www.turkishstudies.net/language

Turkish Studies - Language and Literature

eISSN: 2667-5641

INTERNATIONAL BALKAN UNIVERSITY
Sponsored by IBU

Research Article / Araştırma Makalesi

Derlem Dilbilim ve Derleme Kuramsal Bir Bakış

Corpus Linguistics and Corpus A Theoretical Perspective

Ramazan Şimşek*

Abstract: Language and language studies have been going on for centuries; each researcher makes new inferences with his knowledge in the historical process. The development of technology and informatics has affected knowledge and everything related to knowledge. Corpus linguistics, which is used to examine, describe and analyze language and language practices, was also affected by this development; Corpus research carried out within institutional structures until the 1970s, with tools and software developed in the following years, decreased to the level of independent researchers. In the process, the data sets of the corpus also transformed into structures of million words. All these developments have clarified the basic principles and research areas of corpus and corpus linguistics. With the corpus studies, which were also started to be carried out by independent researchers, principles such as representativeness, balance, sample, size, machine readability, copyrights were determined for corpus design. In the following process, corpus was also divided into types such as general, special, marked, historical, marked, unmarked, student corpus. Language and language education corpus studies conducted in Turkey in recent years have been emphasized; The contributions of the project-based studies to the field have been highlighted. Within the scope of the corpus, the learning methods used by language learners and researchers were emphasized; the contribution of these methods to the process, their contribution to the concepts of autonomous learning or learning to learn; vocabulary teaching, idiom teaching; Sub-headings such as operating the principle of induction in language education were evaluated. In this framework, it is aimed to form a theoretical background in corpus and corpus linguistics researchers.

Structured Abstract: Technology, the concept of technique that changed the world completely, and the systematization of the technical concept with knowledge; education, society, economy, transport, transport and language; has also affected the research fields of these areas. With the computer, which stands out as the most important tool of the age in systematizing information, academic research, inferences and interpretation activities have been transformed. Language and communication areas were also directly affected by this transformation; A research field has emerged that brings together the concepts of computer and linguistics, which is expressed as "corporate linguistics". Corpus linguistics dates back to the 12th century with the classification studies aimed at determining the page information of the concepts, terms and subjects in the holy books of the clergy who belong to the divine religions. In Turkey, it can be said that vocabulary studies in Turkish and foreign languages form the basis of corpus linguistics studies in our country. The term "corpora", which is formed by the Latin plural suffixes of the Latin word "corp", that is, body / corpse, has been translated into Turkish as "derlem". McEnery and Hardie (2011: 32) defined the corpus as "a machine-readable body of

* Öğr. Gör. Dr., Nevşehir Hacı Bektaş Veli Üniversitesi, Rektörlük
Lect. Dr., Nevşehir Hacı Bektaş Veli Universitiy, Rectorate.
ORCİD 0000-0002-8295-8903
ramazansimsek@nevsehir.edu.tr

Cite as/ Atıf: Şimşek, R. (2021). Derlem dilbilim ve derleme kuramsal bir bakış. *Turkish Studies - Language, 16*(3), 2021-2036. https://dx.doi.org/10.7827/TurkishStudies.51828
Received/Geliş: 20 June/Haziran 2021
Accepted/Kabul: 22 September/Eylül 2021
Checked by plagiarism software
Published/Yayın: 25 September/Eylül 2021
CC BY-NC 4.0

text with the highest level of representation of the language structure under consideration". According to Gries (2009: 7), corpus; defined as the whole of written or verbal data that can be processed by the machine in the communication/information environment. According to Sinclair (2005), corpus is an electronic collection of texts selected according to certain criteria that can reflect a language or different uses of that language as accurately and comprehensively as possible for use in linguistics research. In terms of current research, the technical and linguistic aspects of the corpus process come to the fore.

Corpus Linguistics

Corpus linguistics began to take shape as a sub-discipline of linguistics with the development of computer technology in the second half of the 20th century. Corpus linguistics, which also serves research areas such as etymology, pragmatics, and semantics, focuses on the systematic collection of written oral texts in use. However, it is known that researchers in corpus linguistics have different approaches to the relationship between linguistics and corpus linguistics. In this framework, there are evaluations of corpus applications as a methodology or sub-discipline used in the field of linguistics. In recent years, corpus-based approach has begun to be accepted as a methodology in studies conducted specifically in corpus linguistics; The number of studies has also increased.

Historically, the first corpus-based studies were small sized corpuses based on phonological distinctive features; It is known that corpus-based studies on general grammar were made after the 1960s. After the 1950s, developments in computers and electronics made it possible to create an electronic corpus; Researchers have also turned to this area. The Brown Corpus, which is structured on English and has a data set of 1 million words, is known as the first corpus transferred to electronic media (Francis and Kucera, 1964). With the introduction of sound recorders in the same years, the first studies on oral corpus were carried out. Sinclair created an English oral corpus of approximately 170,000 words between 1963 and 1965 (Krishnamurthy, 2004). After these first studies in corpus linguistics, dictionary studies on English continued. However, in the early periods when corpus linguistics began to transform into a discipline, namely in the 1960s, linguists following Chomsky and his school criticized corpus linguistics and the studies in this field in terms of causality and provability (McEnery et al., 2006: 4); In this period, the field of corpus linguistics paused in terms of development. With the development of computer technology, the number of researches increased in parallel with the use of computers in the field of corpus linguistics and the size of data sets.

Corpus Types

Corpuses, which have a wide range according to the purpose of use, are examined under different headings in terms of types and fields of study. The research question that the body focuses on directly shapes the purpose of the corpus and the data set. The most common corpus types can be expressed as general, special, marked, historical, marked, unmarked, student corpus. General Corpus: General corpus is the most commonly used corpus type in corpus linguistics. The data set of these corpuscles has a wide range. It includes films, literary texts, speech texts, newspaper articles, etc. can be found. Special Corpus: Special corpus, unlike general corpus, has narrower datasets and is focused on a particular field of study or discipline. Parallel corpora are designed based on the translation relationship between texts, so they have the original set of texts and their translations (Tognini-Bonelli & Sinclair 2006). Claridge (2008: 242) defines historical corpus as a corpus created to represent the past phases of a language and/or to study language change. One of the important topics in terms of corpus linguistics is the labeling of the words in the data set. In tagged corpus, word types and syntax features in the texts that make up the corpus were determined.

Result

With the rapid development of information technologies in recent years, software that allows independent researchers to create a corpus has emerged. WordsmithTools, one of the internationally prominent software, is a software that can be used under license and has a lot of analysis modules. AntConc software is one of the most frequently used software by independent researchers and it is free. In both software, context, collocation and frequency studies can be carried out. However, statistical studies can also be done with WordsmithTools. TextStat software, unlike other software, can perform analyzes by including different types of files in the corpus database. It also has the ability to create a web-based database. There are many in-depth resources on corpus linguistics and corpus headings in the international sense. It is foreseen that this study, which was shaped to provide a general idea on the corpus, will make a theoretical contribution to the field; It is aimed that researchers who want to carry out studies in this field are aware of the prominent language,

dictionaries and word types about the corpus, at least on the scale of Turkey. It is certain that much deeper research can be carried out on every concept mentioned in the study, both in theory and in practice.

Keywords: Corpus, Corpus linguistics, technology, text, Corpus types.

