

The nature of the paper is descriptive by making use of scientific methods and techniques such as historical description, systems modeling, and the like. The application of content analysis, linguistic, and information retrieval methodologies will be the focal point of the discussion.[1]

3. Findins

Findings of the project will be delivered starting with a description of the problem of measuring content followed by a description of the historical context and definitions of the subject. A brief summary of the current situation in the field will be explained by discussing approaches and applications by giving special focus on computerized, linguistic, and information retrieval approaches to content analysis. The project 107K209 will be discussed in details eventually.

3.1 The Problem of Measuring Digital Content

Science is analysis. One may never consider some scientific action without analyzing. Analyzing is the central step of scientific inquiry. Either by observing, or by experimenting, or empirically in other words, science is trying to reveal provable data and transform it into information to eliminate consciousnesses in mankind's knowledge. Scientific analysis is the act of acquiring data from empirical inquiry and transforming it into an objective form, i.e. into information. Thus analysis may be distinguished as generating information from empirical inquiry.

Data as a raw knowledge object has no other meaning than itself. For instance, the datum "three" means only "three". And the meaning "three" is scientifically nonsense on its own, until it is interrelated with another datum, such as "mile", "page", "word", "degree", etc. The interrelation of these two data is simply a measure. According to this, data can be "acquired" simply by "counting", whereas information is "generated" by "measuring". As a matter of fact, counting is to *determine the total number of something while measuring is to ascertain the size, amount, or degree of something by using an instrument or device marked in standard units or by comparing it with an object of known size.*

Many scholars of communications fall recurrently into the trap of intermingling counting and measuring in media analyses. Analyzing media is measuring media texts, or content in other words, by using an

* Based on the research project 107K209 supported by TUBITAK.

[1] Along with the discussion, the special impact of Computer Assisted Text/Content Analysis (CATA) on digital media analysis will be studied by delivering results of the project 107K209.

instrument or device marked in standard units or by comparing it with and object of known size. In other words, to analyze a communication text, or content, we either need instruments or devices marked in standard units, or we need objects of known sizes for comparison. This is the basic reason that lies behind the usage of content analysis as a research technique for analyzing media.

Arthur Asa Berger points out two critical tribulations of content analysis among a range of problems (2005, 134):

1. Representation problems of samples.
2. Finding measurable units.

Both tribulations are remarkably critical in measuring digital content, because of its distributed, multi-lateral (various types and kinds) and enormously large structure. Therefore, digital content is easy to collect but difficult to store; it is easy to group but difficult to classify, and very easy to count but very difficult to measure.

As to be confirmed by the historical context and definition proposals delivered by various researchers over a long period of time, content analysis enables inexpensive, easy, unobtrusive, quantitative, media research for both past and current events, according to Berger (2005, 134).

3.2 Historical Context and Definitions

Historically, the progress of the idea of analyzing content is reflected on definition proposals principally delivered in the twentieth century. Krippendorff traces back early efforts towards the applications of systematic content analysis to "inquisitorial pursuits by the Church in the 17th century" (Krippendorff, 2004, 3) predating modern content analyses of the 20th century for some three centuries.

Albeit, as a matter of fact, Max Weber offers a "large-scale content analysis of the press" "at the first meeting of German Sociological Society in 1910 (Krippendorff, 2004, 4), prior to this, the first quantitative use of content analysis is published by Gilmer under the name "Do newspapers now give the news?" in 1893 (Krippendorff, 2004, 49).

In contrast to these original accomplishments the term does not appear in English until 1941 (Waples & Berelson, 1941, 2; cited in Krippendorff, 2004, 3). Another remarkable application of the subject in the same episode is the circumstantiation of an US magazine accused of making Nazi propaganda parallel to a German radio station in the 1940's (Tavsancil & Aslan, 2001, 27).

Lasswell's contribution to the field, by analyzing newspapers' content to identify propaganda (Lasswell, 1927), assisted Berelson to develop a first definition of the topic (Tavsancil & Aslan, 2001, 27) as "a research technique for the objective, systematic, and quantitative description of the manifest content of communication" elucidating that the extent of the this definition may be expanded as "…description of human behavior, particularly linguistic" (Berelson, 1952, 18).

1953 Dorwin Cartwright accentuated that content analysis is a systematic, quantitative, and objective method to describe all kinds of symbolic behavior and to analyze their content (Cartwright, 1953; cited in Bilgin, 2006, 2).

On the other hand, Kerlinger defines content analysis at 1964 as "the study of a researcher on communication materials created by others, according to specific criteria" (Kerlinger, 1964, 539; cited in Tavsancil & Aslan, 2001, 18).

1966 Stone, Dunphy, Smith, & Ogilvie define Content Analysis as "any research technique for making inferences by systematically and objectively identifying specified characteristics with text (Stone et al., 1966, 5; cited in Neuendorf, 2002, 10).

Henry and Moscovici describe the method in 1968 as "an analytic tool with not much challenge and no theoretic basis" (Henry & Moskovici; cited in Bilgin, 2006, 2).

