Bilgisayar Destekli Türkçe Tabanlı Medya İçerik Çözümleme Sistemi – 107K209 Projesi: Bir Olgu İncelemesi*

Computer Assisted Turkish Based Media Content Analysis System - Project 107K209: A Case Study

*Aykut ARIKAN***

Öz

İçerik çözümleme, çözümlenen medya içeriği nesnesine, yani iletişim metnine yönelik, iyi tanımlanmış sistematik bir bilgi sorgulama stratejisidir. Araştırma sürecinin ilk aşamalarından itibaren neyin aranacağının iyi tanımlanmış olması gereklidir. İçerik çözümlemenin kuramsal açıklamaları kimi zaman uygulamadan farklılık gösterebilir. İçerik çözümlemeye dayalı araştırmalar ilgili çalışmanın uygun bir temelde gerçekleştirilmesi için lazım olan aşamaları, teknikleri ve süreçleri gerekli kılarken diğer yandan zaman ve işgücünden tasarrufu da gerektirir. Yeditepe Üniversitesi bünyesinde görev yapan bir grup araştırmacı, Türkçe medya içeriği için bilgisayar destekli bir içerik çözümleme sistemi geliştirme kararına varmışlardır. Geliştirilen proje önerisi, Mayıs 2007'de TÜBİTAK'a 1001 kodlu bilimsel araştırma projesi olarak sunulmuş ve 107K209 kodu ile proje olarak destekleme kararı verilmiştir. Projenin amacı, Cumhuriyet Dönemi Türk Basınının içeriğini çözümleyecek, Türkçe tabanlı bir içerik çözümleme sistemini geliştirmektir. Bu projeyle geliştirilen Türkçe tabanlı bilgisayar destekli içerik çözümleme sistemiyle, Cumhuriyet Döneminde 1928'den bu yana yeni Türkçe harflerle yayımlanmış olan Türk gazetelerinin içeriklerinin çözümlenmesi için aşılması gereken en önemli altyapı engeli olan Türkçe içeriği çözümleyebilecek bilgisayar destekli sistem ihtiyacı karşılanmıştır. Bu sonuç raporunun amacı, proje hakkında birinci bilimsel bilgi sağlamak ve projenin metodolojik, bilimsel ve uygulama kaynaklı bulgularını tartışarak ortaya koymaktır. Raporun yapısı bu nedenle daha ziyade betimleyicidir ve tarihsel betimleme, sistem modelleme vb. bilimsel yöntem ve tekniklerden yararlanır. İçerik çözümleme, dilbilimsel ve bilgi erişim yöntemleri tartışmanın odağında yer almaktadırlar.

***Anahtar sözcükler**: İçerik analizi, Medya, Bilgi teknolojisi, Bilgi erişim*

Abstract

Content analysis is a well-described systematic strategy of inquiry to the analyzed media content object, or in other words, to a communication text. From the very beginning of the research it must be described what is looked after. Theoretic explanations of content analyses

* Makale TÜBİTAK tarafından desteklenen 107K209 kodlu 1001 grubu bir araştırmanın sonuç raporundan alınmıştır.

** Yrd. Doç. Dr.; Yeditepe Üniversitesi, Bilgi ve İnovasyon Yönetimi Programı Yürütücüsü, 26 Agustos Yerleşimi Kayışdağı 34755 İstanbul, (aarikan@yeditepe.edu.tr)

sometimes differ from practical usage. Research based on content analysis, requires steps, techniques and processes to fix the research on an adequate foundation and to save time and work. A group of researches 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 under the project number 107K209. 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 purpose of this report is to deliver primary information on the project and to discus methodological, scientific and applied problems and issues related to the project. The nature of the report 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.

***Keywords**: Content analysis, Media, Information technology, Information retrieval*

Giriş

Yeditepe Üniversitesi bünyesinde görev yapan bir grup araştırmacı, Türkçe medya içeriği için bilgisayar destekli bir içerik çözümleme sistemi geliştirme kararına varmışlardır. Geliştirilen proje önerisi, Mayıs 2007'de TÜBİTAK'a 1001 kodlu bilimsel araştırma projesi olarak sunulmuş ve TÜBİTAK tarafından 107K209 kodlu olarak destekleme kararı verilmiştir.

Projenin amacı, Cumhuriyet Dönemi Türk Basınının içeriğini çözümleyecek, Türkçe tabanlı bir içerik çözümleme sistemini geliştirmektir. Bu projeyle geliştirilen Türkçe tabanlı bilgisayar destekli içerik çözümleme sistemiyle, Cumhuriyet Döneminde 1928'den bu yana yeni Türkçe harflerle yayımlanmış olan Türk gazetelerinin içeriklerinin çözümlenmesi için aşılması gereken en önemli altyapı engeli olan Türkçe içeriği çözümleyebilecek bilgisayar destekli sistem ihtiyacı karşılanmıştır.

Bu makalenin amacı, 107K209 Projesi özelinde, içerik çözümlemesi konusunu ve sorunlarını ele alarak, konuya dilbilimsel, bilgi erişim ve istatistiksel yöntemler açısından getirilebilecek yeni bir bakış açısını tartışmaktır.

Makalenin 'araştırma sorusu' şöyledir: "Türkçe tabanlı iletişim metinlerinin, bilgisayar destekli çözümlemesinde, dilbilimsel, bilgi erişim ve istatistiksel yöntemlerin etkileri nelerdir? Makale, tarihsel/deskriptif yöntemi kullanmaktadır. Bilgi toplama tekniği olarak katılımlı gözlem tekniği kullanılmıştır. Bu çerçevede, TÜBİTAK destekli 1001 kodlu bir bilimsel araştırma projesi olan 107K209 Projesi, bir olgu çözümlemesi olarak ele alınmıştır.

Genel Bilgiler, Gereçler ve Yöntem

Bu proje iki temel bilimsel yöntemden faydalanmaktadır: İçerik çözümleme ve bilgi erişim yöntemleri. Bu iki yöntem bu projede bir araya getirilerek kullanılmaktadır. Bunun ötesinde, bu proje etkin ve etkili sonuçlara ulaşmak amacıyla kalite yönetimi yöntemi de benimsenmiştir. Bütün bu yöntemsel yaklaşımlar sistemci yaklaşımı içinde bir sistem tasarımı ile bir araya getirilmişlerdir.

Kalite Yönetimi

Bu projede, etkin ve etkili sonuçlara ulaşmasını sağlamak için gerekli olan tüm etkinliklerin güvence altına alınması amacıyla, kalite yönetimi yöntemi kullanılmıştır. Bu amaçla, proje sürecinin her aşamasında, PDCA (Plan-Do-Check-Act=Planla-Uygula-Denetle-Önlem Al) olarak da bilinen Schewart Döngüsü uygulanmıştır. Bu kapsamda proje sürecinin her aşaması, proje önerisi doğrultusunda uygulanmış, denetlenmiş ve denetim sonucu elde edilen bulgu ve ara-sonuçlara göre projede gerekli uyarlamalar yapılmıştır. Ayrıca projenin tüm süreçleri ayrıntılı olarak dokümante edilmiştir.