Öz: Dil ve dile yönelik çalışmalar yüzyıllardan beri devam etmekte olup her araştırmacı tarihsel süreçteki bilgi birikimi ile yeni çıkarımlarda bulunmaktadır. Teknoloji ve bilişimdeki gelişim bilgiyi ve bilgiyle ilişkili olan her şeyi etkilemektedir. Dil ve dil pratiklerini inceleme, betimleme, analiz etmede kullanılan derlem dilbilim alanı da bu gelişimden etkilenmiş; 1970'lere kadar kurumsal yapıların bünyesinde yapılan derlem araştırmaları takip eden yıllarda geliştirilen araçlarla, yazılımlarla bağımsız araştırmacılar tarafından da yürütülmeye başlanmıştır. Süreçte derlemlerin veri setleri de dönüşerek milyon kelimelik yapılara bürünmüştür. Tüm bu gelişmeler derlem ve derlem dilbilimin temel ilkelerini ve araştırma alanlarını belirginleştirmiştir. Bağımsız araştırmacılar tarafından da yürütülmeye başlanan derlem çalışmaları ile derlem tasarımına yönelik temsil edicilik, denge, örnekçe, boyut, makine tarafından okunabilirlik, telif hakları gibi ilkelerin öne çıktığı saptanmıştır. Derlemler oluşturma amaçlarına göre genel, özel, tarihsel, etiketlenmiş, etiketlenmemiş, öğrenci derlemleri gibi türlere de ayrılmıştır. Bu araştırmada da Türkiye'de son yıllarda yapılan dil ve dil eğitimi ile alakalı derlem çalışmaları üzerinde durulmuş; proje tabanlı yapılan çalışmaların alana katkıları burgulanmıştır. Derlemler kapsamında dil eğitimi alanların ve araştırmacıların kullandığı öğrenme yöntemleri değerlendirilmiş; bu yöntemlerin sürece kakıları, özerk öğrenme ya da öğrenmeyi öğrenme yaklaşımına etkileri; kelime öğretimi, deyim öğretimi; dil eğitiminde tüme varım ilkesinini işletilmesi gibi alt başlıklar değerlendirilmiştir. Bu çerçevede derlem ve derlem dilbilime yönelik araştırmacılarda kuramsal bir alt yapının oluşumu hedeflenmiştir.

Anahtar Kelimeler: Derlem, Derlem dilbilim, teknoloji, metin, Derlem türleri

Giriş

Yazılı tarih öncesi dönemlerden günümüze kadar siyaset, diplomasi, ticaret, eğitim, kültür vb. toplumla ilgili kavramların odağında dil bulunmaktadır. İletişim ve bildirişim fonksiyonunun yanında kitlesel, kültürel ve siyasal değişimler öncelikle kendini dilde göstermiştir. Bu çerçevede tüm dünyayı etkileyen Rönesans, Reform ve ekonomi dizgesinde gerçekleşen dönüşümden dil de doğal olarak etkilenmiş; filoloji çalışmaları arşivler üzerinden dijital ortama doğru kaymaya başlamıştır. Bu köklü değişimde "teknik" kavramının uygulamaya yansımasındaki farklılıklar etkili olmuştur.

Dünya'yı baştan sona değiştiren teknik kavramı ve teknik kavramının bilgiyle sistematize hali teknoloji; eğitimi, toplumu, ekonomiyi, dili ve araştırma sahalarını da etkilemiştir. Bilgiyi sistematize etmede çağın en önemli aracı olarak öne çıkan bilgisayarlar ile akademik araştırmalar, çıkarımlar ve yorumlama faaliyetleri de dönüşüm geçirmştir. Dil, iletşim alanları da bu dönüşümden doğrudan etkilenmiş; "derlem dilbilim" olarak ifade edilen bilgisayar ve dilbilim kavramlarını bir araya getiren araştırma alanı ortaya çıkmıştır.

Derlem dilbilim, tarihsel olarak semavi dinlere mensup din adamlarının kutsal kitaplardaki kavram, kelime, terim, konuların geçtiği sayfa bilgisini belirlemeye yönelik tasnif çalışmaları ile 12. yüzyıla kadar uzanmaktadır. Türkiye'de ise Türkçe ve yabancı dillerdeki söz varlığı çalışmalarının ülkemizdeki derlem dilbilim araştırmalarının zeminini oluşturduğu söylenebilir.

Latince "corp" yani vücut/ceset kelimesinin Latince çoğul ekleriyle oluşturulan "corpora" terimi Türkçeye "derlem" olarak çevrilmiştir. McEnery ve Hardie (2011: 32) derlemi "Ele alınan dil yapısını en üst düzeyde temsil yeteneğine sahip, makine tarafından okunabilen ve belli sınırları olan metin bütünü" şeklinde tanımlamışlardır. Gries'a (2009: 7) göre derlem; iletişim/bildirişim ortamında makine tarafından işlenebilen, yazılı ya da sözlü verilerin bütünü olarak tanımlamıştır. Sinclair'e (2005) göre derlem, dilbilim araştırmalarında kullanılmak üzere bir dili veya o dilin farklı

kullanımlarını mümkün olduğunca doğru ve kapsamlı yansıtabilecek, belirli ölçütlere göre seçilmiş metinlerin elektronik ortamda bir araya toplanmasıdır. Güncel Türkçe Sözlük'te ise "derlem" sözcüğü Fransızca kökenli bir sözcük olan "koleksiyon" olarak tanımlanmış; dilbilimsel karşılık olarak ise "Bir dilin türlü kullanım alanlarından derlenmiş örneklerinin dil bilgisi ve kuramsal dil bilimi araştırmalarında kullanılmak üzere bilgisayar tarafından okunabilecek biçimde bir araya getirilmiş kümesi" ifadesi tercih edilmiştir. Ancak alanyazında "bütünce" kavramı da derlem yerine kullanılmaktadır. Genel çerçevede derlemi, dijital ortama aktarılarak üzerinde işlemlerin yapılabildiği özel ya da genel bir amaçla bir araya getirilen yazılı, sözlü metinlerden oluşan veri seti olarak tanımlayabiliriz.

Günümüz araştırmaları açısından derlem sürecinin teknik ve dil boyutu öne çıkmaktadır. Teknik bağlamda bilgisayar; dil bağlamında ise iletişim, bildirişim esnasındaki veriler derlemin iki ana çıkış noktasını teşkil etmektedir. Bu çerçevede araştırmacı tarafından bir amaç etrafında şekillendirilen veri setinde istatistik, sıklık, kümelenme, eşdizimlilik çalışmaları gerçekleştirebilir. Ancak derlem çalışmaları bu çalışma türlerinin çok ötesinde bir kapsama sahiptir. Bu bağlamda Sinclair'e (2005) göre derlem; sıklık listesi, sezgisel olarak bir araya getirilmiş metin yığınları ya da internet ağından çekilmiş veri yığınlarından ibaret değildir. Daha öncede ifade edildiği gibi derlemi oluşturmada bir amaç, kapsamı belirleyen bir ölçüt, verinin işleneceği bir veritabanı ya da yazılım gerekmektedir.

Derlem Dilbilim

Derlemdilbilim, 20. yüzyılın ikinci yarısında bilgisayar teknolojisinin gelişimi ile dilbilimin alt disiplini olarak şekillenmeye başlamıştır. Etimoloji, pragmatik, semantik gibi araştırma alanlarına da hizmet eden derlem dilbilim, kullanımda olan yazılı sözlü metinlerin sistematik olarak toplanmasına odaklanmıştır. Ancak derlem dilbilim alanındaki araştırmacılar dilbilim- derlem dilbilim arasındaki ilişki konusunda farklı yaklaşımlara sahiptir. Bu çerçevede derlem uygulamalarının dilbilim alanında kullanılan bir metodoloji ya da alt disiplin dalı şeklinde değerlendirmeler mevcuttur (McEnery & Hardie, 2011: 7). Son yıllarda derlem dilbilim özelinde yapılan çalışmalarda "derlem tabanlı yaklaşım" metodoloji olarak kabul görmeye başlamış; yapılan çalışmaların sayısı da artmıştır. Ancak literatür anlamında derlem dilbilim, dilbilimin bir alt dalı olarak değerlendirilmektedir.

