TARİHSEL DERLEM KAVRAMI, ESKİ TÜRKÇE VE KARAHANLI TÜRKÇESİNİN TARİHSEL DERLEMİ (7.-13.YY.)*

Yrd. Doç. Dr. Engin ÇETİN
Çukurova Üniversitesi
ecetin@cu.edu.tr

Yrd. Doç. Dr. Bülent ÖZKAN
Mersin Üniversitesi
ozkanbulent@gmail.com

ÖZET

Bilişim teknolojisindeki hızlı gelişmeler, başta dilbilim olmak üzere birçok bilim dalında bilgisayarın kullanılmasını zorunlu hale getirmiştir. Bu durumun temel sebebi, bilgisayarlar aracılığıyla bilginin daha kolay işlenebilmesi ve sınıflandırılabilmesidir. Bu açıdan bakıldığında, dil çalışmalarında da bilgisayarlı uygulamaların oldukça önem kazandığını görülmektedir. Dünyada 1900'lü yılların ikinci yarısından itibaren *derlem dilbilim* (corpus linguistics) araştırmalarının önem kazanmasıyla birlikte söz konusu alanın içeriği genişlemiş ve alt dallara ayrılmıştır. Bu anlamda, derlem dilbilim uygulamalarında bilgisayarlar aracılığıyla işlenen insan dili, farklı alt araştırma alanlarını (sözlükbilim, dil öğretimi, söylem çözümlemesi vb.) kendine konu edinmeye başlamıştır. Türkçe içinse bu alan henüz oldukça yenidir.

Literatürde derlemler, amaç ve içlemleriyle koşut olarak farklı uygulamalarla karşımıza çıkmaktadır. Bu derlem türlerinden biri de *art süremli/tarihsel derlemler* (diachronic/historical corpus)'dir. Art süremli/tarihsel derlemler, belirli bir dilin tarihsel metinlerini sayısallaştırarak bilgisayarlar aracılığıyla işlenebilir-okunur bir platforma taşımayı ve araştırmacıların hizmetine sunmayı amaçlar. Bu bildiride Türkçe için tarihsel derlem oluşturma çabalarının önemli ve ilk adımlarından biri olan Eski Türkçe ve Karahanlı Türkçesinin Tarihsel Derlemi (ETKT-D) tanıtılacaktır.

Anahtar Sözcükler: Derlem dilbilim, tarihsel derlem, Eski Türkçe, Karahanlı Türkçesi, söz varlığı.

ABSTRACT

Fast improvements in informatics resulted in the vast usage of computer sciences in more sciences especially in linguistics. As corpus linguistics researches (the researches on corpus linguistics) gained importance in throughout the world since late 1900s, sub-sections of this sub-section also have begun to occur. Human language,studied in corpus linguistics applications by means of computers, has been taken as subject in different sub-research areas (lexicology, language teaching, pronunciation analyses, etc.). This is however rather new for Turkish.

Corpora are defined in different types parallelly with their meanings and purposivenesses in literature. One of these types is diachronic/historical corpus. This corpus type aims at carrying the texts written in a definite language to a workable-legible platform thanks to computers and offering them to the researchers' service by making numerical.

Keywords: Corpus linguistics, historical corpus, vocabulary of Old Turkic, vocabulary of Karakhanid.

1.Giriş

Bilişim teknolojisinin ve bilgisayar bilimlerinin hızla gelişimi, bilgisayarı her alanda daha yaygın kullanılır bir araç haline getirmiştir. Öyle ki, günümüzde belirli alanlar için bilgisayarın vazgeçilmez olduğu görülmektedir. Bilgi teknolojilerinde ve bilgisayar bilimlerinde yaşanan bu hızlı gelişmeler, dil araştırmalarının da bilgisayar destekli olmasını gerektirmiş, tüm bu gelişmeler, Doğal Dil İşleme (DDİ) çalışmalarının bir mühendislik alanına dönüşmesini sağlamıştır. Günümüzde Türkiye'de de bilgisayar destekli dil çalışmalarının ivme kazandığı gözlenmektedir. Son on yılda Türkiye'de, güvenilir ve hızlı erişilebilir bir ortamda sözlükbilim ve sözdizimi ile ilgili çalışmalar bu anlamda öne çıkmaktadır. Bu çalışmalar, Türkçenin sıklığı yüksek yapılarını ortaya koymayı ve böylelikle özellikle yabancı dil ve ana dili öğretiminde öncelikli yapıları ortaya koymayı amaçlayan çalışmalar olarak dikkat çekmektedir.