Sistem Tasarımı

Görüntü Tarama ve İşleme Sistemi : Görüntüsü taranacak nesneler ciltler biçiminde düzenlenmiş gazete sayfalarıdır. Gazete sayfaları bu iş için özel olarak geliştirilmiş, sayfa görüntüsünü belirli bir açıyla, gazete cildine zarar vermeden tarayabilen, yüksek çözünürlüklü ve uygun makro-objektifli dijital bir fotograf makinesiyle yapılmıştır. Bu aksamdan elde edilen dijital görüntüler, geçici bir süre, yedekleme amacıyla bilgisayar sisteminde tutulmuştur. Görüntülerin işlenmesinde, optik karakter tanıma (Optical Character Recognition-OCR) uygulaması kullanılmış, sisteme aktarılan her yeni görüntü otomatik olarak OCR uygulamasına aktarılıp, dijital metne çevrilerek tarama ve görüntü işlemede oluşacak zaman kayıpları en aza indirgenmiştir. OCR uygulaması Türkçe uyumlu olacak şekilde seçilmiştir. OCR uygulamasına aktarılamayan fotoğraf, resim, çizim, karikatür, vb. görsel malzeme, veri tabanına görüntü formatında alınmıştır. OCR işleminden geçen her metin, medya tarama sorumlusu tarafından özgün metinle karşılaştırılarak çevrim sırasında oluşabilecek olası hatalar giderilmiştir. Özel bir uygulamayla gazetelerin 1996 yılından bu yana internet üzerinde yayımlanan baskıları da taranmış, bu taramalardan elde edilen dijital metinler veri tabanına aktarılmıştır. Hatadan ayıklanmış ham metinler, medya tarama sorumlusu tarafından gazetenin günlük sayısı bağlamında düzenlenmiştir. Bu düzenleme sırasında, haber başlığı, spot, resim altı, haber metni vb. haber parçaları kendi içlerinde eşleştirilmiş, köşe yazıları, makaleler, fıkralar, rubrikler, yazı dizileri vb. metin içerikleri de, kendi aralarında gruplandırılarak veri tabanına kaydedilmiştir. Veri tabanına aktarılan her bir görsel ve her bir metin, tarih, sayfa sayısı vb. yönlerinden adreslenmiştir.

Bilgi Erişim Sistemi: Bilgi erişim sisteminin oluşturulmasında, veri tabanında bulunan metinler temel alınmıştır. Bu metinlerde başlıktan KWIC (Key-Word-In-Context = Bağlam Kökenli Anahtar Kelime) ve metin gövdesinden KWOC (Key-Word-

Out of Context = Bağlam Dışı Anahtar Kelime) yöntemleriyle anahtar kelime dizinleri, bilişim sistemi tarafından otomatik olarak üretilmiştir. Veri yönetim uzmanı tarafından bu otomatik dizinler, gene veri yönetim uzmanının oluşturacağı konu başlıkları, kişi adları, kurum adları temelli, denetimli dağarcık dizinleriyle birlikte normalize edilmiştir. Normalizasyon işlemi sırasında dizinlerde oluşan eşsesliler ve eşanlamlılar çapraz göndermelerle ("bkz." ve "ayr. bkz." göndermeleri) ayıklanarak, tekrarlar, eksik yazımlar, yanlış yazımlar vb. sorunlar da, uygulamanın dizin yönetimi modülüyle düzeltilmiştir. Bilgi Erişim Sistemi, çok kullanıcılı ve VPN (Virtual Private Network = Sanal Özel Ağ) üzerinden erişim seçeneklidir. Bilgi Erişim Sistemi aylık olarak güncellenmiş, söz konusu güncellemeler, iki ayrı mekânda arşivlenerek yedeklenmiştir.

İçerik Çözümleme Sistemi: İçerik çözümleme sistemi veri tabanında dizilenmiş halde bulunan haberler üzeriden istatistiksel analiz teknikleri ile gerçekleştirilir. İçerik çözümleme sistemi de, bilgi erişim sistemi gibi çok kullanıcılı ve VPN üzerinden erişim seçeneklidir. Sistem aylık olarak güncellenir; söz konusu güncellemeler, iki ayrı mekânda arşivlenerek yedeklenir.

Bilişim Sistemi: Bilişim sisteminin merkezinde SQL veri tabanı bulunmaktadır. Bu veri tabanı, OCR uygulamasından veri aktarımına açıktır. Yazılım uygulaması, SQL veri tabanı ile bütünleşmiş olarak çalışmaktadır. SQL veri tabanı günlük (tek ve çift günler için iki ayrı set olmak üzere) haftalık ve aylık rutinler halinde tamamı ayrı setlerde olmak üzere yedeklenerek, aylık setler çift kopya olarak iki farklı mekânda saklanmıştır. Uygulama istemci/sunucu mimarisine dayalı ve VPN seçeneklidir. Sistemin güvenliği Yeditepe Üniversitesi'nin güvenlik duvarı tarafından sağlanmaktadır. Sistemin bakımı, güncellemesi vb. yönetsel işlemleri, sistem yöneticisi tarafından sürekli ve düzenli olarak yapılmıştır.

İçerik Çözümle Yöntemi

İçerik çözümlemesi, çerçevesi açıkça belirlenen ve çözümleme nesnesine istisnasız uygulanan sistematik arama stratejisidir. Araştırmanın başında neyin aranacağı açıkça belirlenmek zorundadır. İçerik çözümlemesinin kuramsal olarak açıklanması ile pratikte uygulaması birbirinden farklılık gösterir. İçerik çözümlemesi uygulanarak yapılan araştırmalarda, araştırmanın sağlam bir zemine oturtulması, zaman ve emek tasarrufu için takip edilmesi gereken aşama, teknik ve süreçler vardır.

Bilgi Erişim Yöntemi

"Bir bilim disiplini olarak yaklaşık 50 yıllık bir geçmişi olan bilgi erişim (information retrieval) ... "bilgi toplama, sınıflama, kataloglama, depolama, büyük miktardaki verilerden arama yapma ve bu verilerden istenen bilgiyi üretme (veya gösterme) teknik ve süreci" olarak tanımlanmaktadır" (Tonta, 2001).

Konunun, tarihsel perspektif içindeki gelişimini irdelemek, tanım denemeleri bakımından da yararlı olacaktır. "Kavramın ortaya çıktığı tarihsel kesit olan II. Dünya

Savaşı sonrasında, özellikle bilgisayar sistemlerinin kullanılmaya başlanması, bu yaklaşımın tarihsel bağlam içindeki temellerini de açıklamaktadır. Zira bilgisayar öncesinde de kütüphane sistemleri, kataloglar, bibliyografyalar gibi, bibliyografik sistemlerin içinde gerçekleştirilen bilgiye erişim işlemlerinin, yeni bir kavram doğuracak biçimde kurumsallaşmasının temel nedeni bilgisayar sistemleridir (Arıkan, 2006, ss.25-26).

Bu noktada, bilgisayar teknolojisinin konu kapsamına etkisini de vurgulamak gerekir. "Bilgisayara dayalı bir sistem mantığı içinde tanımlanan bilgi erişim kavramı, kendisini biçim açısından bilgi erişim adı verilen sürecin içinde tanımlar ve açıklar. Bu sürecin kavramları, araçları ve yöntemleri de, bir sorun alanı olarak, bilgi erişimin bilimsel bir disiplinin olarak çerçevesini çizer" (Arıkan, 2006, s.26).

Bilgi erişimin bilimsel sorun alanının çizilmesi konusu da özellikle irdelenmelidir. Bilgi toplamak, işlemek ve bir yerden bir yere iletmek için kullanılan bilgi ve iletişim teknolojilerinin gelişmiş olması, bilgi kaynaklarının çok küçük bilgisayar yongaları (chips) üzerine depolanması, hatta söz konusu yongaların insan beyninin bir uzantısı haline getirilmesi bilgi erişim sorununun çözümü için yeterli değildir. Çünkü kayıtlı bilgilere bir şekilde erişim sağlamak gerekmektedir.

Diğer yandan Tonta'nın (2001) da vurguladığı gibi, "bilgi erişim ise çoğu zaman, bilgi ihtiyacımızı tanımladığımız terimler ile bu ihtiyacımızı karşılaması muhtemel belgelerde geçen terimlerin eşleştirilmesine dayanmaktadır. Bu eşleştirme süreci ise henüz çok iyi anlaşılamadığından mükemmel bir biçimde işlememektedir. Bu bakımdan bilgi erişim sorununu çözmek üzere geliştirdiğimiz bilgi teknolojileri ve entelektüel erişimi kolaylaştıran dizinleme ve sınıflama sistemleri henüz bilgi erişim sorununa çözüm bulmaktan uzaktır."