Özkan (2010), bilgisayar aracılığıyla dile ait olan verilerin işlenmesinin yöntemsel olarak ortaya çıkardığı olumlu sonuçların dilbilimin çalışma alanlarında belirgin olarak görüldüğünü; zamanla yaygınlaşma alanını genişleten bilgisayarlı dil çalışmalarının derlem dilbilim, sözlükbilim, biçimbilim, sesbilim, makineli çeviri, anlambilim vb. dilbilim alanlarında hissettirdiğini ifade etmiştir. Tarihsel açıdan ilk derlem tabanlı çalışmaların ses bilimsel ayırt edici özellikler üzerine kurulan küçük boyutlu derlemler olduğu; genel dil bilgisine yönelik derlem tabanlı çalışmalar 1960'lardan sonra yapıldığı bilinmektedir. 1950'li yıllardan sonra bilgisayar ve elektronik ile ilgili gelişmeler elektronik derlem oluşturmayı mümkün kılmış; araştırmacılar da bu alana yönelmiştir. İngilizce üzerine yapılandırılan, 1 milyon kelimelik veri setine sahip *The Brown Corpus* elektronik ortama aktarılan ilk derlem olarak bilinmektedir (Francis & Kucera, 1964). Ses kayıt cihazlarının aynı yıllarda piyasaya çıkışıyla sözlü derlemler üzerine ilk çalışmalar gerçekleştirilmiştir. Sinclair, 1963-1965 yılları arasında yaklaşık 170 bin kelimeden oluşan İngiliz sözlü derlemini oluşturmuştur (Krishnamurthy, 2004). Derlem dilbilim alanında yapılan bu ilk çalışmaların ardından İngilizce ile ilgili sözlük çalışmaları devam etmiştir. Ancak derlem dilbilimin disiplin haline dönüşmeye başladığı ilk dönemlerde, yani 1960'lı yıllarda Chomsky ve ekolünü takip eden dilbilimciler derlem dilbilim alanını ve bu alanda yapılan çalışmaları nedensellik, kanıtlanabilirlik açısından eleştirmiş (McEnery vd., 2006: 4); bu dönemde derlem dilbilim alanı gelişim açısından duraklamıştır. Ancak ilerleyen yıllarda bilgisayar teknolojisinin gelişimi ile derlem dilbilim alanında bilgisayar kullanımı ve veri setlerinin büyüklüğüne paralel olarak araştırma sayıları da artmıştır.

Derlem dilbilim açısından 1970'li yıllar alanın kendine has çalışma şekli, veri tabanı oluşturma kriterleri, derlem türleri arasındaki farklılıkların keskinleştiği yıllar olarak bilinmektedir. Teknik imkânların henüz istenilen düzeye gelmediği bu yıllarda 1 milyon kelimelik derlemler oluşturulmuştur. *Lancaster-Oslo/Bergen* derlemi, *Brown* derlemi model alınarak 1970'li yıllarda yapılandırılmıştır. Veri seti açısından bu derlem *Brown Derlemine* göre daha dengeli bir veri setine sahiptir (Kenneddy, 1998: 26). *Collins Cobuild English Language Dictionary* 10 yıllık tasarım sürecinin ardından 1987 yılında yayınlanmıştır. Bu çalışma dil eğitimi bakımından derlem temelli ilk sözlük olarak da literatüre geçmiştir(Teubert & Čermáková, 2004: 111).

1990'lı yıllarda derlem dilbilim teknik anlamda da veri seti anlamında da çok büyük ilerlemeler kaydetmiş; derlem çalışmaları Batı dillerinin büyük bölümünde yapılmaya başlanmıştır. 1990 yılı Ocak ayında, 365 milyon sözcükten oluşan tek dilli derlemler 16 Avrupa dili için oluşturulmuştur (Leech, 1991: 21). Derlem dilbilim araştırmalarında adından sıkça söz edilen *British National Corpus (BNC),* 1991-1995 yılları arasında ve İngiliz hükümetinin de proje maliyetinin yarısını desteklemesiyle tamamlanmıştır. 100 milyon sözcük içeren BNC, 1960-1975 arası ölçünlü İngiliz İngilizcesini temel almakta, %90'ı yazılı ve %10'u sözlü dilden derlenen malzemeyi içermektedir. Veri setinin %60'ı kitaplardan, %25'i dergilerden, %5'i broşürlerden ve %5'i de yayımlanmamış mektuplardan oluşturulmuştur (akt. Tahiroğlu, 2010; Kennedy, 1998). Takip eden süreçte derlem dilbilim disiplin açısından metodolojik ve ilkesel tartışmaları aşmış; kendine has bir araştırma haline dönüşmüştür. Bu bağlamda istatistiki alt yapıya yönelik MI, MI3, Loglikelihood, Logdice gibi parametrelerle analizler de geçerlik güvenirlik düzeylerini arttırmış; derlem dilbilim odaklı bağımsız yazılımlar tasarlanmıştır.

Derlem Tasarımı

Derlem dilbilim, dilbilimden ayrı bir alan, metodoloji olarak değerlendirilmesi üzerine yapılan tartışmalar sonrasında alanın sınırları netleşmiştir. Bu kapsamda veri seti oluşturmadan araştırma şekillerine kadar benimsenenen temel ilkeler bulunmaktadır. Yapılan çalışmalar bu ilkeler etrafında şekillendirilmekte olup dilbilimin diğer alt dallarında yapılan çalışma usul ve yaklaşımlarından farklılıklar görülmektedir. Conrad'ın (2002: 77) ifade ettiği gibi bu ilkeler derlem dilbilimin süreç içindeki gelişimi ile belirlenmiştir. Derlemler genel olarak bu ilkeler etrafında şekillendirilmiştir. Sinclair (2005) derlem tasarımına ilişkin temel ilkeleri şu şekildedir:

1. Bir derlemin içeriği içinde doğdukları topluluktaki iletişimsel işlevlerine göre seçilmelidir.
2. Derlem araştırmacıları derlemi oluşturdukları dil açısından temsil edicilik düzeyini yüksek tutmalı; veri tabanında gerekli çeşitliliği sağlamalıdır.
3. Derlem yapısını belirleme kriterleri sayıca az olmalı, birbirinden açıkça ayrı olmalı ve incelenen dili veya çeşitliliği temsil eden bir bütünceyi betimlemede bir grup olarak etkili olmalıdır.
4. Bir metinle ilgili kelimelerinin alfanümerik dizisi ve noktalama işaretleri dışındaki her türlü bilgi, düz metinden ayrı olarak saklanmalı ve uygulamalarda gerektiğinde birleştirilmelidir.
5. Bir derlem için dil örneklemi, belgelerden veya tam konuşma olaylarının transkripsiyonlarından oluşmalı veya bu hedefe mümkün olduğunca yakın olmalıdır.
6. Derlem araştırmacısı, veri setinde denge ve bütünü temsil etme ilkesini devamlı gözetmeli; mümkün olduğunca veri setine bu öncelikler uygulanmalıdır.
7. Bir derlemdeki konunun herhangi bir kontrolü, dâhili değil, harici kriterlerin kullanılmasıyla uygulanmalıdır.
8. Bir derlem, kapsamı öncelerken veri setinin içeriğinde homojenliği hedeflemeli; derlemin temsil gücünü azaltan, hipotezin dışında kalan verileri içermemelidir.