Bilgisayar bilimlerinin sunduğu olanaklar doğrultusunda bilgisayarlı dilbilimde *derlem* (corpus) ya da *bütünce* olarak adlandırılan yapı, 'belirli amaçlar doğrultusunda yazılı veya sözlü dile dayalı ürünlerin birtakım ölçünlenmiş yöntem ve işaretlemelerle bir araya getirilmesinden oluşan bütün' olarak dil çalışmalarında yerini almıştır (Say *vd*. 2002). Var olan biçimiyle derlemler amaçlılıkları ve içlemleriyle koşut olarak yazılı ve/ya sözlü içeriğe sahip ve belirli bir dilde dilsel çeşitliliği yansıtabilen *genel derlemler* (general corpora); petrokimya derlemi ya da bilgisayar bilimleri derlemi benzeri *özel alan derlemleri* (specialized corpora); belirli bir yazı dilini, belirli dönem ve örneklemlerle içeren *yazılı derlemler* (written corpora); genel derlemlerin içleminde ya da bağımsız olarak sözlü dili temsil için tasarımlanan *sözlü derlemler* (spoken corpora); eş süremli olarak dilsel veriler içeren *eş süremli derlemler* (synchronic corpora); belirli bir dilin farklı zaman aralıklarıyla içlemlendiği *art süremli derlemler* (diachronic or historical corpora); yine ana dili ve yabancı dil olarak belirli bir dilin öğretiminde temel verilerin derlenmesi amaçlı tasarımlanan *öğrenici derlemleri* (learner corpora) ve diğer derlem türlerinden daha hacimli ve ayrıntılandırılmış, kapsamlı

* Bu çalışma, TÜBİTAK tarafından desteklenen 110K048 numaralı, "Eski Türkçe ve Karahanlı Türkçesinin Tarihsel Derlemi (7.-13. yy.)" başlıklı Ulusal Araştırma Projesi kapsamındadır.

derlemlerden olan *izlem derlemleri* (monitor corpora) olarak karşımıza çıkmaktadır (McEnery *vd*. 2006).

Tarihsel derlem, özetle herhangi bir dilin herhangi bir dönemine ait metinleri derlem dilbilimin yöntemleriyle bilgisayarlar aracılığıyla okunur-işlenir bir yapıyla bir araya getirilmesi ve araştırmacıların hizmetine sunulmasıdır.

Bu alana özgü farklı dünya dilleri üzerine yürütülmüş çalışmaların bazıları şunlardır:

Helsinki Derlemi olarak da bilinen Helsinki Üniversitesi İngiliz Dili Bölümü'nde 1984-1991 yılları arasında hazırlanan 1,572,800 sözcüklük *The Helsinki Corpus of English Texts* (http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/generalintro.html), Eski İngilizce, Orta İngilizce ve Modern Öncesi Dönem İngilizcesi (730-1710 yılları arası) ile yazılmış 450 civarında metni içermektedir.

Zürich Corpus of English Newspapers (ZEN) (http://es-zen.unizh.ch/), Zürich Üniversitesi İngiliz Dili Bölümü'nce 1993-2003 yılları arasında hazırlanmıştır. 1661-1791 yılları arasındaki 349 gazetenin söz varlığından oluşan 1,6 milyon sözcüklük bir derlemdir.

The ARCHER Corpus (a corpus of British and American English from 1650-1990) (http://www.nau.edu/english/ling/), 1,7 milyon sözcük içerir. 1037 metin ve 10 kayıttan (drama, mektup, şiir vb.) oluşmaktadır.

Corpus Del Español (http://www.corpusdelespanol.org), 1200'lü yıllardan 1900'lü yıllara kadar İspanyolca yazılmış 13,926 metinden oluşan 101,311,682 sözcüklük bir tarihsel derlemdir.

Konuyla ilgili bir başka çalışma ise Cambridge Üniversitesi Dilbilim Bölümünce 2001-2004 yılları arasında hazırlanan *A Historical Corpus of the Welsh Language 1500-1850* (http://people.pwf.cam.ac.uk/dwew2/hcwl/menu.htm)'tir. Bu derlem, 1500 ile 1850 yılları arasında Gal dili ile yazılan 30 metni içeren yaklaşık 420.000 sözcük barındırmaktadır.

The Corpus Of Historical American English (Coha) ise Amerika Birleşik Devletleri'ndeki Brigham Young University'de gerçekleştirilen 1810-2009 yılları arasındaki 107 bin metni 400 milyonu aşkın sözcüğü kapsayan bir derlemdir.

Türkçeyi tarihsel olarak konu edinen, bir çalışma Johann Wolfgang Goethe Üniversitesi'nde yürütülen *Vorislamische Alttürkische Texte: Elektronisches Corpus* (http://vatec2.fkidg1.uni-frankfurt.de/), çözümlemeli bir derlemdir ve Eski Türkçenin Uygurca dönemine ait metinleri içermektedir. 1999-2003 yıllarında oluşturulan derlem projesi Alman Araştırma Vakfı (Deutschen Forschungsgemeinschaft) tarafından desteklenmiştir.