Temel bilgi erişim yöntemlerinden olan "konu başlıkları, bibliyografik kayıtlara ilişkin bir yöntemdir. Burada amaç, belgelerin içerdikleri konuları denetimli bir sözcük dağarcığı içindeki terimlerle karşılayarak, bunları alfabetik erişim düzeni içinde erişilebilir kılmaktır. Konu başlıklarında, konular arası kavramsal ilişkiler çapraz göndermelerle sağlanır. Ancak bu düzen, konular arası kavramsal ilişkileri, sistematik düzendeki kadar derinlemesine yansıtamadığı için eleştirilmektedir. Konu başlığı olarak seçilecek denetimli sözcük dağarcığı terimlerinin oluşturulmasında, terimlerin okura yönelik olması; her konuya tek bir başlık vermesi (eşanlamlılar arasından bir terimi tercih etmek), kullanılacak terimin konunun kavramsal karşılığını tam ve özgül bir şekilde karşılaması (geniş veya dar anlam taşımaması), tercih edilecek terimin yaygın kullanımlı olması ve yabancı terimlerden sakınılması gibi temel ilkelere dikkat edilmelidir" (Arıkan, 2006, s.53).

Dahası konu başlıklarının örgütlenmesi de başlı başına bir sorundur. "Konu başlıkları, bu yöntemden yararlanan her bilgi erişim sistemi için özel olarak sistemin kendi gereksinimleri doğrultusunda bir yetke dizimi içinde örgütlenmelidirler. Bunlara örnek olması amacıyla, günümüzde iki standart liste yoğun biçimde kullanılmaktadır: Küçük ve orta ölçekteki kütüphaneler için 'Sears' konu başlıkları listesi ve daha

büyükleri için de 'Kongre Kütüphanesi Konu Başlıkları' listesi (Library of Congress Subject Headings - LCSH)'. Dikkat edilmesi gereken nokta, bu standart listelerin bir konu başlıkları yetke dizini gibi kullanılamayacağı, sadece konu başlıkları yetke dizinlerinin türetilebileceğidir (Arıkan, 2006, s.53).

Bu noktada, konu başlıkları açısından dil sorunu ortaya çıkmaktadır. Başka bir deyişle, "iki listenin İngilizce olduğu da gözden kaçırılmamalıdır. Zira yabancı dildeki konu başlıklarını kullanmak hem kütüphaneci, hem de kullanıcı için oldukça zor ve zahmetli bir iştir. Bu başlıkları çevirerek kullanmak ise terimler arası ilişkiler ve kapsamlar, diller arasında farklılıklar gösterebileceğinden neredeyse imkânsızdır" (Arıkan, 2006, s.53).

Diğer bir bilgi erişim yöntemi olan "anahtar sözcük yöntemi, bilgi erişim alanında kişisel bilgisayarların yoğun olarak kullanılmaya başlaması ile birlikte önemini yitirmiştir. Bunun nedeni, kişisel bilgisayar öncesi dönemde, belgelerin yayın adlarında bulunan sözcüklerin bilgisayar yoluyla dizinlenerek bilgisayar çıktısı haline getirilmesi, bunların da yayımlanarak son kullanıcının hizmetine verilmesidir (Arıkan, 2006, s.54). Günümüzde çoğu kütüphanede bunların görülmemesinin temel nedeni budur.

Kişisel bilgisayarların kütüphanelere girişi birçok şeyi etkilediği gibi anahtar sözcük yaklaşımını da derinden etkilemiştir. Yani "kişisel bilgisayarlar ile birlikte, bu işlem artık her tür sorgu için otomatik olarak yapılabildiğinden, herkes tarafından kullanılmıştır. Anahtar sözcük yöntemi önceleri KWIC (Key Word In Context) adı verilen anahtar sözcüklerin yapıt adının içinden seçildiği bir biçimde, daha sonra da KWOC (Key Word Out of Context) adı verilen, anahtar sözcüklerin denetimli bir sözcük dağarcığı içinden seçildiği bir biçimde kullanılmıştır" (Arıkan, 2006, s.54).

Bulgular ve Tartışma

Projenin bulguları konunun tarihsel bağlamının betimlenmesi tanımlarla başlayarak aktarılacaktır. Alanda yaşanan güncel durumun kısa bir özeti, içerik çözümlemesine bilgisayar destekli, dilbilimsel ve bilgi erişimsel yaklaşımlara özel bir önem verilerek uygulamaların ve yaklaşımların tartışılmasıyla yansıtılmaya çalışılacaktır. Son olarak 107K209 Projesi'nin bulguları ayrıntılı olarak ele alınacaktır.

Tarihsel Bağlam ve Tanımlar

Tarihsel bağlamda, içeriğin çözümlenmesi düşüncesi özellikle 20. yüzyılda ortaya konulmuş tanım önerilerinde kendini gösterir. Krippendorff sistematik içerik çözümlemeye yönelik erken yaklaşımları, 20. yüzyılın çağdaş anlamdaki içerik çözümlemesinin üç yüzyıl öncesine, "Kilise'nin 17. yüzyıldaki engizisyon soruşturmalarına" kadar götürür (Krippendorff, 2004, s.3).

Max Weber'in Alman Sosyoloji Derneği'nin 1910 yılındaki toplantısında, "basının içeriğinin geniş ölçekli bir çözümlemesini" önermesine karşın (Krippendorff, 2004, s.4), içerik çözümlemesinin ilk niceliksel kullanımı, bundan çok önce, daha 1893'te Gilmer

tarafından yayımlanan "Do newspapers now give the news?" adlı yazısında yer alır (Krippendorff, 2004, s.49).

Bu öncü girişimlere karşın kavramın İngilizcede ilk kullanımı 1941 yılında gerçekleşir (Waples ve Berelson, 1941, s.2; aktaran Krippendorff, 2004, s.3). Aynı dönemin diğer bir ilginç uygulaması da 1940'larda bir Alman radyo istasyonunun paralelinde yayın yapan bir Amerikan dergisinin Nazi propagandası yapmakla suçlanmasında görülür (Tavşancıl ve Aslan, 2001, s.27).

Güncel Durum: Yaklaşımlar ve Uygulama

İçerik çözümleme, günümüzde iletişim fakültelerinde, lisans ve lisansüstü düzeyde, doğrudan iletişim metinlerine yönelik, analitik ve nesnel yapısı ve özellikle elci uygulamalarda kolaylık sağlamaktadır. Bu nedenle çok incelenen ve öğretilen araştırma yöntemlerinden biridir. Ancak buna karşın, bu örnekler özellikle aylar düzeyinde kısıtlı bir dönemin ele alınması veya tek yayın ya da kısıtlı konuların incelenmesi türünden ciddi kısıtlara sahiptirler.

Öte yandan içerik çözümleme, birçok değişik iletişim ortamı için kullanılabilir. Örneğin Ogilvie, Stone ve Shneiderman 1966'da tarafından ortaya konulduğu gibi el yazması notlar için dahi kullanılabilir (Weber, 1990, s.20). Araştırmacılar içerik çözümleme yöntemiyle elle yazılmış olan değişik intihar notlarını çözümleyerek sonuçlar elde etmeye çalışmışlardır.

Bu alandaki önemli ve ses getiren araştırma örneklerinden biri, ABD'ndeki CIRCLE (The Center for Information & Research on Civic Learning & Engagement) adlı kuruluşun "News for a New Generation Report 1: Content Analysis, Interviews, and Focus Groups" başlıklı raporudur (Sherr ve Staples, 2004). Avrupa'dan başka bir örnek ise "Textpack" (http://www.gesis.org/en/zuma/index.htm, erişim 09.05.2008) adı altında bilgisayar destekli içerik çözümleme sistemi geliştiren "The Centre for Survey Research and Methodology (ZUMA)" (Alman Sosyal Bilimler Altyapısı Servisi'nin bir kolu) kuruluşunun katkılarıdır. Afrika'dan bir katkı da Güney Afrika'daki Univesity of Stellenbosch tarafından yürütülen ve University of Botswana (Mogotsi, 2007) tarafından desteklenen bir araştırmayla kaydedilmiştir. Günümüzde bu yöntem örneğin pazar araştırmaları gibi geniş bir yelpazeye yayılı araştırma alanlarında kullanılmaktadır (Anderson ve Song, 2008). Türkiye'deki uygulama örnekleri arasında, Doğan Atılgan'ın "Kataloglamada Standardizasyon Açısından Türkiye Bibliyografyası'nın İçerik Analizi" başlıklı doktora teziyle (1992), Yontar ve Yalvaç'ın "Problems of Library and Information Science Research in Turkey: A Content Analysis of Journal Articles 1952-1994" başlıklı makalesi (2000) sayılabilir.