Derlem tasarımı ile ilgili temel ilkeler, araştırmacıların bulgularına yönelik geçerlik ve güvenirlik açısından dikkat edilmesi gereken hususlardır. Derlem ile ilgili ilkelerden sonra bu ilkelerin kavramsallaştırılması üzerinde durulacaktır. Bu kavramlar; temsil edicilik, denge, örnekçe, boyut, makine tarafından okunabilirlik, telif hakları gibi başlıklardan oluşmaktadır.

a) Temsil Edicilik (Representativeness): Derlem araştırmacıları derlemi oluştururken yapılandırdıkları problem cümleleri oldukça önemlidir. Derlemi oluşturmadaki amaç, kapsam temsil ediciliği doğrudan etkilemektedir. Belirli bir alana yönelik özel bir derlem veya geneli ilgilendiren bir genel derlem tasarımı "temsil edicilik" kavramının ele alınış şeklini değiştirir. Leech'e (1991:25-30) göre derlemi oluşturmak için ele alınan dil kesiti/değişkesi, araştırılan konu ile ilgili genellenebilir çıkarımlarda bulunabilmeyi sağlamalıdır. Genellenebilir çıkarımlarda bulunmayı sağlayan veri çeşitliliği temsil edicilik olarak ifade edilmektedir. McEnery ve arkadaşları (2006) derlemin veri seti açısından yeteri kadar çeşitliliğe sahip olmasının çıkarımları ve genellemeleri daha güvenilir kıldığını belirtmiştir.

b) Denge (Balance): Temsil edicilik gibi derlemde denge ilkesi de hayati bir öneme sahiptir. Araştırmacı, oluşturacağı derlem türüne göre denge ilkesini gözetir. Genel amaçlı derlemlerde dil kesitini doğru örnekleme adına farklı türlerden verilerin derleme taşınması gerekmektedir. Bu kapsamda sadece gazete yazıları, sadece siyaset haberleri ya da sadece roman türündeki metinlerden oluşan bir derlem denge ilkesini ihlâl eder. Dolayısıyla derlemi oluşturan veri setindeki türler ve dağılımlarında bir tür lehine ya da aleyhine ağırlık olmamalıdır. Ancak özel derlemlerde odaklanılan türe yönelik veri seti bir tür üzerinde yoğunlaşabilir.

c) Örneklem/örnekçe (Sample): Derlem oluşturma sürecinde en temel problemlerinden biri de örneklemdir. Clear'a (1992) göre örneklem oluşturulurken derlemin kapsamı ile ilgili sınırların belirlenmesi; genellemelere varmak için örnekçenin sınırlarının belirlenmesi önemli bir problemdir. Çünkü "dil" olarak ifade ettiğimiz sınırsız yapıyı tamamıyla veri setine dâhil etmek mümkün değildir. Dolayısıyla örnekçeler, büyük veri setlerini temsil edebilen, dengeli içerikler olmak durumundadır. Örneğin hukuk makaleleri üzerine inşa edilen özel bir derlemde tüm makaleler derlem dilbilim tabiri ile "popülasyon" , derlemin veri setini oluşturulması için belirli sayıda seçilen hukuk dergileri ise örnekçeyi oluşturmaktadır.

d) Boyut (Size): Derlem oluşturma sürecinde en çok tartışılan unsurlar arasında yer alan kavram, boyuttur. İlk derlemlerin veri seti 2.000 kelimeden 20.000 kelimeye kadar değişmekteydi. Oostdijk (1991) 20.000 kelimelik bir örneklem boyutunun, belirli bir çeşidi temsil edecek kadar büyük olduğunu belirtmiştir (Oostdijk 1991: 50). Ancak British National Corpus (BNC) gibi projelerde 40.000 kelimelik örneklemler tercih edildi. Derlem dilbilim açısından ilerleyen süreçte araştırmacının araştırma yaptığı alana, derlemi oluşturduğu alana göre derlem boyutunun değişkenlik gösterdiği bilinmektedir. Dolayısıyla standart bir boyuttan bahsetmek mümkün değildir.

e) Makine Tarafından Okunabilirlik (machine-readibility): Son yıllarda bilgisayar teknolojisi ve bilişimin gelişmesi ile derlemlerin boyutu kapsamı artmıştır. Bu çerçevede hazırlanan derlemlerin veri setleri yazılımlar ve bilgisayarlar tarafından okunabilir niteliğe sahip olmak durumundadır. Böylece milyonlarca kelimelik veri setlerinde kısa sürede çok kapsamlı taramalar gerçekleştirilebilmektedir.

f) Telif: Derlem için oluşturulan veri setinde kullanılacak yazılı sözlü metinler, içeriklerine ve kapsamına göre araştırma sürecinde kullanılmadan izin alınması gerekebilir. İzin alınarak kullanılabilecek eserler nedeni ile araştırmacı telif hakkı ile ilgili hukuki problemler yaşayabilir. Bu bağlamda veri setini oluştururken örnekleme dâhil edilecek metinlerin niteliğine dikkat edilmelidir.

Derlem Türleri

Kullanım amacına göre geniş yelpazeye sahip olan derlemler, tür ve çalışma alanları açısından da farklı başlıklar altında incelenmektedir. Derlemin odaklandığı araştırma sorusu, amaç derlem türünü ve veri setini doğrudan şekillendirmektedir. En sık karşılaşılan derlem türleri genel, özel, tarihsel, etiketlenmiş, etiketlenmemiş, öğrenci derlemleri şeklinde ifâde edilebilir.

a) **Genel Derlemler:** Genel derlemler, derlem dilbilim alanında en sık kullanılan derlem türüdür. Bu derlemlerin veri seti oldukça geniş bir yelpazeye sahiptir. İçeriğinde filmler, edebi metinler, konuşma metinleri, gazete yazıları vb. bulunabilir. Genel derlemler veya referans derlemler sözlü, yazılı veri setinden oluşabilir; bir bütün olarak dil hakkında izlenim edinmeyi amaçlar. Genel derlemler milyonlarca kelimeden oluşan derlemler olup araştırmalarda referans derlemler olarak kullanılır (Kübler ve Aston 2010: 504; Baker, 2010: 12). 50 Milyon kelimeden oluşan *Türk Ulusal Derlemi*, 700 Milyon kelimeden oluşan Türkiye merkezli *TS CORPUS* derlemi, 100 milyon kelimeden oluşan *İngiliz Ulusal Derlemi (BNC)* genel derlemlere örnek olarak verilebilir.

b) **Özel Derlemler**: Özel derlemler, genel derlemlerin aksine daha dar kapsamlı veri setlerine sahip olup belirli bir çalışma alanı ya da disipline odaklıdır. Özel derlemler, daha küçük veri setleri ile çalışmaları nedeniyle eleştirilmiştir (Lee 2010: 114). Ancak araştırmalar, yüksek sıklık değerine sahip ögeleri araştırırken de güvenilir sonuçlar verebildiklerini ve bir derlemin her zaman milyonlarca kelimeden, çok sayıda metinden oluşması gerekmediğini göstermiştir (Biber, 1990). Örneğin hukuk, tarih, biyoloji gibi alanlarda yazılan makalelerden oluşturulan bir derlem özel derlem başlığı altında incelenir. *Profesyonel Sözlü Amerikan İngilizcesi Derlemi (Corpus of Professional Spoken American English)* örnek olarak verilebilir.

c) **Paralel Derlemler**: Tek dilli derlem bir dil içerirken, çok dilli derlem iki veya daha fazla dil içeren veri setine sahiptir ve iki dilli derlemler iki kategoride incelenir: paralel ve karşılaştırılabilir derlemler. Paralel derlemler, metinler arasındaki çeviri ilişkisine dayalı olarak tasarlanır, böylece orijinal metin grubuna ve bu metinlerin çevirilerine sahip olur (Tognini-Bonelli ve Sinclair 2006). Karşılaştırılabilir bir bütünce, metinlerin çevirilerini içermez; birkaç dilde toplanan ve aynı iletişimsel işleve (Kenning 2010) dayanan metinleri içerir. İlk paralel derlem, İngilizce ve Kanada Fransızcası hükümet belgelerinden oluşan *Kanada Hansard Corpustur* (Tognini-Bonelli ve Sinclair 2006). En ünlü paralel derlemlerden biri, her iki dilde de orijinal ve tercüme edilmiş metinler içeren 2,6 milyon kelimelik *İngilizce-Norveç Parallel Corpus'tur (ENPC*). *Oslo Çok Dilli Derlemi (OMC), ENPC*'nin İngilizce, Norveççe, Almanca, Fince, İsveççe, Felemenkçe, Fransızca ve Portekizce metinleri içeren bir uzantısıdır.