Görüldüğü gibi, dünyada tarihsel derlem oluşturma çalışmaları bilgisayar bilimleri paralelinde düşünüldüğünde çok eskilere götürülemez. Ancak, bu alanda kısa zamanda önemli çalışmalar yapılmış, tarihsel derlem konusunda büyük aşamalar kaydedilmiştir. Türkçe için ise tarihsel derlem çalışmalarının yeterince yapıldığını söylemek pek mümkün olmasa da derlem çalışmaları konusunda Türkiye'de önemli çalışmaların yapıldığını söylemek mümkündür. Derlem oluşturma çalışmalarının Türkiye'de bilinen ilk örneği Say, B., *vd*.'nin Bilgisayar Ortamında Bir Derlem Geliştirme Çalışması (Enformatik Enstitüsü Bilişsel Bilimler Ana Bilim Dalı, Orta Doğu Teknik Üniversitesi)'dır. Bu çalışma dilbilim ve doğal dil işleme çalışmalarına kaynak olmak üzere elektronik ortama geçirilen günümüz Türkçesini yansıtan metin örneklerinin işaretlenmesiyle 2 milyon sözcükbirim içeren bir derlemdir (METU-Sabancı Turkish Treebank). Söz konusu derlem XML (Extansible Mark-up Language) ile işaretlenmiştir. Mersin Üniversitesi İngiliz Dili ve Edebiyatı Bölümü'nün 50 milyon söz içeren Türkiye Türkçesinin Ulusal Derlemi'ni oluşturma çabalarının sürdürüldüğü de bilinmektedir (http://www.tudd.org.tr/).

Türkiye'de, Türkçe için bir derlem oluşturmak bugün özellikle bilgisayar bilimleriyle uğraşan araştırmacıların üzerinde durdukları bir konudur. Özellikle Türkçenin morfolojik çözümlemesini yapmayı hedefleyen bu çalışmaların en önemlileri Kemal Oflazer'in Two-level description of Turkish morphology, Literary and Linguistic Computing ve *vd*. Tagging and Morphological Disambiguatıon of Turkish Text başlıklı çalışmalarıdır.

Türkçe için bir derlem oluşturmak ve Türkçenin söz varlığını bilgisayar destekli araştırmalarla ortaya koymak konusunda Türk dili uzmanları da yeni ve değerli çalışmalar yapmaktadır. Bunlar arasında 12 milyon söz içeren ve Cumhuriyet Dönemi Türk Yazın Dili'ni temsil eden TÜRKÇE DERLEM-1, halen devam eden TÜBİTAK destekli projelerden olan "Türkiye Türkçesi Söz Varlığında Sıfatların Eşdizimliliği -Derlem Tabanlı Bir Uygulama-" (http://turkcederlem.mersin.edu.tr/) ve "Türkiye Türkçesi Sözvarlığında Fiillerin Derlem Denetimi ve

Derlem Tabanlı Sözlüğü" (http://derlem.mersin.edu.tr/derlem516/) alan için önemli çalışmalar olarak dikkat çekmektedir.

2.Eski Türkçe ve Karahanlı Türkçesinin Tarihsel Derlemi (ETKT-D)

ETKT-D'de, Eski Türkçenin Orhon Türkçesi ve Uygurca dönemlerine ait metinler ile Karahanlı Türkçesi dönemi metinler yer alacaktır. Bu üç dönem bilindiği gibi, Türklerin bağlı bulundukları kültürel alanlar dolayısıyla etkileşimde olunan diller açısından farklılıkların olduğu dönemlerdir. Orhon Türkçesi dönemi göçebe (ya da yarı göçebe) olan Türk topluluklarının dillerinin de dış etkilere -sonraki dönemlere oranla- daha kapalı olduğu bir dönemdir. Uygurca döneminde ise özgün metinlerin yanında Budizm, Maniheizm ve Hıristiyanlık dinlerinin etkisiyle bu dinlere özgü metinlerin çevrilmesi sırasında Türkçeye çok sayıda sözcük girmiştir. Bu dönemdeki yoğun çeviri etkinlikleri nedeniyle Türkçenin çok sayıda yeni Türkçe sözcük de kazandığını belirtmek gerekir. Karahanlı Türkçesi döneminde ise Türklerin İslamiyet'i kabulü dolayısıyla bu dönem metinleri İslami çevre metinleridir. Bu dönemde de Uygurca döneminde olduğu gibi çeviri ve özgün eserler verilmiştir. Bu dönemde Türkçe, Arapça ve Farsça ile etkileşim içindedir. Bu üç döneme ait metinler edebi türler açısından da çeşitlilik arz etmektedir. Bu dönemlerde, söylev niteliğindeki yazıtlar, fal kitabı, sutra çevirileri, dini içerikli çeşitli anlatılar, masallar, ilahiler, tövbe duaları, Kur'an tercümesi, şiirler ve şiir parçaları, didaktik metinler, din dışı konular içeren hukuk belgeleri, sözleşmeler, vasiyetnameler, mektuplar vb. türlerde eserler verilmiştir. Eski Türkçe ve Karahanlı Türkçesi metinlerinin genel olarak 7.-13. yy.'da yazılmış metinler olduğunu söylemek mümkündür.