Konunun özel bir uygulama alanı da elbette CATA (Computer Assisted Text/Content Analysis = Bilgisayar Destekli Metin/İçerik Çözümleme)'dir. CATA kapsamında geliştirilen özel yazılımlar kendilerine hem bilimsel, hem de ticari çevrelerde uygulama alanı bulmuştur. Örneğin, Kimberly Neuendorf (2008) the Content Analysis Guidebook Online adlı web sitesinde kapsamlı bir CATA yazılımları listesi sunmaktadır. ZUMA

bünyesinde faaliyet gösteren Züll ve Landmann içerik çözümleme konusundaki yayınların kapsamlı ve açıklamalı bir listesini yayımlamışlardır (Züll ve Landmann, 2002, ss. 23-98).

Bilgisayar Destekli Yaklaşım: CATA (Computer Assisted Content/Text Analysis = Bilgisayar Destekli İçerik/Metin Çözümleme)

Klaus Krippendorff CATA konusunda iki temel düşünceyi ortaya koyar (Krippendorff, 2004, ss.258-259):

1. Bilgisayarların geniş ölçekli veriyi yüksek hızda işleme yeteneği;
2. Bilgisayarların metinsel içeriği işlemedeki güvenilirlikleri.

İlk gerekçe, araştırmacılara içerik çözümlemenin kapsamında, zaman, konu ve değişik iletişim ortamlarının niteliksel ve niceliksel yönlerden incelenmesi konusunda geniş olanaklar sağlar. CATA ile gazete dermelerinin on yılları hatta yüzyılları kapsayan çözümlemelerinin yapılabilmesine olanak sağlar. Diğer bir yandan, niteliksel olarak farklı olan değişik kitaplar veya niceliksel olarak farklı olan değişik konulardaki çeşitli dergiler dahi artık çözümlenebilir hale gelmektedirler. Artık elci sistemle kodlama hiç bir şekilde CATA ile karşılaştırılabilecek konumda değildir. Elbette CATA sistemleri için ek bir makine öğrenme süreci gereklidir. Ancak bu süreç, CATA sistemleri tarafından sağlanan güvenirlikle kıyaslandığında önemsiz kalır.

CATA sistemlerinin kayda değer bir özelliği de içerik çözümlemesini istatistiksel çözümleme olanaklarıyla birleştirebilmesidir. Zira içeriğin çözümlenmesiyle sağlanan veri, istatistiksel olarak değişik korelasyonların yapılması ve derinlemesine yorumlar sağlanması amacıyla da incelenebilir. Bu becerilerin kombinasyonu elbette güçlü ve uzmanlaşmış veri-sözlüklerinin dizinler ve atlanacak kelime (stop-word) listeleri olarak entegrasyonunu gerekli kılar ki; bu durum çözümlenecek içeriğin artmasıyla imkânsız olmasa bire giderek daha zor hale gelir. Bu nedenle gelişkin CATA sistemleri hem uygun dizinler ve atlanacak kelime listeleri için güçlü alıştırma verisi (training data) gerektirir, hem de çözümlenecek içeriğin artmasıyla birlikte, dizinlerin ve atlanacak kelime listelerinin sürekli geliştirilmesi esnekliğini de gerektirir.

İçerik Çözümlemeye Dilbilimsel Yaklaşım: NLP (Natural Language Processing = Doğal Dil İşleme) ve Gövde (Corpus) Geliştirme

İçerik çözümlemesinde, çözümlenecek içeriğin (veya iletişim metinlerinin) kayda değer bir bölümü, yani sözgelimi her türden basılı materyal (müzik notaları hariç), el yazmaları, ses ve görüntü kayıtları dile dayalıdır, daha doğrusu doğal dile dayalıdır. Bu nedenle NLP (Natural Language Processing = Doğal Dil İşleme) içerik çözümlemeyle doğrudan ilişkili olan önemli bir alandır.

Çözümlenecek dile dayalı içerik sözlüksel (leksikal) bir kaynaktır: metinler, sözlükler, kavram sözlükleri (thesaurus) ve bunlara dayalı işleme araçları (Manning ve Schütze, 1999, s.19).

Gövde (Latince: Corpora, Corpus'un çoğulu) olarak da adlandırılan metin dermeleri (Manning ve Schütze, 1999, s.6), çözümlenecek nesneyi oluşturur. Bu gövde teknik bir bakışla CATA sistemi içinde yer alan veri tabanından başka bir şey değildir.

Veri sözlükleri veya veri dilleri Krippendorff'a göre "çözümleyicilerin verilerini biçimlendirdikleri betimleyici aygıtlardır" (Krippendorff, 2004, s.150); Weber'e göre ise Bağlam-İçi-Anahtar-Kelime (KWIC) listeleri (Weber, 1990, s.44) veya diğer bir deyişle, kavramsal dizinlerdir. Ancak bunları (yapay dillerin özel bir türü olarak) bilgi erişim dili olarak da adlandırmak mümkündür (Arıkan, 2006, s.79).

Bilgi erişim dillerinin yapılandırılmasında karşılaşılan sorunların temelinde doğal dile dayalı sorunlar yatmaktadır (Arıkan, 2006, s.151). Dolayısıyla içeriği çözümlenecek materyalin taşıdığı doğal dilin yapısı ve özellikleri, veri sözlükleri ve dizinler gibi uygun bilgi erişim dillerinin yapılandırılmasındaki temel sorun kaynağıdır. İngilizceden türetilmiş bilgi erişim dilleri için uygun terimler sağlayan, WordNet vb. türünden, sözlüksel veri tabanları biçiminde bir dizi kavram sözlüğü mevcuttur (Krippendorff, 2004, s.279).

Ural-Altay dil ailesinin bir üyesi olarak Türkçe, İngilizce ve diğer Hint-Avrupa dillerinden tamamen faklı bir morfolojik bir yapıya sahiptir. Dolayısıyla, Türkçe içeriğin çözümlenmesi için geliştirilecek kavram sözlükleri, dizin vb. bilgi erişim dilleri Türkçe tabanlı içeriğin çözümlemesine yönelik olarak ayrıca geliştirilmek zorundadır.

İçerik Çözümlemeye Bilgi Erişim Yaklaşımı

Metin gövdelerinin çözümlenmesine yönelik olan tanımlayıcı (descriptive) yapılar bilgi erişim birimleri olduklarından dolayı, bilgi erişim içerik çözümlemenin merkezi öneme sahip bir parçası halini almıştır. İçerik çözümleme ve istatiksel çözümleme özelliklerinin karmaşık bir CATA sistemi içinde entegrasyonu için, içerik veya metin gövdeleri dizinlenmek zorundadır, ya da diğer bir deyişle, çözümlemeli olarak betimlenmelidir. Bilgisayar tarafından üretilen anahtar kelime listeleri, eş anlamlılar, kısaltmalar vb. bir dizi dil sorundan ötürü oldukça kısıtlı ve esnek olmayan özel dizin türleridir. Bu nedenle, bilgi biliminde "denetimli dağarcık" (contrelled vocabulary) olarak da adlandırılan insan tarafından üretilen dizinler, bilgisayar tarafından üretilenlere göre her zaman daha güvenilir ve esnektirler. Ancak insan müdahalesinin en belirleyici yanı, işlem hızındaki azalmadır. Bilgisayar destekli morfolojik ve semantik çözümleme insan müdahalesini azaltacak ve işlem hızını artıracak çözümlerdir, ancak bu yaklaşım, yapay zekâya doğru güçlü bir yönelim ve uzun çalışmaları gerektirmektedir. Türkçe konusundaki öncü ve temel çalışmalar Kemal Oflazer'e aittir (Oflazer, 2009).