d) **Tarihsel Derlem:** Claridge (2008: 242) tarihsel derlemi, bir dilin geçmiş aşamalarını temsil etmek ve/veya dil değişimini incelemek için oluşturulmuş bir derlem olarak tanımlar. Alan yazın içinde artsüremli ve eşsüremli derlem olarak iki başlıkta incelenen bu derlem türü örneklemi belirli bir kesitte inceler. Artsüremli derlem tarihsel bir noktanda günümüze ilgili dilin gelişimi, dil pratiklerini inceler. Eşsüremli derlemde ise başlangıcı ve bitişi belli olan 2 nokta arasındaki dil gelişimi ve dil pratiklerinin örnekleme dolayısıyla derleme yansımaları incelenir. Özellikle İngilizce üzerine oluşturulan artsüremli derlemler derlem dilbilim alanında öne çıkmaktadır. Artsüremli derlem dilin zaman içindeki evrimini içerir. 1923'ten günümüze veri toplayan *Time* gazetesi derlemi artsüremli bir derlem olarak kabul edilebilir (Gatto, 2014: 15).

e) **Çok Yönlü (Multimodal) Derlemler:** Çok yönlü bir derlem, orijinal ses veya görsel kayıtlarla bağlantılandırılmış veya senkronize edilmiş transkriptleri içeren derlemlerdir (Lee 2010: 114). Bu tür derlem hem metinsel hem de metinsel olmayan verileri içerir. Veri setinde metinler, videolar, konuşma kayıtları yer alabilmektedir. Akademik

bağlamda tek konuşmacı ve ikili konuşmalardan toplanan kayıtları ve yazıları içeren 250.000 kelimelik *Nottingham Multimodal Corpus (NMMC)* çok yönlü (multimodal) derlemdir.

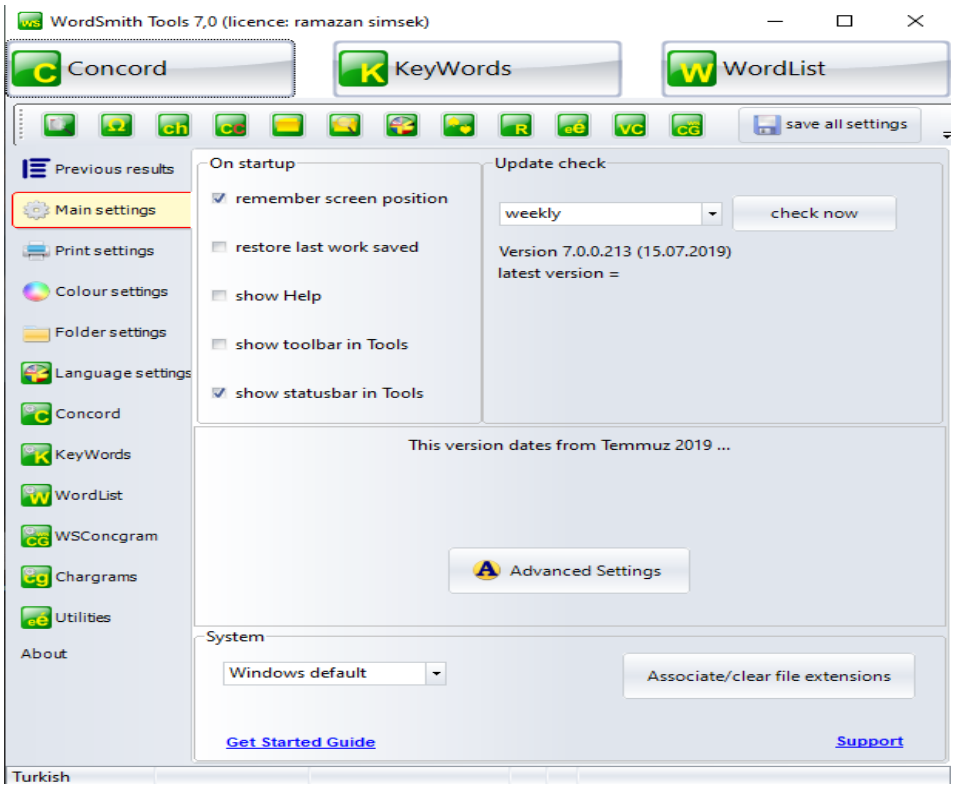
f) **Etiketlenmiş (İşaretlenmiş) ve Etiketlenmemiş (İşaretlenmemiş) Derlemler:** Derlem dilbilim açısından önemli konu başlıklarından biri de veri setinde yer alan kelimelerin etiketlenmesidir. Etiketlenmiş derlemlerde, derlemi oluşturan metinlerdeki sözcük türleri, sözdizim özellikleri belirlenmiştir. Etiketlenmemiş derlemlerde ise veri setini oluşturan metinler kelime, sözdizimi gibi faktörler açısından düzenlenmemiştir.

g) **Öğrenci Derlemleri:** Derlem dilbilim alnında en sık rastlanılan derlem türlerinden biridir. Yabancı dil eğitimi alan bireylerin oluşturdukları metinlerden oluşan öğrenci derlemleri dil öğrenmeye başlayan bireyler açısından rehber konumundadır. Öncelikli hedef kelime listeleri bu derlemler aracılığı ile tespit edilerek öğrencilere sunulabilir (Şimşek, 2021).

Derlem Araçları

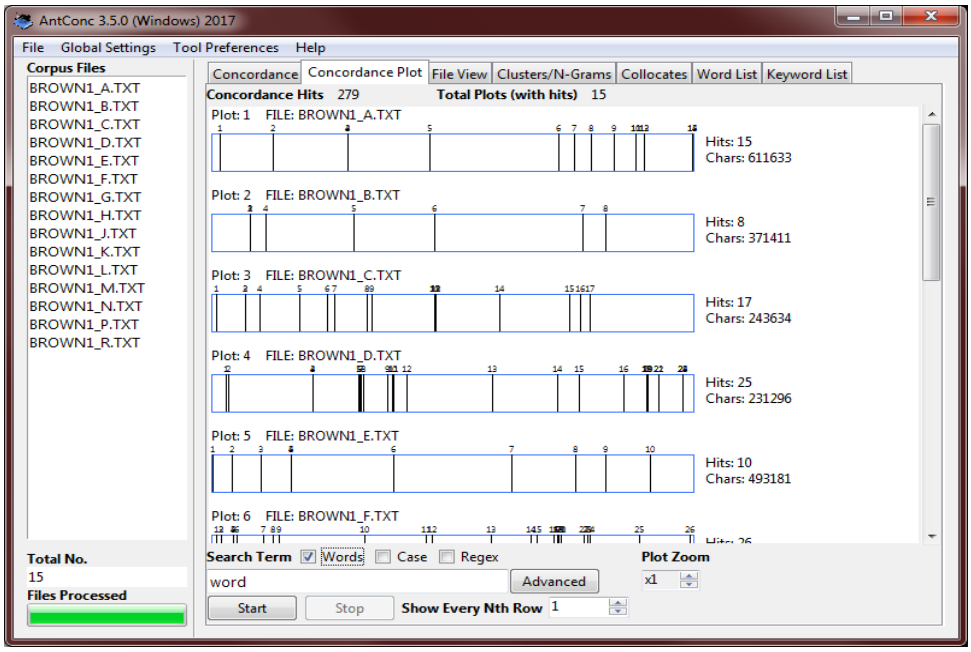
Derlem dilbilim araştırmalarına 1970'li yıllardan itibaren ilgi giderek artmış; derlemlerin veri setleri de milyon kelimeleri aşmaya başlamıştır. 1980'li yıllarda bilgisayar ve elektronik alanlarındaki gelişmeler derlem dilbilim araştırmalarına bilgisayarı da dâhil etmiş; ilerleyen yıllarda ise bağımsız araştırmacıların çalışmaları için yazılımlar tasarlanmıştır. Bu bağlamda serbest ve kurumsal araştırmalar için AntConc, Wordsmith Tools, Textstat yazılımları en bilinen yazılımlardır.