1969 Kolasa Blair describes the method as "a scientific approach to study the 'what' of different types of communication" (Blair, 1969; cited in Tavsancil & Aslan, 2001, 20), while David Fox describes it as "the classification, summarizing, and

the measurement of specific variables and concepts, according to a specific problem or intention, as well as the scanning and categorization of these to deliver a specific meaning" (Fox, 1969; cited in Tavsancil & Aslan, 2001, 20).

Subsequent to this in 1971, Thomas Carney describes content analysis as "the general purpose technique for posing questions to a 'communication' in order to get finding which can be substantiated. The 'communication' can be anything: A novel, some paintings, a movie, or a musical score-the technique is applicable to all alike and not only to analysis of literary material" (Carney, 1971, 52; cited in Neuendorf, 2002, 10)

1977 Laurence Bardin proposes that content analysis is the summing up of all analyses of communications (Bardin, 1971; cited in Bilgin, 2006, 2). Where Ghuiglione and Matalon emphasized that content analysis is a theoretically weak method but with remarkable outcomes in applications (Ghuiglione & Matalon, 1978; cited in Bilgin, 2006, 2).

As one of the most significant names of the field, Klaus Krippendorff formulated the definition of content analysis in 1980 as "a research technique for making replicable and valid inferences from data to their context" (Krippendorff, 1980, 21; cited in Neuendorf, 2002, 10) and then reformulated it in 2004 as "a research technique for making replicable and valid inferences from texts (or meaningful matter) to the contexts of the use" (Krippendorff, 2004, 18).

1983 Klaus Merten proposes the definition "a method that researches social reality by deriving non-manifest features of content from manifest features of contents of social reality" (Merten, 1983, 15; cited in Gokce, 2006, 18).

Robert Philip Weber defines content analysis in 1990 as "a research method that uses a set of procedures to make valid inferences from text" (Weber, 1990, 9) by referring to the study of Ogilvie, Stone, and Shneiderman (1966) that distinguished 'real' suicide notes from 'simulated' ones by using a computer-assisted content analysis system (Ogilvie, Stone, & Shneiderman, 1966; cited in Weber, 1990, 20).

1991 Arthur Asa Berger proposes the definition "a research technique that is based on measuring the amount of something (violence, negative portrayals of women, or whatever) in a representative sampling of some mass-mediated popular art form" (Berger, 1991, 25; cited in Neuendorf, 2002, 10).

Riffle, Lacy, and Fico define quantitative content analysis in 1998 as "the systematic and replicable examination of symbols of communication, which have been assigned numeric values according to valid measurement rules, and the analysis of relationships involving those values using statistical methods, in order to describe the communication, draw inferences about its meaning, or infer from the communication to its context, both of production and consumption" (Riffle, Lacy, & Fico, 1998, 20; cited in Neuendorf, 2, 10). A significant contribution that emphasizes on the statistical approach and semantics by centralizing context.

As one of the key contributions to the field, Kimberly Neuendorf delivers in her Content Analysis Guidebook the definition "a summarizing, quantitative analysis of messages that relies on the scientific method (including attention to objectivity-intersubjectivity, a priori design, reliability, validity, generalizability, replicability, and hypothesis testing) and is not limited as to the types of variables that may be measured or the context in which the messages are created or presented"(Neuendorf, 2002, 10).

Krippendorff delivers the latest and most current definition in 2004 as "a research technique for making replicable and valid inferences from texts (or meaningful matter) to the contexts of the use" (Krippendorff, 2004, 18).

A multiplicity of further definition proposals in writings on research methods of psychology or education may be added to these definitions (Tavsancil & Aslan, 2001, 19-21).

In Turkey, most significant contributions to the field are literature provided by Tavsancil and Aslan (2001), Orhan Gokce (2006), and Nuri Bilgin (2006). Not to be overseen is the compilation of translated texts on content analysis by Murat S. Cebi (2003)

3.3 Current Situation: Approaches and Applications

Content analysis is one of the mostly trained research methods in schools of communication at undergraduate and graduate levels, due to its analytic and objective character dealing exactly with communication texts, and ease of application especially in the manual approach; even though these examples show a major constraint of capacity usually covering short periods in months-scale or single publications or limited subjects.

On the other hand, content analysis is used for an assortment of media types as described in Carney's definition; books, newspapers, magazines, radio and TV broadcasts, musical scores, drama scenarios or screenplays, or even handwritten notes as used by Ogilvie, Stone, and Shneiderman.

A noteworthy example-study from the United States is CIRCLE's (The Center for Information & Research on Civic Learning & Engagement) report named "News for a New Generation Report 1: Content Analysis, Interviews, and Focus Groups" by Sherr and Staples (2004).

Outstanding contributions from Europe are delivered by "The Centre for Survey Research and Methodology (ZUMA)" (a branch of the German Social Science Infrastructure Service), which also developed a computer-assisted content analysis system under the name "Textpack" (http://www.gesis.org/en/zuma/index.htm, as retrieved of 09.05.2008).

A contribution from Africa is a study performed in the University of Stellenbosch, South Africa funded by the University of Botwana (Mogotsi, 2007).

Nowadays the method is also applicable to a wide range of research fields for instance such as market research (Anderson & Song, 2008).