İstatistiksel İlişkilendirme: Pearson Dizi-Zaman Korelasyon Katsayısı

Konu başlıkları arasındaki istatistiksel ilişkilendirmeyi bulmak için zamana bağlı değer dizilerinde yaygın olarak kullanılan Pearson dizi-zaman korelasyon katsayısının (PMCC) hesaplanması sisteme dahil edilmiştir.

Buna göre:

X= anahtar kelime

Y= ilişkisi aranan kelime

n= dönem (ay, yıl)

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

Burada s_x ve s_y standart sapma olup; standart sapma

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

formülüyle hesaplanabilir. İşlem sonunda bulunan değerin -1, 1 aralığında olması gerekir. Bu değerler Tablo 1'deki şekilde yorumlanır:

Tablo1. Korelasyon Değerleri

Korelasyon	Negatif	Pozitif
Zayıf	−0.3, −0.1	0.1, 0.3
Orta	−0.5, −0.3	0.3, 0.5
Güçlü	−1.0, −0.5	0.5, 1.0

Sonuçların negatif olması ters korelasyonu ifade eder, yani bir dizi artış gösterirken diğeri azalma göstermektedir.

107K209 Numaralı Araştırma Projesi

Bu projenin amacı, Cumhuriyet Dönemi Türk Basınının içeriğini çözümleyecek, Türkçe tabanlı bir içerik çözümleme sistemini geliştirmektir. Bu projeyle geliştirilen Türkçe tabanlı bilgisayar destekli içerik çözümleme sistemiyle, Cumhuriyet Döneminde 1928'den bu yana yeni Türkçe harflerle yayımlanmış olan Türk gazetelerinin içeriklerinin çözümlenmesi için aşılması gereken en önemli altyapı engeli olan Türkçe içeriği çözümleyebilecek bilgisayar destekli sistem ihtiyacı karşılanmıştır. Bu amaçla, özellikle kişi adı, kurum adı, konu başlığı, anahtar kelime vb. dizinlerin makine öğrenmeyle geliştirilmesi için, bir kısım içerik taranacak olmakla birlikte Cumhuriyet Dönemi Türk Basınının 1928'den bu yana tamamı bu kapsamın dışındadır. Bunun nedeni, yapılan ön değerlendirmeler sonucunda, söz konusu içeriğin eksiksiz olarak

taranması yaklaşık 6250 adam/gün'lük çalışmaya gerek duyulacağı olgusudur. Bu sistemle yapılacak içerik çözümlemesi ve bu çözümlemenin dayalı olacağı içerik taraması bu projenin kapsamının dışındadır. Literatürde, kimi araştırmacıların içerik çözümle yönteminin bir uzantısı olarak değerlendirdikleri söylem çözümlemesi yöntemi de proje kapsamının dışında tutulmuştur. Bilgisayar destekli içerik çözümleme tekniği, içerik analizi yönteminin diğer tekniklerini aştığından, elle veri toplamaya dayalı olan bu teknikler de kapsam dışı bırakılmıştır.

Projenin genel özgün değeri, içerik çözümleme alanındaki uygulama eksikliğini gidermeye katkı ve uygulama alanının önünü açmasıdır. Bu eksikliğin temel nedeni ise Türkiye'de yapılan tüm içerik çözümleme araştırmalarında, doğal dil ve doğal dile dayalı bilgi erişim sorunlarının ihmal edilmiş olmasıdır.

Projeyle geliştirilmesi amaçlanan, sistem, teknolojik olarak kendi alanında bir öncüdür. Sistemin bileşenlerinden olan, büyük boyutlu (A2) görüntü tarama, Türkçe optik karakter tanımlama (OCR) ve Türkçe bilgi erişim teknolojileri, birbirlerinden bağımsız ve ilişkisiz olarak kullanılmaktadırlar. Projenin teknolojik açıdan özgün değeri bu teknolojileri bir şemsiye altında bir araya getirmesidir. Bu sayede, Türk Basınının içeriğini çözümleyebilecek teknolojik altyapı ortaya çıkacaktır.

Projenin yöntemsel özgün değeri, içerik çözümleme yöntemini bilgi erişim yöntemiyle desteklemesidir. Günümüze dek, özellikle Türkçe içerik için tarif edilmiş özgün bir içerik çözümleme sistemi mevcut değildir. Dahası içerik çözümleme yöntemi konusunda, Türkçe doğal dile dayalı hiç bir ayrıntı ele alınmış değildir. Bu nedenle bilgi erişim konusu, özellikle Türkçe literatürde ihmal edilmiştir (Bilgin, 2006; Gökçe, 2006; Tavşancıl ve Aslan, 2001). Oysa doğal dil, bilgi erişim sorunlarının temelinde yatan olgudur (Arıkan, 2006, s.20). Doğal bir dil olan Türkçeden yapay bir dil olan bilgi erişim diline çevrimde yaşanan dil sorunları, bu alandaki temel sorunlardır. Bu sorunlar aşılmadan, Türkçe içeriğe yönelik içerik çözümleme sorunlarını yöntemsel açıdan aşmak ve sonuca ulaşmak olası değildir. Bu alandaki uygulama eksikliğinin temelinde yatan neden budur.

Projenin kuramsal özgün değeri, başta iletişim bilimleri kuramı olmak üzere, diğer sosyal bilimler kuramlarında da yoğun olarak kullanılan içerik çözümlemesine, bilgi erişim kuramıyla bir yaklaşım getirmesidir. Proje kuramsal açıdan disiplinlerarası nitelik taşımaktadır. Temel alınan iki ana disiplin iletişim ve bilgi erişim disiplinleridir.

Yukarıda belirtildiği gibi, içerik çözümleme konulu literatürde, bilgi erişim konusu, özellikle Türkçe literatürde ihmal edilmiştir. Bu açıdan, projenin literatüre özgün katkısı, özellikle Türkçe doğal dil kaynaklı sorunların aşılmasında, bilgi erişim disiplininin bulgu ve sonuçlarının kullanımıdır. Özellikle Türkçe içerik çözümleme literatüründe eksik kalan Türkçe içeriğe yönelik uygulamalar, bu özgün katkı sayesinde tamamlanabilecektir.

Bu projenin "genel yaygın etkisi", başta iletişim ve tarih araştırmaları olmak üzere, tüm sosyal bilim araştırmalarını ve iletişim endüstrisini doğrudan etkilemesi, bu alanlardaki araştırmalara maliyet ve zaman tasarrufu yönünden katkı sağlaması ve araştırmaların derinliğini artırmasıdır.

İçerik çözümleme, iletişim bilimleri alanının temel yöntemlerinden biridir. Her iletişim fakültesi öğrencisi, eğitimi süresince birden fazla içerik çözümlemesi ödevi hazırlamaktadır. Ortaya konan iletişim konulu yüksek lisans ve doktora tezlerinin de bir bölümü bu yöntemi kullanmaktadır. Bu alanda, Türkçe tabanlı bir içerik çözümleme sisteminin eksikliği hem araştırmaların kapsamını daraltmakta, hem de gereksiz maliyet ve işgücü kayıplarına yol açmaktadır. Projenin iletişim araştırmalarına katkısı, ortaya çıkacak sistemin, bu alandaki araştırmalarda maliyet ve zaman tasarrufu sağlamasıyla, araştırmaların hızlanması, bilgisayar teknolojisinin imkânlarıyla, araştırmaların derinliğinin artacak olmasıdır.