a) **WordSmith Tools:** WordSmith Tools derlem dilbilim alanında en sık kullanılan yazılımlardan biridir. Lisanslı olarak kullanabilen bu yazılımın günümüze kadar 8 farklı versiyonu bulunurken bünyesinde entegre derlemleri de barındırmaktadır. Kod dönüşümü, referans derlem, işitsel-görsel niteliğe sahip içerikleri veri tabanına aktarma, derlem tabanlı istatistik verileri oluşturma gibi özelliklere sahip olan WordSmith Tools çok yönlü bir yazılımdır. Kendi içinde 3 ana bölümden oluşmaktadır. *Wordlist* modülü ile araştırmacının oluşturduğu derlemdeki kelimelerin sıklık değerleri; *Concordance* modülü ile hedef kelimenin bağlam temelli dizinleri betimlenmektedir. *Keywords* modülü ile referans derlem üzerinden araştırmacının hazırlamış olduğu derlem; veri tabanının tür özelliklerine yönelik çıkarımlar yapmak mümkündür.



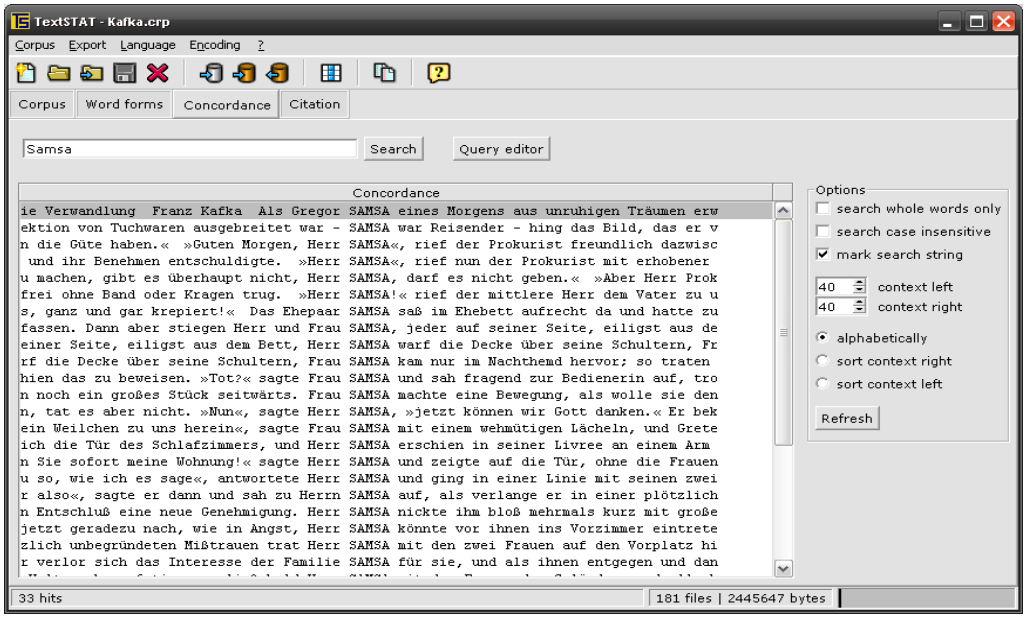
Görsel 1. WordSmith Tools Yazılımına Ait Program Ara Yüzü

b) **AntConc:** Bağımsız araştırmacılar tarafından en sık kullanılan yazılımlardan biri de AntConc'tur. Ücretsiz bir yazılım olması ve programın düşük konfigürasyonlu sistemlerde dahi rahat çalışabilmesi tercih edilmesinde önemli etkenlerdir. Yazılım aracılığı ile veri tabanında sıklık, bağlam, anahtar kelime (freq, concordance, keywords) gibi sorgulamalar yapılabilmektedir



Görsel 2. WordSmith Tools Yazılımına Ait Program Ara Yüzü

c) **TextStat:** TextStat yazılımı diğer yazılımlardan farklı olarak farklı türdeki dosyaları derlem veri tabanına dâhil ederek analizler gerçekleştirebilmektedir. Ayrıca web tabanlı veri tabanı oluşturabilme imkânına da sahiptir. Diğer yazılımlarda olduğu gibi veri tabanında sıklık, bağlam, anahtar kelime (fren, concordance, keywords) gibi sorgulamalar yapılabilmektedir



Görsel 3. TextStat Yazılımına Ait Program Arayüzü

Derlem yazılımlarında bulunan modüllerde ve analiz araçlarında sabit hale gelmiş terimler bulunmaktadır. Derlem araçları ile ilgilenirken modülleri, fonksiyonları anlamlandırma adına bu modül ve analiz araçlarındaki terimlerin işlevleri aşağıdaki gibi tarif edilebilir:

- **Sıklık Listeleri (Frequency Lists):** Bir derlemde sözcüklerin veri setindeki görünüm değeridir
- **Toplam Kelime Sayısı/Örnekçeler (tokens):** Derlemi oluşturan veri setindeki toplam sözcük sayısıdır.
- **Farklı Kelime Sayısı/Örnekler (types):** Derlemi oluşturan veri setinde birbirinden farklı sözcüklerin sayısıdır.
- **Ortalama Sözcük ve Tümce Uzunluğu (average word and sentence length):** Derlemi oluşturan veri setindeki kelimelerin ortalama hece sayısı ve cümlelerin ortalama kelime sayısı.
- **Bağlamlı dizinler (Concordances):** Derlemi oluşturan veri setindeki bir ifadenin, geçtiği her durum ve içinde bulunduğu bağlamların listesidir.
- **Eşdizimler (Collocations):** Derlemi oluşturan veri setindeki herhangi bir kelimenin birlikte kullanımları, yaygın olan sözcük kümeleridir.

Türkiye'de Derlem Çalışmaları

Türkçede derlem çalışmaları son 10 yılda büyük ivme kazanmış; bağımsız ve kurumsal anlamda birçok çalışma gerçekleştirilmiştir. Türkçenin bilgisayarla işlenmesi üzerine ilk çalışma 1976'da Aydın Köksal'ın otomatik biçim birimsel çözümleme konusunda hazırladığı "Automatic Morphological Analysis of Turkish" adlı doktora tezidir. 1981'de Sagay'ın İngilizce-Türkçe arasında, 1987'de de Stoop'un Almanca ve Türkçe arasında otomatik çeviri sistemi hazırlama çalışmaları ilk çalışmalar olarak görülmektedir. 1990'lı yıllarla birlikte bilgisayar destekli dil bilimi çalışmalarının tüm dünyada internetin gelişmesiyle Türkçe bilgisayar destekli çalışmalarının da proje düzeyinde yapılmaya başlandığı görülür (Tahiroğlu, 2010). Bilgisayar altyapısı kullanılarak Türkiye'de yapılan ilk derlem Orta Doğu Teknik Üniversitesi (ODTÜ) bünyesinde 1999-2003 yılları arasında tamamlanan ODTÜ Türkçe Derlemidir. Yazılı kaynakların veri setini oluşturduğu bu derlem genel derlem türünde tasarlanmıştır. Derleme internet ağından erişim imkânı bulunmamaktadır.