A particular field of application of the subject is certainly computer-assisted content analysis (CATA). Special software developed for CATA has found itself a playground in scientific as well as commercial circles. Kimbely Neuendorf (2008) provides on the Content Analysis Guidebook Online web site an inclusive list of software for CATA. Züll and Landmann of ZUMA provide a comprehensive and annotated list of literature on applications of content analysis (Züll & Landmann, 2002 23-98).

3.4 Computerized Approach: CATA (Computer Assisted Content Analysis)

Klaus Krippendorff accentuates two basic reasons for CATA (Krippendorff, 2004, 258-259):

1. The ability of computers to process large volumes of data at high speed.
2. Their reliability in processing textual material.

The first argument provided analysts and researchers with the ability to enlarge the scope of content analysis beyond time, subject and different media sources both qualitatively and quantitatively. With CATA it is possible to include collections of newspapers covering decades or even centuries into the scope of analysis. On the other hand, qualitatively separated media types such as different books, and quantitatively separated media types such as magazines on different subject became due to be analyzed.

Reliability of CATA is in no way comparable to human coding. Of course, an extra period for machine learning is necessary for any CATA system. But this becomes insignificant compared to the reliability provided by CATA.

Another significant reason for CATA may be regarded as its ability to combine content analysis with statistical analysis capabilities. Because data delivered by the analysis of the content is due to be analyzed statistically to deliver correlative results and in-depth interpretations. This combination of abilities, of course, requires the integration of strong and specialized data-dictionaries as indices and stop-word lists, which becomes more-and-more difficult if not impossible with the increase of the content that is analyzed. Therefore a sophisticated CATA system requires powerful training data for adequately developed indices and stop-word lists as well as the flexibility to further train indices and stop-words with the expansion of content that is analyzed.

Regarding Berger's tribulations mentioned above, the most crucial contribution of CATA is its ability to process wide ranges of content, which makes sampling unnecessary. A simple reason that makes any kind of digital or digitized content measurable entirely. On the other hand, CATA itself delivers many options for measuring units, due to inter-relationality options.

3.5 Linguistic Approach to Content Analysis: NLP (Natural Language Processing) and Corpus Construction

A most significant amount of content (ort texts of communication) to be analyzed is based on language, i.e. natural language, such as all types of printed material (excluding musical scores), handwritings, audio recordings and broadcasts. Therefore natural Language Processing (NLP) is an important field related to content analysis.

Language based content to be analyzed is a lexical resource: texts, dictionaries, thesauri, and processing tools related to these (Manning & Schütze, 1999, 19). Collections of texts, designated as corpora (plural of the Latin term corpus) (Manning & Schütze, 1999, 6) make up the body or the object that is analyzed. Seen technically, for a CATA system this corpus is nothing but a database.

Data dictionaries, or data languages are according to Krippendorff, "descriptive devices in which terms analysts cast their data" (Krippendorff, 2004, 150), or Key-word-in-context (KWIC) lists after Weber (1990, 44), or so-called thesauri, are indeed, indices or Information-Retrieval languages (as a specific form of artificial languages) (Arikan, 2006, 79).

Problems in constructing information retrieval languages are basically results of natural language problems (Arikan, 2006, 151). Thus, the nature of the natural language of the material that is due to a content analysis is the basic source of problems in constructing an adequate information retrieval language, i.e. data dictionaries and indices.

There are several concept dictionaries as lexical databases that provide adequate terms for information retrieval languages derived from English, such as WordNet, and the like (Krippendorff, 2004, 279).

Turkish as a member of the Ural-Altaic language family owns a totally different morphological structure than English and other Indo-European languages. Therefore, any kind of information retrieval language, i.e. concept dictionaries, thesauri, indices, etc., must be uniquely developed for the analysis of Turkish based content.

3.6 Information Retrieval Approach to Content Analysis

Since descriptive devices to analyze text corpora are information-retrieval units, information retrieval becomes a central part of content analysis. Content, or text corpora are to be indexed, i.e. described analytically, for an adequate integration of content analysis and statistical analysis features into a sophisticated CATA system.

Of course machine generated lists of keywords are specific types of indices, although quite restricted and inflexible because of a series of language problems such as synonyms, acronyms, ant the like. Therefore human generated indices, described as "controlled vocabulary" in information science are always more reliable and flexible than computer generated ones. But the most deterministic consequence of human interaction is the decrease of processing speed. Computer-assisted morphological analysis and semantic analysis are solutions to minimize human interaction and to increase processing speed, but this approach requires a strong devotion to artificial intelligence.

Regarding Berger's tribulations mentioned above, indices are the very answer to the problem of finding measurable units, for indices are predefined, linguistic units for measuring.

3.7 Measuring Digital Content in Research Project Nr. 107K209

The purpose of the project 107K209 was to develop a Turkish based computer assisted content analysis system to analyze the content of the Turkish press of the Republican Era. The Turkish based computer

assisted content analysis system provides a basic approach to meet the requirement of a computer assisted system for the analysis of Turkish content which has ever been an infrastructural problem for the analysis of Turkish newspapers using modern Turkish script after 1928.

The original value of the project is its contribution to applications of content analysis. Lack in this field is a result of the negligence of natural language and problems of information retrieval based on natural language issues in content analysis research in Turkey. The project shall open frontiers in applications with its overall original contribution to the field.