İletişim endüstrisini oluşturan basın, medya, internet, reklam, halkla ilişkiler vb. sektörlerdeki şirketler ve bunların müşterisi olan kişi ve kuruluşlar içerik çözümleme yöntemini ticari amaçlarla sıkça kullanmaktadırlar. Bu alandaki faaliyetlerde, diğer tüm içerik çözümlemesine dayalı araştırmalarda olduğu gibi, tekrarlardan kaynaklanan zaman ve işgücü kayıpları ortaya çıkmaktadır. Bu araştırmalar kapsam kısıtlılığı nedeniyle zaman aralığı yönünden de eksik durumdadır. Projenin iletişim endüstrisine katkısı, ortaya çıkacak sistemin, bu alandaki içerik çözümlemesi faaliyetlerine maliyet ve zaman tasarrufu sağlamasıyla, bu faaliyetlerin hızlanması, bilgisayar teknolojisinin imkânlarıyla, çözümlemenin derinliğinin artacak olmasıdır.

Yazılı kaynaklar, tarih araştırmalarının temel araştırma nesneleridir. Bu kaynaklarda yapılan içerik analizlerinin derinliği ve etkinliği, tarih araştırmalarının niteliğini doğrudan belirler. Bu nedenle diğer tüm sosyal bilimler araştırmalarında olduğu gibi tarih araştırmalarında da, gazetelerde yapılacak içerik çözümleme yöntemi kullanılan başlıca yöntemlerden biridir. Ancak özellikle Türkçe basın için böyle bir sistemin eksikliği, bu çözümlemeleri yapacak öğrenci ve araştırmacıları çözümlemeleri elle yapmaya yönlendirmektedir. Bu da zaman ve iş gücü kayıplarına yol açtığı gibi, çözümlemeler daha kısıtlı bir zaman aralığında gerçekleşmektedir. Dahası, çözümleme elle yapıldığından toplanan veriler sayısallaştırılamamakta, bu çözümlemeler kimi zaman tekrar edilebilmektedir. Projenin diğer tarih araştırmalarına etkisi, ortaya çıkacak sistem sayesinde, bu alandaki araştırmalarda maliyet ve zaman tasarrufu sağlaması ve araştırmaların hızlanması, bilgisayar teknolojisinin imkânlarıyla da araştırmaların derinliğinin artmasıdır.

Bir sosyal nesne olarak gazetelerin incelenmesi, tüm sosyal bilim araştırmalarına uygulanabilecek bir veri toplama yöntemidir (Babbie, 2004, ss.96-97). Bu sayede, hem toplumsal sorunların tarihsel kökeni, hem de bu konulardaki temel açılım ve değişimler belgelenebilir. Proje süreci, 12 ay içinde tamamlanan altı temel aşamadan oluşmuştur: Medya tarama, yazılım geliştirme, istatistiksel çözümleme şablonlarının geliştirilmesi, dizin geliştirme, dizin yönetimi ve deneme çözümlemeleri.

Emek yoğun bir çalışma yürüten projenin ekibi, bir yürütücü (Yrd.Doç.Dr. B. Aykut Arıkan), iki araştırmacı (Yrd.Doç.Dr. M. Deniz Tansi ve Yrd.Doç.Dr. Nilüfer Hatemi), bir bursiyer, bir istatistik uzmanı, bir sistem yöneticisi, bir veri yönetim uzmanı, bir medya tarama uzmanı, bir medya tarama elemanı, bir kalite yönetim uzmanı ve bir bilgi teknolojileri uzmanı olmak üzere toplam 11 kişiden oluşmuştur. CATA sisteminin

geliştirilmesi hizmeti, Peremeci Dijital Çözümler şirketinden Erman Peremeci'den alınmıştır.

Projede İstanbul Üniversitesi Kütüphane ve Dokümantasyon Daire Başkanlığı ve Dragoman Dil Teknolojileri ve Danışmanlık Ltd. Şti. ile bilimsel ve teknolojik işbirliği kurulmuştur.

Proje başlangıcında Yeditepe Üniversitesi Bilgi Merkezi (Merkez Kütüphane) koleksiyonunda yer alan gazeteler, medya tarama açısından yetersiz kalmış, bu nedenle, İstanbul Üniversitesi Kütüphane ve Dokümantasyon Daire Başkanlığı'yla işbirliğine gidilerek, söz konusu taramalar/çekimler burada yapılmaya başlanmıştır. Bu işbirliği çerçevesinde, çekimi yapılan gazetelerin elektronik kayıtları İstanbul Üniversitesi Kütüphane ve Dokümantasyon Daire Başkanlığı'na teslim edilmiştir.

107K209 Projesi daha hayata geçmeden sanayi/sektörle önemli işbirliğinin önünü açmıştır. Bu çerçevede, Dragoman Dil Teknolojileri ve Danışmanlık Ltd. Şti. ile yapılan işbirliği sayesinde, Dragomanos.com elektronik terim sözlüğü metin dizinleme modülü ile bütünleştirilmiştir. Söz konusu sözlüğün entegrasyonu proje önerisinde yer almamasına karşın, bunun sisteme, dizinleme açısından önemli bir kesinlik ve isabet katacağı öngörülmüş ve yapılan denemelerde de bu durum kanıtlanmıştır. Bu işbirliği çerçevesinde, Dragomanos.com'a tüm konu dizinlerinin kullanım yetkisi verilmiştir.

Proje Süreci

Proje TÜBİTAK'a Mayıs 2007'de sunulmuş ve Eylül 2007'de onaylanmıştır. Toplam 12 ayda bitmesi planlanan proje 15 Ekim 2007'de başlamış ve öngörülen takvime uygun olarak 15 Ekim 2008'de başarılı bir şekilde tamamlanmıştır.

Projenin yalnızca yazılım geliştirme sürecinde toplam iki aylık kısa bir sapma söz konusudur. İlerleme raporunda TÜBİTAK'a bildirilen bu sapma, projenin genel gidişini etkilememiş ve belirtilen sürede başarılı bir şekilde tamamlanmıştır. Kalite dokümantasyonunda yer aldığı üzere, yukarıda belirtilen kısa sapmanın üç nedeni vardır:

1. Planlama, bütçeleme vb. nedenlerle sürecin planlanandan bir ay geç başlaması;
2. Metin dizinleme modülüne Dragomanos.com elektronik terim sözlüğünün entegrasyonu nedeniyle 15 günlük gecikme;
3. Rapor Modülündeki sayfa çözümleme fonksiyonalitesine içerik kombinasyonlarının veri tabanına entegrasyonu nedeniyle 15 günlük gecikme.

Toplam iki aylık bu gecikme, projenin geneli açısından bir risk oluşturmadığı ve diğer süreçleri de etkilemediği için, herhangi bir aksaklığa yol açmamakta ve kabul edilebilir sınırlar içinde kalan bir sapma olarak değerlendirilmektedir. Projenin zamanında herhangi bir sorunla karşılaşılmadan tamamlanmış olması da bunun göstergesidir.

Projedeki temel uyum ve değişiklik faaliyeti, proje öneri formunda yer alan B Planı uyarınca, tarayıcıya dayalı yöntemden fotografik yönteme geçiştir. Yöntemsel açıdan nedenleri yukarıda açıklanan bu uyumlandırma ve değişiklik faaliyeti, projenin sağlıklı işlemesini sağlamıştır.

Projenin idari açıdan ilerlemesinde herhangi bir sorun görülmemiş, proje personeli büyük bir uyum ve işbirliği içinde çalışmıştır. Yeditepe Üniversitesi, projeyi her yönden desteklemiş, İstanbul Üniversitesi Rektörlüğü de Kütüphane ve Dokümantasyon Daire Başkanlığı ile işbirliği yapılması konusunda gerekli izinleri süratle vermiştir.

107K209 Projesinin idaresinde kalite yönetimi çalışmasının büyük yararları görülmüştür. Bu sayede, olası aksaklıklar düzenleyici-önleyici faaliyetler ve iç denetimle ortaya çıkmadan çözülebilmiştir.

Medya içeriği sağlamaya ilişkin çalışmalar, tarayıcıdan fotografik yönteme geçişi takiben gerçekleştirilen sistem geliştirme amaçlı deneme çekimleri ve nihayetinde, İstanbul Üniversitesi Kütüphane ve Dokümantasyon Daire Başkanlığı'nda gerçekleştirilen içerik tarama faaliyetleriyle devam edilmiştir.