Akademik araştırmalar kapsamında son yıllarda en sık kullanılan derlem 50 milyon kelimelik Türkçe Ulusal Derlemi'dir (TUD). Derlemin tasarımcıları tarafından günümüz Türkçesinin çok sayıda farklı alan ve türlerden yazılı ve sözlü örneklerini içeren, geniş kapsamlı, dengeli ve temsil yeterliliğine sahip, genel amaçlı bir referans derlem olarak tanımlanmıştır. Ayrıca TUD temel olarak tek sözcük ya da sözcük grubunun bağlam içinde anahtar sözcük arama işlevine sahiptir. Bununla birlikte, araştırmacılar derlem ana sayfasında bulunan çeşitli dil dışı ölçütlerle (yayın yılı, alan, türev metin biçimi, vb.) araştırma sorularına uygun olarak sorgularını daraltabilir ve bu doğrultuda bağımlı dizin sonuçlarına ulaşabilmektedirler. Derlem hazırlanırken İngiliz Ulusal Derlemi model olarak alınmıştır, bununla birlikte Türkçenin yapısına uygun birtakım değişiklikler yapıldığı ifade edilmiştir. (Aksan vd., 2016).

Veri setinin genişliği ve çevrimiçi erişim imkânı ile öne çıkan bir diğer derlem TS Corpus Derlemidir. Bağımsız araştırmacılar tarafından oluşturulan bu derlem veri seti bakımından en kapsamlı derlem olarak dikkat çekmektedir. Yaklaşık 500 milyon kelimelik bir veri setinden oluşan derlem Taner Sezer ve Türker Sezer tarafından 2011 yılında oluşturulmuştur. BNC ve COCA gibi büyük veriye sahip ve modern derlemlerle ortak özellikleri sunmaktadır. Derlem diğer derlemlerden farklı olarak alt derlemleri de erişime açık durumdadır (Sezer & Sezer, 2013).

Derlem dilbilim üzerine TÜBİTAK destekli araştırma projeleri de yürütülmüştür. Mersin Üniversitesi bünyesinde Prof. Dr. Bülent Özkan ve arkadaşları (2014) tarafından yürütülen ve

tamamlanan projeler şunlardır: "Türkiye Türkçesinin Yazın Dili Derlemi", "Türkiye Türkçesinde Fiilimsiler", "Türkçe için 'Kendi Kendine Derlem' Platformu Oluşturma", Bir Özel Alan Derlemi Olarak ”Türk Çocuk Yazını Derlemi” ve “Türk Çocuk Yazını Üzerine İleri Uygulamalar Projesi”, "Türkiye Türkçesinde Fiillerin Derlem Denetimi ve Derlem Tabanlı Sözlüğü", "Derlem Tabanlı Çevrimiçi Çok dilli Türkçe Öğrenici Sözlüğü"

Derlem dilbilim alanında çalışan bağımsız araştırmacıların filoloji ve edebiyat ağırlıklı çalışmaları da son yıllarda giderek artmaktadır. Tez boyutunda yapılan çalışmalarda Tahiroğlu (2010) sözlük çalışmaları, Gündoğdu(2012) çocuk edebiyatı, Tüfekçioğlu (2013) sözlükte A madde başı, Kurtoğlu (2015) isim fiillerin eşdizimliliği, Sert (2018) eş anlamlı sıfat durumları, Tüfekçioğlu (2018) sosyal bilimlerde akademik ve teknik söz varlığı, Halat (2020) görsel Türkçe sözlük kullanımı, Şimşek (2020) derlem tabanlı eleştirel söylem analizi üzerine araştırmalar gerçekleştirmiştir.

Türkçe üzerinde derlem dilbilim odaklı ya da derlem tabanlı yapılan makalelerin de sayısı artmaktadır. Bu çalışmalardan öne çıkanları şu şekildedir: Özkan (2013), Türkiye Türkçesi söz varlığı üzerine derlem tabanlı çalışmalar gerçekleştirimiş; İnce ve Çınar (2015) Türk kültürü ders kitaplarını derlem tabanlı incelemiş; Özkan (2020) terimleri derlem tabanlı olarak incelemiş; Tüfekçioğlu (2021) yabancı dil olarak Türkçe öğretiminde kültürel söz varlığı üzerinde derlem tabanlı bir araştırma gerçekleştirmiş; Şimşek ve Gün (2021) yabancı dil olarak Türkçe öğretiminde A1 düzeyinde en sık kullanılan sıfat, fiil ve isimler üzerinde derlem tabanlı çalışmış; Çalışkan (2016) öğrenir derlemleri üzerine araştırmalar yapmış; Çalışkan (2019) Türkçedeki görevli sözcükler ve işlevleri üzerine derlem tabanlı bir araştırma gerçekleştirmiş; Karadağ (2020) Türkçede ses-harf sıklığı üzerine çalışmıştır. Burada ele alınan ve taranan eserler doğal olarak derlem dilbilim alanı ile ilgili yapılan çalışmaların bütününü temsil etmekten ziyade araştırmacılarda bir izlenim oluşturmayı amaçlamaktadır.

Derlem Dilbilim ve Veri Yönetimli Öğrenme

Veri yönetimli öğrenme, dil öğreticileri tarafından son yıllarda sık kullanılmaya başlanan bir yöntemdir. Veri setini derlemlerin oluşturduğu bu öğrenme yöntemi öğreticiler için de öğrenciler için de önemli avantajlara sahiptir. Milyonlarca kelimelik derlemlerin erişime açıldığı düşünülüp bu derlemler aracılığı ile dilin işlevsel olarak öğretimi, dil bilgisi kurallarının betimlenmesi, tümevarım gibi imkânlar sayesinde veri yönetimli öğrenmenin popülaritesi artırmaktadır.

Veri yönetimli öğrenmenin kurucusu, dil öğrenenlerin aynı zamanda dil araştırmacıları olduğuna inanan Birmingham Üniversitesi akademisyenlerinden Tim Johns'tur (1989). Bu çerçevede Johns (1991) hedef dili etkili bir şekilde öğrenebilmek için öğrencilerin mümkün olduğu kadar özgün dilsel veriye erişmesi gerektiğini belirtmektedir. Dil öğrenimine yönelik bu yenilikçi yaklaşıma “veri yönetimli öğrenme” tanımlaması yapılmış; sürecin odağına ise otantik dil malzemesi konumlandırılmıştır. Böylece araştırmacılar için de öğrenciler içinde öğrenme süreçlerini yönetebilme imkânı sunulmuştur.

Veri yönetimli öğrenme geleneksel dil öğrenme ve öğretme yaklaşımlarından farklı olarak aktif olarak kullanılan yazılı ve sözlü dil içeriklerini hedef kitleye sunmakta; dile yönelik geniş perspektifli bir bakış açısına imkân sağlamaktadır (Chambers, 2010: 345). Derlemler aracılığı ile hedef kitleye çıkarımlarda bulunma, parçadan bütüne gitme imkânı sağlanır; hedef kitlenin öğrenme özerkliği ya da kendi kendine öğrenme becerisi de desteklenmiş olur. Johns’a (1991: 30) göre veri yönetimli öğrenmenin derlemler aracılığı ile diğer öğrenme yöntemlerinden farkı, ortadaki bilgi ve semantik yükü aktaran aracıların ortadan kaldırılmasıdır. O'Sullivan (2007), veri yönetimli öğrenme ile sadece dil öğrenimi için değil, gözlemleme, eleştirel düşünme, analiz etme, yorumlama, karşılaştırma, çıkarımlarda bulunma, teori oluşturma ve hipotez kurma gibi bir dizi genel becerileri geliştirdiğine dair önemli ve sıklıkla gözden kaçan bir noktaya da değinmektedir.

Veri yönetimli öğrenme ile hem dil öğrenenler hem de araştırmacılar veriye bütün olarak bakabilmekte, veriler arasındaki değişimi ve bağlantıları yorumlayabilmektedir. Bu da dili işlevsel olarak kullanma ve öğrenme becerileri açısından hedef kitleye derin bir bakış açısı sağlamaktadır. Bu modelde hedef kitle ele aldığı kavram ya da kelimenin anlam katmanlarını, kullanım özelliklerini, gönderim şekillerini saptayabilir. Sık tekrar eden hatalarını derlem üzerinden tespit ederek düzeltme yapabilir. Dilbilgisi kurallarını neden sonuç ilişkisi içinde betimleyerek genellemelere ulaşabilir.