The prevailing effect of this project shall be its direct effect to all social research starting with communications and historical research and to communications industries, by contributing to these with cost and time saving and affecting the depth of research.

The added value of the project may be taken in two ways: to accelerate research by enabling cost and time savings and by contributing to the depth of research with opportunities provided by information technology. Both added values have serious contributions to research and industries. As a result interdisciplinary research especially in social sciences will be fostered and new areas of service and new production lines shall come to life.

The project makes use of two scientific methods: content analysis and information retrieval. In addition to these, quality management is adopted, to reach effective results. All these methodological approaches are driven together as a system design within the systems approach.

The project process consists of six basic steps to be taken in 12 months: data (image) acquisition, software development, development of statistical analysis templates, index development, index management and test analyses.

The 107K209 project that is basically labor-intensive, consists of 11 people: one principal researcher (Asst. Prof. Dr. Aykut ARIKAN), two adjunct researchers (Asst. Prof. Dr. Deniz TANSI and Asst Prof. Dr. Nilufer HATEMI), one scholar, one statistical specialist, one systems manager, one data management specialist, one media scanning specialist, one media scanning employee, one quality management specialist and one information technologies specialist. The development of the CATA system was outsourced to Mr. Erman PEREMECI, of Peremeci Digital Solutions.

During the project, two strategic partnerships were established. First is the partnership with the Library and Documentation Center of the Istanbul University. The Center is a legal deposit library of Turkey that collects all printed material in Turkey. The second strategic partnership was established with Dragoman Language Technologies, a company that specializes in language and translation technologies. The company owns the Dragomansos.com electronic dictionary, which was an excellent substitute for a concept dictionary and lexical database. According to the cooperation agreement, the project 107K209 got granted the right to make use of the dictionary and further develop a thesaurus, which will be available for the use of the Company.

3.7.1 Project Cycle

The project was proposed in May 2007 to TUBITAK (The Scientific and Technological Research Council of Turkey, as he leading agency for management, funding and conduct of research in Turkey). A total budget of approx. 120 thousand YTL (equals approx. 100 thousand USD) for a total project plan of 12 moths was requested from TUBITAK and TUBITAK approved the total budget in September 2007.

The project started operating on October 15th 2007. Since that date, until the termination of the project, a team o two people continues acquiring newspaper content as photographic images. The project was terminated October 15th 2008, as scheduled.

The CATA system is developed and became fully operational in a timeframe of six months. The system consists of a Spider module to collect electronic news from the Internet, a Text Indexing module, an Image (Page) Analysis module, and a Statistical Analysis

and Reporting module. An OCR (Online Character Recognition) tool (ABBYY FineReader vers. 9) is converting photographically acquired newspaper images into text that is stored in a database together with news texts collected from the Internet as the text corpus of the project.

3.7.2 Research Constraints and Limitations

The basic constraint of the project was the amount of training data to be incorporated. The Printed newspapers Hürriyet, Milliyet, Sabah, Cumhuriyet and Türkiye were chosen for photographical content acquisition, and all national-scale newspapers providing their online content on the Internet starting with Zaman Online, as the first online newspaper of Turkey, were chosen for online news acquisition through the Spider module, to develop training data. This amount may seem relatively limited compared to printed news content available since 1928 (in modern Turkish script based on Latin letters). But even this relatively limited amount of training data causes a large workload for the project team.

3.8 Observations and Results

All modules of the CATA system became fully developed and operational as of May 2008. The project was basically carried out according to the project plan with minor adjustments and amendments that did not alter the project cycle. The main reason for this orderliness is the quality management approach used in management of the project. The PDCA (Plan-Do-Check-Act) approach, also known as the Shewhart cycle, was adopted as the main tool for quality assurance. Monthly quality control procedures were established and performed by the quality management specialist.

The project proposal prescribed the use of a drum-scanner provided by Yeditepe University. But this tool was inadequate for bounded newspaper volumes. Therefore printed content acquisitions were changed to photographic acquisition due to the alternative plan as proposed to TUBITAK. A digital DSL camera with a macro objective was purchased by Yeditepe University and assigned to the project, together with PC-workstation with high disc capacity.

Fig.1. Screenshot of the Spider Module.

The Spider module (please refer to Figure 1 for a screenshot) is still collecting online news from the Internet. The total number of news collected from the Internet is 441,168 as of May 2008.

Indices of the Text Indexing module are fully operational. The total number of keywords is 116,069 as of October 2008. For refractions of keywords please refer to Table 1 and Figure 2.

Type	Number	Percentage
Controlled Vocabulary (Indices)	65,586	43 %
Stop-words	50,483	57 %
TOTAL	116,069	100 %

Table 1. Refraction of Keywords

Controlled vocabulary, i.e. indices, consists of subjects, persons, places, and institutions. The number of subjects builds a strong block of 89 % of all indices, followed by persons with 6 %, regions with 3 %, and institutions with 2 %. For details please refer to Table 2 and Figure 3. For a screenshot of the Text Indexing Module please refer to Figure 4. Please notice that these figures are due to change with the increase of training data and the development of the corpus of news texts.