Veri tabanı tasarımında içeriği çözümlenen sayfalarda yer alan terimlerin, kavram, kişi adı, kurum adı ve yer adı bağlamındaki tüm kombinasyonları da veri tabanına işlenmiştir. Bu sayede, söz konusu ilişkiler daha haber çözümlenirken veri tabanına işlenmiş olacağından, bu ilişkilerin raporlamada tekrar kurulmasına gerek kalmayacaktır. Böylece özellikle karmaşık içerik çözümlemelerinde, çözümleme veri tabanında var olan ilişki kombinasyonları üzerinden gerçekleştirileceği için, bu tür içerik çözümlemelerinin hızı kayda değer şekilde artacaktır.

Yeditepe Üniversitesi'nde bulunan Kyocera KM-3650W tarayıcı, sadece kâğıt tabakalarının taranmasına uygun olduğundan ciltli gazetelerin işlenmesi sırasında gereksiz zaman kayıplarına yol açmaktadır. Bu nedenle tarayıcı yerine fotografik yönteme geçiş kararı verilmiştir. Bu çerçevede, Sony DSC-R1 fotoğraf makinesiyle yapılan deneme çekimleri olumlu sonuçlanmış ve makro objektifli profesyonel bir sistemin seçimi uygun görülmüştür. Yeditepe Üniversitesi tarafından tedarik edilen Canon EOS 5D fotoğraf makinesi ve UV filtreli Canon F 85mm f/1.2L II USM objektifli fotografik sistem başarılı bir şekilde uyarlanmıştır. Yansıtmasız cam, dolaylı ışıklandırma ve rahle düzeneği kurularak esas çekimler başlatılmıştır.

Kısıtlar

Projenin temel kısıtı, sisteme aktarılacak alıştırma verisinin yoğunluğudur. Hürriyet, Milliyet, Sabah, Cumhuriyet ve Türkiye gazetelerinin basılı nüshalarının sisteme fotografik yöntemle aktarımına karar verilmiştir. Bunun yanı sıra, ulusal basına ait internet üzerinde bulunan tüm çevrim-içi nüshaların da Spider (örümcek) Modülüyle, ilk çevrim-içi gazete olan Zaman Online ile başlatılarak sisteme deneme verisi olarak aktarımı kararlaştırılmıştır. Bu ölçekteki bir veri, 1928'den bu yana basılmış olan Türkçe gazetelerle karşılaştırıldığında, göreceli olarak kısıtlı görülebilir, ancak bu miktar bile yoğun bir iş yükü getirmiştir.

Gözlem ve Bulgular

CATA sisteminin tüm modülleri çalışır durumdadır. Proje, temel olarak öngörülen plana uygun şekilde, küçük değişiklikler ve eklemelerle sonuçlanmıştır. Proje önerisinde tarif edilen tüm hedefler ve sonuçlar gerçekleşerek, tüm başarı ölçütlerinin tamamı hayata geçirilmiş olup, sistem eksiksiz olarak çalışmaktadır.

Sistemin ilk modülü olan "Spider" (örümcek) Modülü tam olarak geliştirilmiş ve başarıyla çalıştırılmıştır. Modül halen internet üzerinden veri toplamaya devam etmektedir.

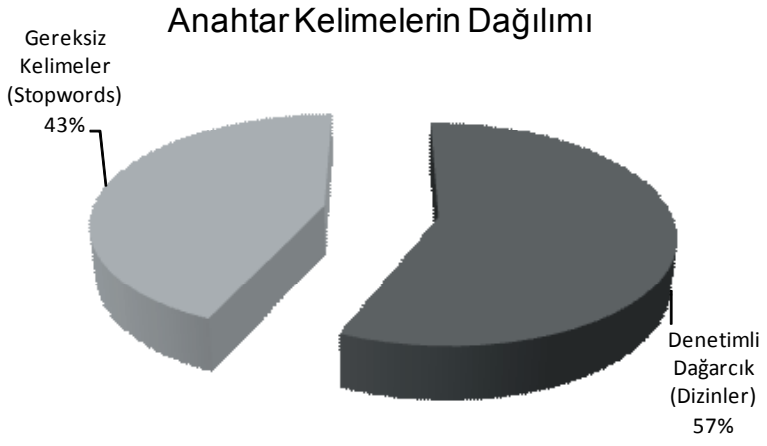
Sayfa çözümleme Modülü Abby Fine Reader tarafından işlenerek HTML haline getirilmiş ham grafik verilerinin, "haberler" halinde veri tabanına kaydedilmesini sağlar. Program bu amaçla kullanıcıya haber öğelerini ilişkilendirebileceği bir grafik arayüz sunar.

Yapılan incelemede Abby Fine Reader programının işlediği sayfalardaki metin ögelerinin yazı biçimlerini algıladığı ve bunları ürettiği HTML sayfalarında "style"lar kullanarak orijinaline yakın bir şekilde işlediği tespit edilmiştir. Öte yandan Fine Reader yazıdaki paragrafları bir bütün olarak değil, birbirinden bağımsız birer sayfa öğesi olarak algılamaktadır. Ayrıca paragrafların sırası her zaman yazının içeriğindeki sırada gelmemekte, Fine Reader paragrafları sayfanın dizimindeki sıraya göre sıralamaktadır. Bu sorunun çözümü için, birbirini takip eden paragrafları bir bütün olarak algılayan, kopuk haber ögelerini de operatörün birleştirmesine olanak sağlayan bir arayüz tasarlanmıştır. Haber öğeleri birleştirildikçe renklenmekte ve başlarına numaraları gelmektedir. Ayrıca operatör devamı başka sayfada olan haberleri işaretleyerek ilgili sayfaya geçtiğinde, bunları da ilişkilendirebilmektedir.

Bölüm 2.5'te ele alınan, Türkçenin morfolojik yapısından kaynaklanan sorunların aşılması amacıyla, çevrimiçi ve ve çevrimdışı ortamlardan elde edilen Türkçe metin gövdeleri, sistemin metin dizinleme modülüyle işlenerek dizinlenmiştir. İlgili haber metinleri, sözkonusu modülle operatörün ekranına gelmektedir. Operatör destekli metin dizinleme işlemi "Metin Dizinleme Modülüne" kategorize edilmiş geniş ölçekli bir denetimli dağarcık veri tabanı, yani bir anahtar kelimeler dizini kazandırmıştır. Sistemden üretilen haberlerden elde edilen yüz bini aşkın kelime operatör destekli olarak kategorize edilmiştir. Bu noktada sistemin yeterli miktarda çalıştırma verisine sahip olduğu değerlendirilerek, Metin Dizinleme Modülü robot modunda çalıştırılmaya başlanmıştır. Modül, robot modunda veri tabanında yer alan haberleri kelime-kelime inceleyerek, oluşturulan sözlükle karşılaştırarak işlemektedir. Metin Dizinleme Modülünün dizinleri tam olarak geliştirilmiş ve kullanılabilir durumdadır. Anahtar kelimelerin toplam sayısı, Ekim 2008 itibarıyla 116.069'dur. Anahtar kelimelerin dağılımı Tablo 2 ve Şekil 1'de yer almaktadır.