Amaç

Projenin amacı, 7. yy.'ın sonlarından başlayarak yazılı ürünlerle izleyebildiğimiz Türk yazı dilinin Eski Türkçe ve Karahanlı Türkçesi dönemlerinde yazılan metinleri derleyerek, söz konusu dönem metinlerinin söz varlığını, bilgisayarlar aracılığıyla araştırmacıların kolaylıkla erişebileceği dil malzemesi haline getirmek ve Türkçe için *art süremli/tarihsel derlem* bir oluşturmak ve devam çalışmalarında önemli bir altyapı olanağını araştırmacılara sunmaktır.

Önem

Türkçenin yazıya ilk geçirildiği dönem olan Eski Türkçe ve Karahanlı Türkçesinin söz varlığını bilgisayarlarca okunabilir-işlenebilir hale getirmek; söz varlığı ve sözdizimi açıdan ilgili dönemin Türkçesini sayısal platforma taşımak; 7.-13. yüzyıllar Türkçesiyle ilgili olarak araştırmacıların erişimine açık bir platform oluşturarak araştırmacılara dönemle ilgili dilsel malzemeye erişim kolaylığı sağlamak; söz konusu dönemlerin söz varlığına ve sözdizimi özelliklerine ilişkin karşılaştırmalı ve istatistiksel çalışmalar yapmaya olanak sağlanmak; kültür tarihi araştırmalarında söz varlığı izleğinde katkıda bulunmak; *"Türkçenin Tarihsel Sözlüğü"*nün hazırlanmasında Türkçenin ilk ürünlerini bu anlamda yordanabilir hale getirmek amaçla bağlantılı olarak çalışmanın önem taşımaktadır.

Kapsam

Temsil gücü yüksek bir derlemin oluşturulabilmesi için yöntemle ilgili ilk aşama kuşkusuz dönemi temsil edebilecek sayıda ve nitelikte eserlerin seçilmesidir. Bu, projenin başarılı olmasının en temel koşullarından biridir. Bu amaçla, farklı türlerde ve yeterli sayıda yazılı ürünün seçilmesi dönemin genel söz varlığının yeterli derecede temsil edilmesine olanak sağlayacaktır.

Yukarıda da belirtildiği gibi, dönem ürünleri tür açısından çeşitlilik arz etmektedir. *Orhon Türkçesi* döneminde çoğunluğu taşa, bir ölünün ardından yazılan metinler, kimi zaman salt kendini geleceğe taşıma çabasıyla yazılmış bir tür not görünümündeyken kimi zaman da ölünün geçmişte yaptıklarını, savaşlarını, başarılarını anlatan, gelecek kuşakların ders almasının beklendiği söylev niteliğinde ürünlerdir. Bu yazıtlar, hacimce sınırlı olmalarına karşın dil özellikleri ve söz varlığı açılarından oldukça gelişmiş ve geniş kapsamlı bir yapıya sahiptir.

Uygurca döneminde yazılan metinler ise Orhon Türkçesindeki metinlere oranla daha hacimli, dil özellikleri ve söz varlığı açılarından daha gelişmiş bir görünüm arz etmektedir. Bu dönemde kabul edilen dinlerin etkisiyle, bu dinleri anlamak, öğrenmek ve yaygınlaşmasını sağlamak amacıyla dini içerikli metinleri Çince, Sanskritçe, Toharca ve Soğdca gibi dillerden çeviren Uygurlar, özgün yapıtlar da meydana getirerek son derece gelişmiş bir edebiyat dili yaratmışlardır. Bu dönemde, masallar, hikâyeler, biyografi türünde eserler, dualar, tövbe duaları, fal kitabı, dini bilgiler içeren kitaplar,

sutralar, ilahiler gibi dini içerikli eserler yanında vasiyetname, kira sözleşmesi, satış sözleşmesi, mektup gibi din dışı konularda düzyazı ve şiir türünde ürünler meydana getirilmiştir.

Karahanlı Türkçesi döneminde İslamiyet'in Türklerce kabulü dolayısıyla Kur'an tercümelerinin yapıldığı gözlenmektedir. Bunun yanında bu dönemin en önemli ürünleri Türk dilinin de en önemli yapıtları olan Kutadgu Bilig ve Divanü Lugati't-Türk'tür. Atebetü'l-Hakayık ise Karahanlı Türkçesi döneminde kaleme alınmış didaktik türde bir eserdir.