A special feature of the Text Indexing module is that interrelations resulting from combinations of all four

Index	Number
Personal Names	3,959
Institutions	1,411
Regions	1,970
Subjects	58,246
TOTAL	65,586

Table 2. Controlled Vocabulary (Indices)

indices, namely Persons, Regions, Institutions, and Subjects are stored numerically into the database while a news text is indexed. This is expected to cause an increase of processing speed of statistical analyses. On the other hand, association of indices seems to be what Vannevar Bush projected with its idea of the Memex; because associative indexing is according to Bush the basic idea behind the Memex (Bush, 1837; reprinted in the Atlantic.com, retrieved 06.05.2008).

To define statistically significant correlations between subjects is based on the Pearson product-moment correlation coefficient (PMCC). Whereas following formula was used:

X= index term as independent variable
Y= index term as dependent variable
n= time period (month, year)

$$r = \frac{1}{n-1} \sum \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

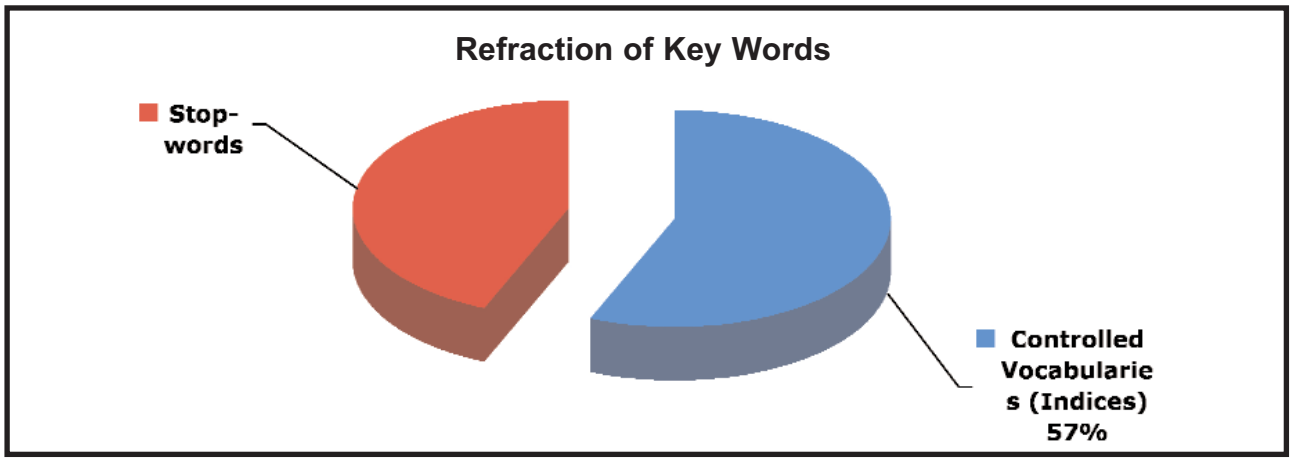

Fig.2. Refraction of Keywords

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

Where sx and sy stands for standard deviation and standard deviations may be found with following formula:

The result must be between -1 and 1. Results may be interpreted according to Table 3, where negative results indicate inverse correlation.

Correlation	Negative	Positive
Weak	-0.3, -0.1	0.1, 0.3
Medium	-0.5, -0.3	0.3, 0.5
Strong	-1.0, -0.5	0.5, 1.0

Table 3. Interpretation of Results

The Statistical Analysis and Reporting Module (Figure 5, Figure 6, Figure 7) is based on the analysis of the Pearson product-moment correlation coefficient approach. The corpus of the system does still not incorporate enough data do deliver significant statistical results, and will most probably not even within the lifetime of this project, because collected data in the corpus is only for data training purposes. To reach statistically significant results, the corpus must consist of a closed, complete, and well-defined data set, such as the full coverage of the Turkish Press at national level since 1928 until nowadays, which is far beyond the scope of this project. This project delivers a tool for content analysis of the Turkish Press, but not the analysis itself, which is a quite huge, costly and time-consuming process.

3.9 Future Outlook of the Subject

The full coverage content analysis of the Turkish Press at national level since 1928 until nowadays, requires, according to manpower and processing speed calculations of the 107K209 project, a huge amount of manpower, i.e. approximately 150 man/month. Therefore the researchers proposed a consequential project covering the period of 1950-2010 with three photographic acquisition facilities operating in double-shifts. The decreasing number of pages backward in time is a strong proof that this project organization scheme looks compatible with TUBITAK's 36 months restriction for research projects. Most probably this time frame will provide