Tablo 2. Anahtar Kelimelerin Dağılımı

Tür	Adet	Oran
Denetimli Dağarcık (Dizinler)	65.586	% 43
Gereksiz Kelimeler (Stopwords)	50.483	% 57
TOPLAM	116.069	% 100

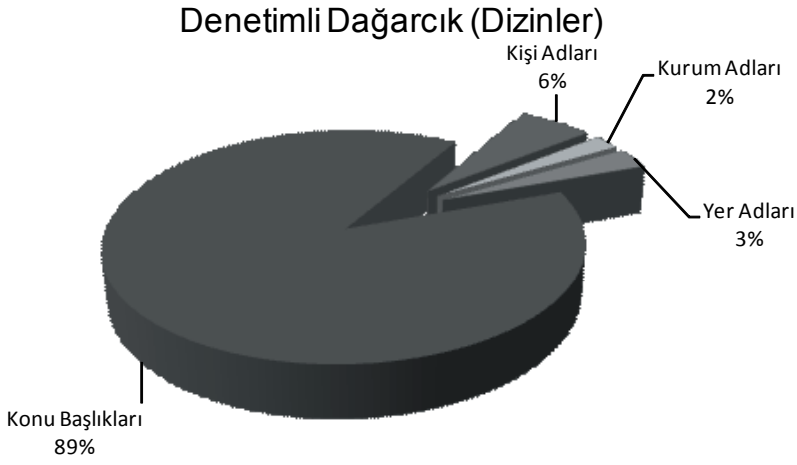


Şekil 1. Anahtar Kelimelerin Dağılımı

Denetimli Dağarcık (dizinler), Kişi Adları, Kurum Adları, Yer Adları ve Konu Başlıkları'ndan oluşur. Konu başlıkları % 89'luk bir payla, dizinlerin en büyük kısmını oluşturur; bunu % 6'lık bir oranla Kişi Adları, % 3'lük bir oranla Yer Adları ve % 2 ile Kurum Adları izler. Ayrıntılar Tablo 3'e ve Şekil 2'ye yer almaktadır.

Tablo 3. Denetimli Dağarcık (Dizinler)

Dizin	**Adet**
Kişi Adları	3.959
Kurum Adları	1.411
Yer Adları	1.970
Konu Başlıkları	58.246
TOPLAM	65.586



Şekil 2. Denetimli Dağarcık (Dizinler)

Metin Dizinleme Modülünün özgün bir özelliği de, bir haberdeki Kişi Adları, Kurum Adları, Yer Adları ve Konu Başlıklarındaki dizin öğeleri arasındaki tüm çapraz ilişkilerin, haber daha işlenirken sayısal olarak veri tabanına aktarılmasıdır. Bu yapı, istatistiksel analizde işlem hızını oldukça artırmaktadır. Diğer bir yandan, dizinlerin ilişkilendirilmesi düşüncesi, Vannevar Bush'un ünlü Memmex'inde yansıtmaya çalıştığı yaklaşıma çok yakın düşmektedir; zira ilişkili dizinleme (associative indexing) Bush'un Memmex'in temeli olarak gördüğü yaklaşımdır (Bush, 1837).

İstatistiksel Çözümleme ve Raporlama Modülü Pearson Çarpım-moment Korelasyon Katsayısı "PMCC" (Pearson Product-moment Correlation Coefficient) yaklaşımına dayalıdır. Ancak sistemde ortaya çıkartılan alıştırma verisi, tam olarak anlamlı bir istatistiksel sonuç vermekten uzaktır. Bunun sağlanabilmesi için, kesin tanımlı bir araştırma evrenindeki gerçek veriden yola çıkılmalıdır ki, bu da 1928'den bu yana basılmış olan tüm Türkçe gazetelerin dizinlenmesiyle mümkün olmakla birlikte, böyle bir çalışma 107K209 Projesinin kapsamı dışındadır. Proje ancak bu türden bir araştırmanın yapılmasını sağlayabilecek altyapının geliştirilmesiyle sınırlıdır ve bu amacını da proje sonunda gerçekleştirmiştir.

Alıştırma verisiyle yapılan test çalışmaları sırasında, örneğin Türkiye ile enflasyon arasındaki PMCC 0,8954 çıkarken, IMF ile enflasyon arasındaki PMCC 0.9051'dir. Yani Türk Basınında enflasyon kavramıyla Türkiye ve IMF arasında kurulan ilişki çok güçlü bir korelasyona sahiptir, ancak IMF ile enflasyon kavramı arasındaki ilişki, Türkiye ile enflasyon kavramı arasında ilişkiden görece daha güçlüdür. Dahası Türkiye ile IMF arasındaki korelastın OMMCC 0,6256 çıkması da ilginç bir bulgudur. Öte yandan son dönemde yoğunlaşan Türkiye AB ilişkileri nedeniyle, PMCC 0.9955'le çok güçlü bir korelasyon ortaya çıkmaktadır.

Sonuç ve Öneriler

107K209 Projesi başarılı bir şekilde sonuçlandırılmıştır. Projenin öneri formunda yer alan tüm başarı ölçütlerinin tamamı hayata geçirilmiş olup, sistem eksiksiz olarak çalışmaktadır. Sisteme araştırma önerilerinin kabulü için T.C. Yeditepe Üniversitesi'nde bir prosedür hazırlanmaktadır. İlgili prosedürle yurtiçi ve yurtdışında gelen araştırma talepleri Yeditepe Üniversitesi yönetimince değerlendirilerek, ilgili araştırmacılar veya kuruluşların sistemden yararlanması sağlanacaktır. Ayrıca sistem, uluslararası planda bir dizi önemli bilimsel işbirliği çalışmasının da önünü açmıştır.

Sistemin geliştirilmesi gereken yönleri arasında özellikle KA Katsayısı'nın sisteme uyarlanması gereği ortaya çıkmaktadır. Bu konuda, bir "Hızlı Destek" projesi hazırlanarak TÜBİTAK'tan yeniden destek istenmelidir.

Projenin yol açtığı yeni bir araştırma alanı da "Söylem Çözümlemesi"dir. Bu alanda ilk araştırma desteği önerisi, AB 7. Çerçeve Programı Fikirler Özel Programına gerçekleştirilmiştir. Sistem bu ve benzeri alanlarda geliştirilmelidir.

Kaynakça

Anderson, T., ve Song, A. (2008). *Next generation projective techniques: Combining psychological content analysis and text mining in market research*. Stamford: Anderson Analytics.

Arıkan, A. (2006). *Bilgi erişim sistemleri: Bilgi erişimde dil sorunları*. İstanbul: Babil.

Atılgan, D. (1992). *Kataloglamada standardizasyon açısından Türkiye Bibliyografyası'nın içerik analizi.* Yayımlanmamış doktora tezi. Ankara Üniversitesi: Ankara.

Babbie, E. (2004). *The practice of social research* (10. bs.). Belmont, CA: Wadsworth.

Berelson, B. (1952). *Content analysis in communication research*. New York: Free Press.

Bilgin, N. (2006). *İçerik analizi: teknikler ve örnek çalışmalar*. Ankara: Siyasal.

Bush, V. (1837). *As we may think*. 6 Mayıs 2008 tarihinde www.theatlantic.com/doc/194507/bush adresinden erişildi.

Çebi, M. S. (Yay. Haz.). (2003). *İletişim araştırmalarında içerik çözümlemesi*. Ankara: Alternatif.

Gökçe, O. (2006). *İçerik analizi: Kuramsal ve pratik bilgiler*. Ankara: Siyasal.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2. bs). Thousand Oaks: Sage.

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

Manning, C. D. ve 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.

Oflazer, K. (2009). *Kemal Oflazer's recent publications*. 2 Nisan 2009 tarihinde http://people.sabanciuniv.edu/oflazer/pubs.html adresinden erişildi.

Sherr, S. ve Staples, M. (2004). *News for a new generation: Report 1: Content analysis, interviews, and focus groups.* New York: CIRCLE.

Tavşancıl, E. ve Aslan A. (2001). *İçerik analizi ve uygulama örnekleri*. İstanbul: Epsilon.

Tonta, Y. (2001). Bilgi erişim sorunu. *21. Yüzyıla Girerken Enformasyon Olgusu Sempozyumu*'nda sunulan bildiri. 04 Mayıs 2007 tarihinde http://yunus.hacettepe.edu.tr/~tonta/yayinlar/tonta-hatay-bildiri.htm adresinden erişildi.

Weber, R. P. (1990). *Basic content analysis* (2. bs.). Newbury Park: Sage.

Yontar, A. ve Yalvac, M. (2000). Problems of Library and Information Science Research in Turkey: A Content Analysis of Journal Articles 1952-1994. *IFLA Journal, 26*, 39-46.

Zhang, P., Huang, Y., Shekhar, S. ve Kumar V. (2002). *Technical report: Correlation analysis of spatial time series datasets: A filter-and-refine approach*. Minsota: Department of Computer Science and Engineering University of Minnesota.

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