Tablo 1. *Derlemin içeriğinde yer alacak metinler (dönem-metin-yüzyıl-tür)*

Dönem	Metin	Yüzyıl	Metin Türü
		7-9.	
Orhon Türkçesi	Çoyr Yazıtı	7. yy. sonu (?)	Düzyazı
	Küli Çor Yazıtı	8. yy	Düzyazı
	Ongin Yazıtı	8. yy	Düzyazı
	Kül Tigin Yazıtı	8. yy	Düzyazı
	Bilge Kagan Yazıtı	8. yy	Düzyazı
	Tunyukuk Yazıtı	8. yy	Düzyazı
	Tes Yazıtı	8. yy	Düzyazı
	Taryat Yazıtı	8. yy	Düzyazı
	Şine Usu Yazıtı	8. yy	Düzyazı
	Süci Yazıtı	9.yy	Düzyazı
		9.-11.	
Uygurca	Irk Bitig	9. yy.	Düzyazı
	Manichaica I-III	Muhtelif	Düzyazı-şiir
	Huastuanift	9.-10.yy.	Düzyazı
	Rüzgâr Tanrısı	8.-9. yy.	Düzyazı
	Die türkischen Yosïpas-Fragmente	9. yy.	Düzyazı
	Maniheist şiirler	Muhtelif	Şiir
	Uigurica I-IV	Muhtelif	Düzyazı-Şiir
	TT I-X	Muhtelif	Düzyazı-Şiir
	Altun Yaruk	10. yy.	Düzyazı
	Maitrisimit Nom Bitig (Hami)	13. yy.(?)	Düzyazı
	Hsüen Tsang Biyografisi	10.-11. yy.	Düzyazı
	Hamam Sutrası		Düzyazı
	Ölüler Kitabı	13.-14. yy.	Düzyazı
	Bahşı Ögdisi	13.-14. yy.	Düzyazı
	Et'öz-üg könül-üg körmek atl(ı)ġ nom bitig	10. yy.	Düzyazı
	Sekiz Yükmek	10., 13-14. yy.	Düzyazı
	Vimalakīrtinirdeśasūtra	?	Düzyazı
	Saddharmapuṇḍarīka Sūtra	9.-10. yy.	Düzyazı
	HSİN Tözin Okıtdaçı Nom	13.-14. yy.	Düzyazı
	Sadd. Dharmagotta Bod. Hikâyesi	13.-14. yy.	Düzyazı
	Kalyanamkara Papamkara	10. yy.	Düzyazı
	Kşanti Kılguluk Nom Bitig	13. yy.	Düzyazı
	Hıristiyanlığa İman Metni (Zieme 1998 Nest.)	13.-14. yy.	Düzyazı
	Dindışı metinler (mektuplar, şiirler vs.)	Muhtelif	Düzyazı-Şiir
	Hukuk Belgeleri (sözleşmeler, vasiyetnameler)	13.-14. yy.	Düzyazı
	Budist Uygur şiirleri	Muhtelif	
		11.-13.	
Karahanlı Türkçesi	Kur'an Tercümesi Rylands Nüshası	11. yy	Düzyazı
	Kutadgu Bilig	11. yy	Şiir
	DLT'deki Türkçe parçalar	11. yy	Düzyazı-Şiir
	Atebetü'l-Hakâyık	12.-13. yy	Şiir

Sınırlılıklar

ETKT-D'nin içeriğini oluşturacak sözcük sayısının 400-500 bin arasında olması beklentisi göz önünde bulundurulduğunda, yurt dışında hazırlanan yabancı dil derlemlerinin sözcük sayısı açısından Türkçe derleme göre oldukça fazla olduğu görülmektedir. Bu durumu, 7.yy. sonu ile 13. yy.'daki Türkçe metinlerin azlığına bağlamak mümkündür. Oysa, yukarıda yer alan derlemlere bakıldığında Helsinki Derlemi ve VATEC dışındaki derlemlerin, 13. yy. ve sonrasında yazılan metinlerdeki söz varlığını içerdiği görülecektir. Bu durum, söz konusu derlemlerin içerdiği sözcük sayısının fazla olmasını sağlamıştır. Çünkü dünyadaki gelişmeler, yazı dillerinin işlenmişliği vb. etkenler 10. yy.'dan sonra her dil için çok sayıda metnin yazılması sonucunu doğurmuştur. Bu açıdan örneğin, 13. yy.'dan sonra Anadolu'da yazılan metinlerden oluşan bir derlem, 13. yy.'dan önce Türk yazı dilleriyle yazılmış metinlerden oluşan derleme göre daha çok sözcük içermesi doğal karşılanmalıdır.

3.YÖNTEM

3.1.Derlem Oluşturma ve Veri İşleme Aşamaları

Bugün için, söz konusu 7-13. yüzyıla ait metinler ve bunların türleri belirlidir. Derlemde yer alacak metinler ilgili dönemde yazıya geçirilmiş olanları kapsamaktadır.

Derlem dilbilimi konusunda daha önce de değinildiği gibi, Türkiye'deki ve Batı'daki çalışmalar karşılaştırıldığında Türkiye için bu alanın henüz çok yeni olduğu görülmektedir. Bu nedenle, Türkçe için temsil gücü yüksek bir derlemin oluşturulması yapılması gereken öncelikli çalışmalar arasındadır.