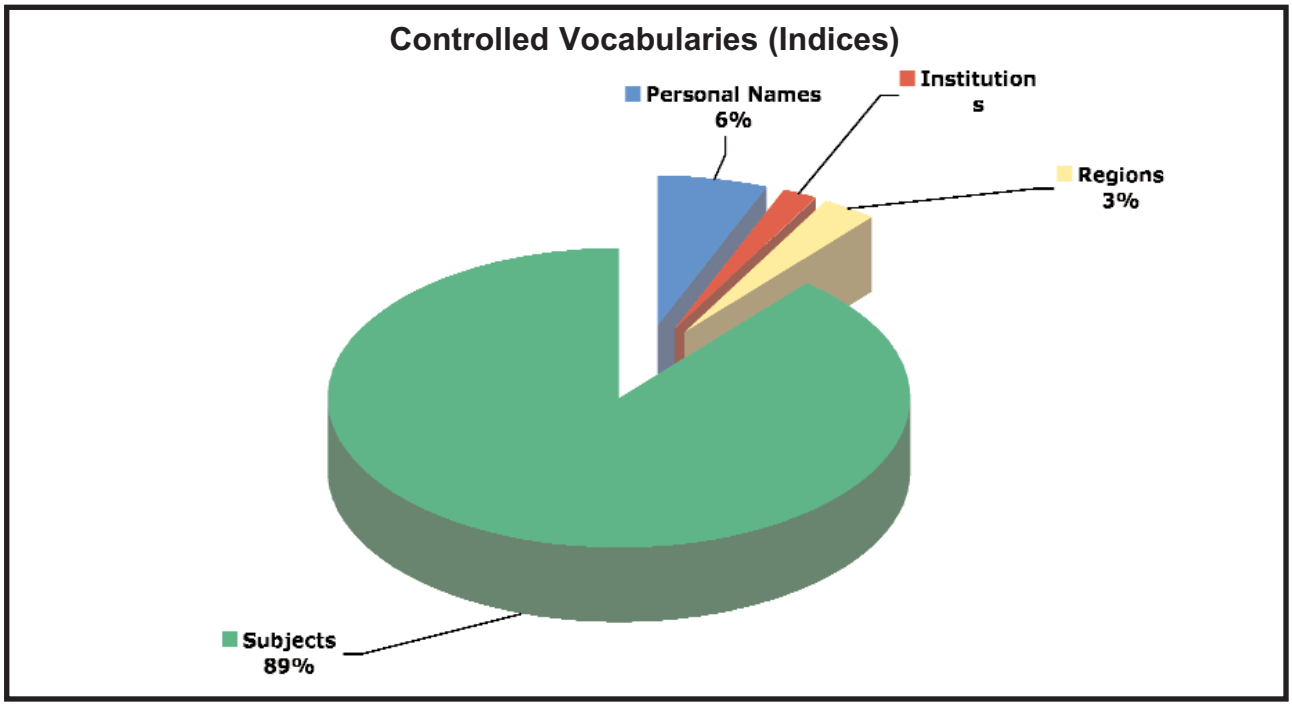

Fig.3. Controlled Vocabulary (Indices)

extra resources to go deeper in time. Extra time needed to cover everything since 1928 will be proposed to TUBITAK as an expansion or as another consequential project.

Meanwhile the researches have formed a research facility called Contextual Data Research Laboratory (Data-Lab), subsequent to the Vice-Rectorate Responsible for Institutional Advancement of Yeditepe University. Data-Lab that has become the hosting institution of the 107K209 project has applied for a FP7 REGPOT Grant.

The CATA system is available for the Turkish as well as for the international research community, operating under Data-Lab. Research request are welcome. One example will be an analysis of theses in social sciences.

Further research will be possible in areas such as trend research, online reputation analysis, environmental scanning, and the like.

4. Conclusion

Measuring digital content has always been a problem of media analysis. To analyze a communication text, or content, we either need instruments or devices

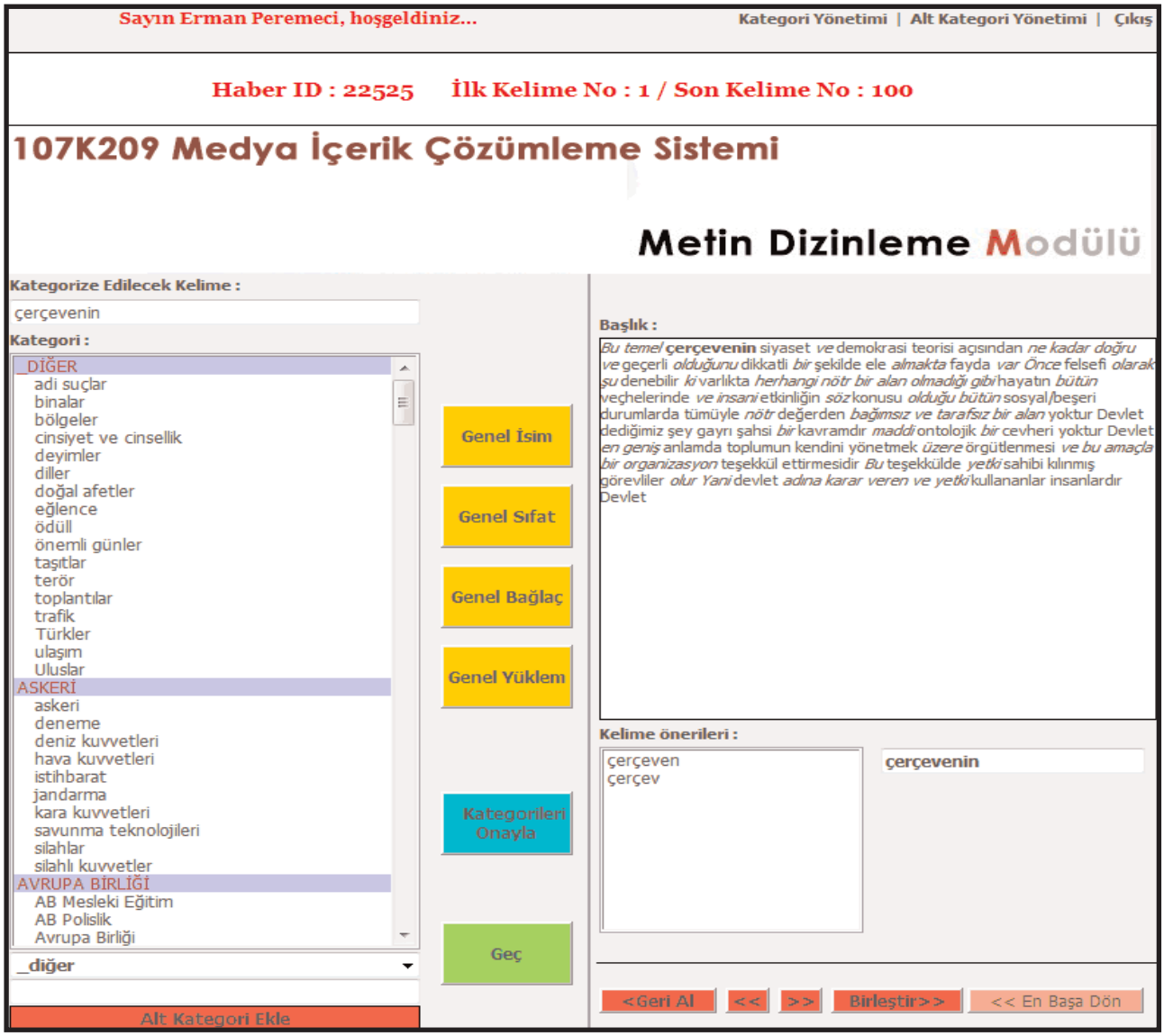

Fig.4. Screenshot of the Text Indexing Module

Sayın Erman Peremeci, hoşgeldiniz...

Frekans Çözümlemeleri | İstatistiksel Çözümlemeler | Haftalık İş Takip Raporları | Çıkış

107K209 Medya İçerik Çözümleme Sistemi

Raporlama Modülü

İncelenen Yayın : Tümü Kelimeler : AB, Türkiye		
Ocak / 2003	425	1097
Şubat / 2003	0	1
Mart / 2003	0	2
Nisan / 2003	56	249
Mayıs / 2003	1	9
Haziran / 2003	10	16
Temmuz / 2003	0	0
Ağustos / 2003	0	2
Eylül / 2003	1	1
Ekim / 2003	0	6
Kasım / 2003	0	0
Aralık / 2003	2	11
TOPLAM	495	1394
İncelenen kelimeler arasında bulunan korelasyon değeri : **0,995462222861292**		

Fig.5. Screenshot of the Statistical Analysis and Reporting Module – Reporting of the PMCC

Sayın Erman Peremeci, hoşgeldiniz...

Frekans Çözümlemeleri | İstatistiksel Çözümlemeler | Haftalık İş Takip Raporları | Çıkış

107K209 Medya İçerik Çözümleme Sistemi

Raporlama Modülü

İncelenen Yayın : Tümü Kelimeler : ABD, Türkiye		
Ocak / 2003	447	1097
Şubat / 2003	0	1
Mart / 2003	1	2
Nisan / 2003	127	249
Mayıs / 2003	1	9
Haziran / 2003	5	16
Temmuz / 2003	0	0
Ağustos / 2003	3	2
Eylül / 2003	0	1
Ekim / 2003	0	6
Kasım / 2003	0	0
Aralık / 2003	4	11
TOPLAM	588	1394
İncelenen kelimeler arasında bulunan korelasyon değeri : **0,998327916106564**		

Fig.6. Screenshot of the Statistical Analysis and Reporting Module – Reporting of the PMCC

marked in standard units, or we need objects of known sizes for comparison. This is the basic reason that lies behind the usage of content analysis as a research technique for analyzing media.

Computerized and linguistic approaches to content analysis are commonly used approaches, as well in fields such as NLP (Natural Language Processing) and CATA (Computer Assisted Content Analysis).

The main scientific contribution of the 107K209 project to the field is the information retrieval approach it proposes with associative indexing. The most distinguishing point of the CATA system of 107K209 is its compatibility to Turkish. The purpose of the project is to develop a Turkish based computer assisted content analysis system to analyze the content of the Turkish press of the Republican Era.

The Turkish based computer assisted content analysis system provides a basic approach to meet the requirement of a computer assisted system for the analysis of Turkish content which has ever been an infrastructural problem for the analysis of Turkish newspapers using modern Turkish script after 1928.Another applied contribution of the project is its integration of statistical analysis to CATA. Findings and figures about the CATA system realized with the 107K209 project that are delivered in this paper may be regarded as a strong proof of these contributions.