Türkçe için temsil niteliği olan bir derlemin oluşturulamamasının nedenlerinden biri, bir derlemin, bilgisayar tarafından okunabilecek yapıda işaretlenmiş olması zorunluluğudur. Bu işaretlemeler, *XML* (Extansible Mark-up Language), *SGML* (Standard Generalized Mark-up Language) ve *TEI* (Text Encoding Initative) gibi işaretleme dilleri aracılığıyla yapılan ayrıntılı işaretlemelerdir. Bunun yanında kimi derlemlerin kendilerine özgü etiketleme sistemi geliştirdikleri de görülmektedir. Derlem işaretleme işlemleri zaman alıcı ve yüksek maliyetli çalışmalardır. Yukarıda sayılan ya da bunlara benzer işaretleme dilleriyle bilgisayarlarca okunabilir hale getirilen derlemler Java, Perl, PHP vb. programlama dilleriyle hazırlanan görsel arayüze sahip yazılımlarla sorgulanabilmektedir. Böylelikle dilsel verilerin görünümlerinin belirlenmesi, sorgulanabilmesi hızlı ve güvenilir biçimde kullanılması dijital bir ortamda mümkün olabilmektedir.

Bu çerçevede *bütünleşik* olarak ETKT-D'nin veri işleme ve sorgu aşamaları şu şekildedir:

1- Derlem oluşturma.
2- Derlem işaretleme (işaretleme ve etiketleme).
3- Derlem sorgu sistemi oluşturma.

3.1.1. Derlem Oluşturma ve İşaretleme (1. ve 2. Aşama)

Genelde derlem oluşturma aşamasında kullanılan OKT (Optik Karakter Tanıma) yazılımları bu tip tarihsel derlemlerin oluşturulması aşamasında işe yaramamaktadır. Bunun nedeni özel karakter setlerinin kullanılma zorunluluğudur. Döneme özgü alfabenin örneğin 'altı noktalı k' 'altı noktalı h' üstü noktalı 'g' vb. çeviri yazı karakterleri bu tip yazılımlarda tanınamadığı için bilgisayarın okuyup işleyebileceği bir formata dönüştürülememektedir. Metinler (*bk.* Tablo 1) bu anlamda yeniden yazılacaktır.

Derlem için seçilen metinlerin ortak yazı karakterleri belirlendikten sonra bilgisayar ortamına aktarılmasıyla birlikte yukarıda söz edilen uygulamalar ve dil programları aracılığıyla metinlerin işaretlenerek bilgisayarlarca okunabilir hale getirilmesi işlemi gerçekleştirilmiş olacaktır. Çalışmanın özelliği gereği çeviri yazı karakterlerinin bilgisayarca tanınması güçlüğü aşağıda anlatıldığı biçimiyle aşılmıştır.

3.1.1.1.Eski Türkçe Çeviri Yazısının Sunucu ve İstemciye Tanımlanması

Eski Türkçe karakterler standart olarak kullanılmadıklarından projede kullanılacak bilgisayarların bu karakter setine göre ayarlanması gereklidir. Bu işlem iki aşamalı olarak aşağıdaki gibidir.

I-) Sunucu Ayarı

Veritabanı Ayarı (MySQL veritabanı için)

*Veritabanı karakter seti "utf8_turkish_ci" olarak ayarlanmalıdır.

*mysql_connect( ) işlevinden sonra mysql_query( ) işlevi aracılığıyla sırasıyla şu sorgular yapılmalıdır:

SET NAMES "utf8"
'SET CHARACTER SET "utf8"'
'SET COLLATION_CONNECTION="utf8_turkish_ci"
Böylelikle veritabanı ve programlama dili arasında uyum sağlanmış olur.

Kodlama Ortamı Ayarı

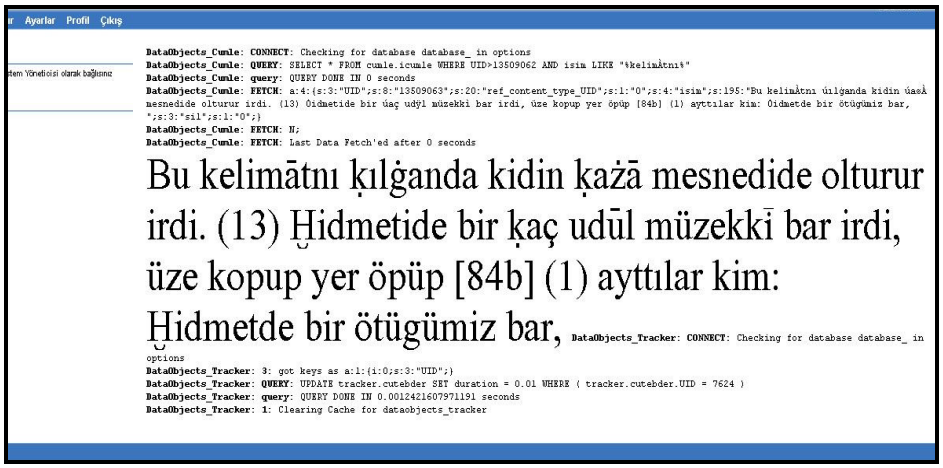
Kod dosyaları utf8 biçiminde kaydedilmelidir.

HTML Çıktısı Ayarı

Eski Türkçe yazılar komutları arasına yazılmalıdır.

II) İstemci Ayarı

İstemci bilgisayarda işlem biraz daha kolaydır. İstemci bilgisayarların font karakter setleri dizinine (Windows XP sürümü için standart yeri "C:\WINDOWS\Fonts" dizinidir) "Ali Sir Nevayi.ttf" fontunu kopyalamamız ve çıkan soruya "Evet eklemek istiyorum" dememiz yeterlidir (*bk.* Şekil 1 Karakter seti ayarları yapılmış sistem çıktısı).



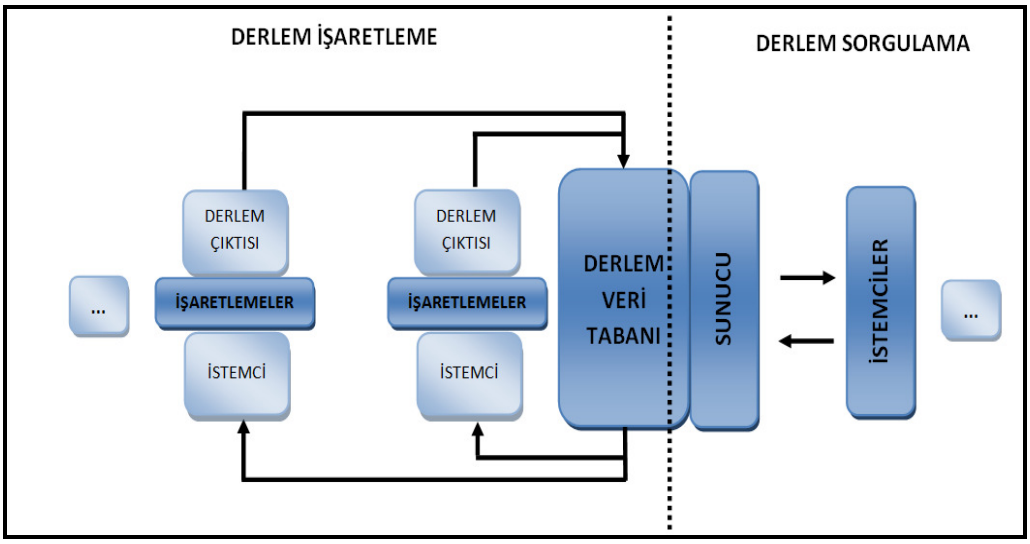
Şekil 1 Karakter seti ayarları yapılmış sistemin veri çıktısı

3.1.2. Derlem Sorgu Sistemi Oluşturma (2. ve 3. Aşama)

Derlem veri tabanı yukarıda belirttiğimiz *derlemde yer alacak metinlerin seçimi* ve *seçilen metinlerin sayısallaştırılması* aşamalarının sonucunda oluşturulacaktır. Oluşturulan bu derlem üzerinden sözcükbirimsel ve sözdizimsel işaretlemeler ile diğer içerik işaretlemeleri (metin türü, saha, yüzyıl vb.) proje için oluşturulacak bir sunucu (server) üzerinden araştırmacı ve bursiyerlerce yapılacaktır.

Öte yandan kurulacak olan sistem eş zamanlı olarak kullanıcılara açılacak, çalışma süresince işlenen veriler yayınlanacak, kullanıcılar sistem üzerinden sorgulama vb. işlemleri yapabileceklerdir (*bk.* Şekil 2 Veri işaretleme ve derlem sorgulama diyagramı).

Projenin üzerinden yürütüleceği söz konusu sistem projede yer alacak olan yazılımcı ve araştırmacılarca sınanmış ve amaca uygunluğu test edilmiştir.