5. Acknowledgements

The researchers would kindly express their deep acknowledgements to Prof. Dr. Nilufer EGRICAN, Yeditepe University's Vice-Rector Responsible for Institutional Advancement, for her support and efforts in developing all proposals, to TUBITAK-SOBAG for their funding of the Project 107K209, to the staff of the Istanbul University Library and Documentation Center for their support and hospitality, to Mr. Umit OZAYDIN form Dragomanos Language Technologies for his kind and supportive cooperation, to Instructor Mr. Devrim BARAN for his contribution to the development of the photographic acquisition system, Mr. Baris DALAN, CEO of ISTEK Service Company for his generous support to the project and finally Mr. Erman PEREMECI, for his excellent work in software development.

References

Anderson, T. & Song, A. (2008). Next Generation Projective Techniques: Combining Psychological Content Analysis and Text Mining in Market Research. Stamford: Anderson Analytics.

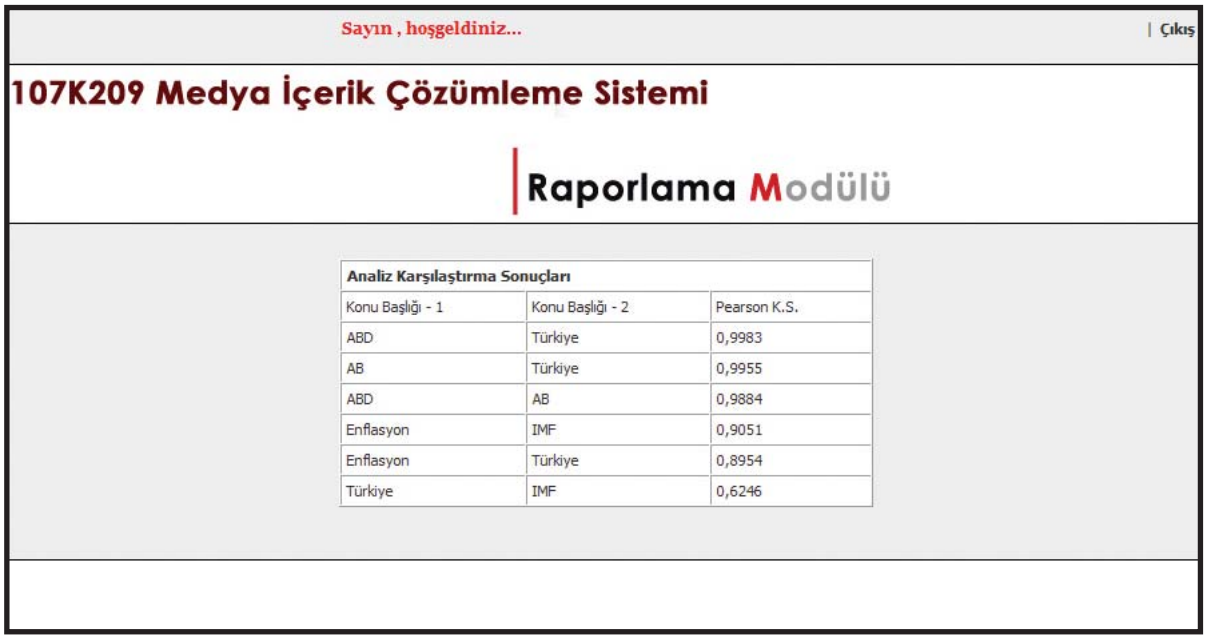

Fig.7. Screenshot of the Statistical Analysis and Reporting Module – Comparison of PMCC

Arikan, A. (2006). Bilgi Erisim Sistemleri: Bilgi Erisimde Dil Sorunlari. Istanbul, Babil.

Berelson, B. (1952). Content Analysis in Communication Research. New York: Free Press.

Berger, A. A. (2005). Media Analysis Tecniques. Thousand Oaks: Sage.

Bilgin, N. (2006). Icerik Analizi: Teknikler ve Ornek Calismalar. Ankara: Siyasal.

Bush, Vannevar. (1837). “As We May Think”, The Atlantic.com (rerrieved on 06.05.2008).

Cebi, M. S. (transl. & ed.) (2003). Iletisim Arastırmalarında Icerik Cozumlemesi. Ankara: Alternatif.

Gokce, O. (2006). Icerik Analizi: Kuramsal ve Pratik Bilgiler. Ankara: Siyasal.

Krippendorff, K. (2004). Content Analysis: an Introduction to its Methodology. (2nd ed.). Thousand Oaks: Sage.

Lasswell, H.D. (1927). Propaganda Technique in the World War. New York: Knopf.

Manning, C. D. & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. Cambridge: The MIT Press.

Mogotsi, I. C. (2007). “News Analysis Through Text Mining: a Case Study”, VINE: The Journal of Information and Knowledge Management Systems, 37, 516-531.

Neuendorf, K. A. (2002). The Content Analysis Guidebook. Thousand Oaks: Sage.

Sherr, S & Staples, M. (2004). News For a New Generation: Report 1: Contetn Analysis, Interviews, and Focus Groups. New York: CIRCLE.

Tavsancil, E. & Aslan A. (2001). Icerik Analizi ve Uygulama Ornekleri. Istanbul: Epsilon.

Weber, R. P. (1990). Basic Content Analysis. (2nd ed.). Newbury Park: Sage.

Züll, C. & Landmann, J. (2002). Computergestützte Inhaltsanalyse: Literaturbericht zu neueren Anwendungen. Mannheim: ZUMA.