Şekil 2 Veri işaretleme ve derlem sorgulama diyagramı

3.1.2.1.Biçimbirimsel Değişkenlerde Sözcükbirimsel Sorgulamalar

Öte yandan, Eski Türkçenin Orhon Türkçesi ve Uygurca dönemi ile Karahanlı Türkçesi dönemleri ses, biçim vb. dil özellikleri açısından benzer pek çok yönünün yanında kimi farklılıkları da barındırmaktadır. Örneğin Orhon Türkçesinin söz içi ve söz sonu *-b-, -b* sesleri Uygurcada *-v-, -v*; Karahanlı Türkçesinde *-w-, -w* olarak gelişmiştir. Dolayısıyla Orhon Türkçesinin *eb* "ev" sözcüğü Uygurcada *ev*; Karahanlı Türkçesinde *ew* biçimindedir. Bu durumda okuyucu derlemde *eb* sözcüğünü aradığında, sözcüğün yanında *ev* ve *ew* sözcüklerini ve bunların ilgisini görebilmelidir. Bunun gibi, Orhon Türkçesinin *anyıg* "kötü, fena" sözcüğü, Uygurcada *anıg* ve *ayıg*; Karahanlı Türkçesinde *ayıg* biçiminde görülmektedir. Okuyucu, *anyıg* sözcüğünü sorguladığında farkı üç biçimi de görebilecektir. Söz konusu güçlük sözcüklerin derleme kayıtları sırasında yapılacak anlam işaretlemeleriyle aşılabilmiştir (bk. Sorgu, anlam eşleşmesi ve çıktı).

<Sorgu>	**<Anlam Eşleşmesi>**	**<Çıktı>**
<anıg>, <ayıg>, <anyıg>	*"kötü, fena"*	*<anıg>, <ayıg>, <anyıg>*
<ew>, <ev>, <eb>	*"ev"*	*<ew>, <ev>, <eb>*

Şekil 2. Sorgu, anlam eşleşmesi ve çıktı.

Tüm bu işlemlerin tamamlanmasının ardından yine derlemimiz için oluşturulacak bir arayüzle derlem kullanıcının hizmetine sunulacaktır. Kullanıcı, projenin tamamlanmasıyla, aradığı sözcüğün türünü, kullanım sıklığını, kullanıldığı metinleri, yazım özelliklerini hızlı ve güvenilir bir biçimde görme olanağına sahip olacaktır.

Sonuç

"Eski Türkçe ve Karahanlı Türkçesinin Tarihsel Derlemi Projesi"nden beklenen sonuçlar şunlardır:

1- Türkçenin yazıya ilk geçirildiği dönem olan Eski Türkçe ve Karahanlı Türkçesinin söz varlığını bilgisayarlarca okunabilir-işlenebilir hale getirilmiş olacaktır.

2- Söz varlığı ve sözdizimi açısından ilgili dönemin Türkçesini sayısal platforma taşınmış olacaktır.

3- 7.-13. Yüzyıl Türkçesiyle ilgili olarak araştırmacıların erişimine açık bir platform oluşturarak araştırmacılara dönemle ilgili dilsel malzemeye erişim kolaylığı sağlanmış olacaktır.

4- Söz konusu dönemlerin söz varlığına ve sözdizimi özelliklerine ilişkin karşılaştırmalı ve istatistiksel çalışmalar yapmaya olanak sağlanmış olacaktır.

5- Kültür tarihi araştırmalarında söz varlığının izleğinde katkıda bulunulmuş olacaktır.

6- "Türkçenin Tarihsel Sözlüğü"nün hazırlanmasında Türkçenin ilk ürünleri bu anlamda yordanabilir hale getirilmiş olacaktır.

Kaynaklar

http://derlem.mersin.edu.tr/derlem516/
http://es-zen.unizh.ch/
http://people.pwf.cam.ac.uk/dwew2/hcwl/menu.htm
http://turkcederlem.mersin.edu.tr/
http://vatec2.fkidg1.uni-frankfurt.de/
http://www.corpusdelespanol.org
http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/generalintro.html
http://www.nau.edu/english/ling/
http://www.tudd.org.tr/

McENERY, Tony, Richard Xiao – Yukio Tono (2006) *Corpus-Based Language Studies, An Advanced Resource Book*, Routledge, Newyork.

OFLAZER, K. (1994), "Two-level description of Turkish morphology, *Literary and Linguistic Computing*, Vol. 9 No. 2.

_____*vd.* (1994), "Tagging and Morphological Disambiguation of Turkish Text", *Submitted to 4. Conference on Applied Natural Language Processing*.

ÖZKAN, Bülent (2008), *Türkiye Türkçesi Söz Varlığında Sıfatların Eşdizimliliği -Derlem Tabanlı Bir Uygulama-*, 109K104- TÜBİTAK-SOBAG Ulusal Araştırma Projesi.

_____ (2009) *Türkiye Türkçesi Sözvarlığında Füllerin Derlem Denetimi Ve Derlem Tabanlı Sözlüğü*, 109K516- TÜBİTAK-SOBAG Ulusal Araştırma Projesi.

SAY, Bilge, Deniz Zeyrek, Kemal Oflazer and Umut Özge (2002) "Development of a Corpus and a Treebank for Present-day Written Turkish", *Current Research in Turkish Linguistics*, *Proceedings of 11th International Conference on Turkish Linguistics. Eastern Mediterranean University*, (ed. Kamile İmer and Gürkan Doğan), s. 183-192, Northern Cyprus.