Greaves ve Warren (2007), veri yönetimli öğrenme modelinde dil öğrenme etkinliklerinin, derlem etkinlikleriyle deyim gibi çok yönlü söz gruplarının anlam tabakalarının daha kolay betimlendiğini belirtmektedir. Bu etkinlikler aynı zamanda bir metnin deyimsel profiline ulaşmak için eleştirel becerilerin geliştirilmesine yardımcı olmakta; aynı türden metinlerle karşılaşıldığında hangi deyimlerin o metne özgü olduğunu belirlenebilmektedir.

Öğrenme modellerinin uygulanabilmesi ve çıktıların değerlendirilebilmesi için temel gereksinimlerin giderilmesi gerekmektedir. Veri yönetimli öğrenme modeli ile derlemleri kullanma ve oluşturma sürecinde yazılımlara ve bilgisayara teknolojilerine ihtiyaç duyulmaktadır. Ayrıca bu yazılımları kullanmak için gerekli eğitimlerin alınmış olması ve rehberlik yapacak uzman kişilerin de bulunması gerekmektedir.

Sonuç

Derlem, yazılı ve sözlü metinlerin çeşitli süreçlerden geçirilerek elektronik ortama aktarılması neticesinde oluşan etiketlenmiş ya da etiketlenmemiş veri olarak tanımlanabilir. Derlem dilbilimde bu verinin işlenmesi, istatistikî verilerin elde edilmesi için çeşitli yazılımlardan faydalanılmaktadır.

Derlem ve derlem dilbilimle ilgili temel kavramlar ve temel terminolojiyi anlatmayı esas alan bu çalışma hedef kitlede kuramsal çerçeveyi yapılandırmayı öncelemiştir. Teknolojinin gelişmesi ile bilgisayarlı dilbilime verilen önem özellikle 1990'lı yıllardan sonra artmıştır. Derlem dilbilimin ayrı bir disiplin olarak ele alınması belirginleşmiştir. 20. yüzyılın ikinci yarısından itibaren milyon kelimelik derlemler alanda kullanılmaya başlanmış; özellikle Batı dillerinde etiketlenmiş, artalanı olan derlemler araştırmacıların erişimine açılmaya başlanmıştır.

Derlem araştırmacılarının derlem oluşturma sürecinde dikkat etmesi gereken ilkeler bulunmaktadır. Bu ilkeler temsil edicilik, boyut, örneklem, denge, telif, makine tarafından okunabilirliktir. Temsil edicilik, oluşturulan veri setinin ilgili alan özelinde çıkarım yapmayı sağlayacak kapsama sahip olmasıdır. Boyut, derlem araştırmalarında uzun yıllardır tartışılan bir kavramdır. Derlemin sahip olduğu veri setinin nicel büyüklüğünü ifade eden boyut kavramı derlem türlerine göre çeşitlilik göstermektedir. 1990'lı yıllardan sonra artan bağımsız derlem araştırmacılarıyla oluşan özel derlemlerin boyutu değişiklik göstermektedir. Örneklem, araştırma yapılacak ilgili alanda temsil gücü olan veri setini ifade eder. Özellikle genel amaçlı derlemlerde "dil" ve "dil pratiği" açısından çeşitliliğin sağlanması gerekmektedir. Bu çerçevede derlemin veri seti dili temsil edecek tür çeşitliliğine sahip olmalıdır. Telif, derlemi oluşturan veri setinde kullanılacak metinlerin hukuki çerçevede alınan izinler neticesinde kullanımını içermektedir. İzinsiz olarak derleme dâhil edilen eserler hukuki problemlere sebebiyet verebilir. Makine tarafından okunabilirlik derlemlerin veri setlerinin dijital ortama aktarılması ile ilgilidir. Teknolojinin gelişimi ile bilginin dijital ortama aktarılması derlemlerin boyutlarını, veri setlerini kapasitesini artırmıştır. Makine tarafından okunan veri setleri aracılığı ile bağımsız araştırmacılar da derlem tasarlayıp araştırmalarını yürütebilmektedir.

Son yıllarda derlem üzerine artan kurumsal ve bağımsız araştırmalarla derlemin çıkış noktasındaki problem cümlelerinin çeşitliliği artmış; derlem türleri arasındaki ayrımda aynı paralelde belirginleşmiştir. Bu çerçevede derlem dilbilim alanında öne çıkan derlem türleri genel, özel, paralel, tarihsel, etiketlenmiş ve etiketlenmemiş, çok yönlü, öğrenci derlemlerinden oluşmaktadır. Genel

derlemler, yazılı ve sözlü metinlerden oluşan oldukça büyük veri setlerine sahiptir. Dile yönelik çıkarımlarda bulunmak, dil ile ilgili tümevarım içeren araştırmalar için tercih edilir. Genel derlemlerin veri setinde her türden metin içeriği bulunmaktadır. Özel derlemler, belirli bir alana hitap eden tematik içeriğe sahip derlemlerdir. Veri setleri genel derlemlere göre daha dar kapsamlıdır. Paralel derlemler, metinler arasındaki çeviri ilişkisine dayalı olarak tasarlanır, hedef dil ve derlemin hazırlandığı dil deki öğeler cümle düzeyinde eşleştirilir. Tarihsel derlem, bir dilin geçmiş aşamalarını temsil etmek ve/veya dil değişimini incelemek için oluşturulmuş bir derlem olarak tanımlanabilir. Etiketlenmiş derlem, derlemi oluşturan metinlerdeki sözcük türleri, sözdizim özellikleri belirlenmiş derlem olarak tanımlanabilir. Etiketlenmemiş derlemlerde ise veri setini oluşturan metinler kelime, sözdizimi gibi faktörler açısından düzenlenmemiştir. Dil eğitimi süresince öğrencilerin yazılı verilerinden oluşturulan derlemler öğrenci derlemleri olarak ifade edilmektedir.

Bilişim teknolojilerinin son yıllardaki hızlı gelişimi ile bağımsız araştırmacılarında derlem oluşturmasına imkân tanıyan yazılımlar ortaya çıkmıştır. Uluslararası anlamda öne çıkan yazılımlardan WordsmithTools lisanslı olarak kullanılabilen, oldukça fazla analiz modülüne sahip bir yazılımdır. AntConc yazılımı bağımsız araştırmacılar tarafından en sık kullanılan yazılımlardan birisi olup ücretsizdir. Her iki yazılımda da bağlam, eşdizimlilik, sıklık çalışmaları yürütülebilmektedir. Ancak WordsmithTools ile istatistik içerikli çalışmalar da yapılabilmektedir. TextStat yazılımı diğer yazılımlardan farklı olarak farklı türdeki dosyaları derlem veritabanına dâhil ederek analizler gerçekleştirebilmektedir. Ayrıca web tabanlı veri tabanı oluşturabilme imkânına da sahiptir.

Uluslararası anlamda derlem dilbilim ve derlem başlıkları üzerine derinlemesine birçok kaynak bulanmaktadır. Derlem üzerinde genel bir fikir edindirme amacıyla şekillendirilen bu çalışma ile alana kuramsal bir katkı sağlanacağı öngörülmüş; bu alanda çalışma yürütmek isteyen araştırmacıların derlemle ilgili öne çıkan dil, sözlük, sözcük türleri odaklı araştırmalardan en azından Türkiye ölçeğinde mevcut durum betimlenmeye çalışılmıştır. Çalışmada değinilen her kavram üzerine kuramsal anlamda da uygulama anlamında da çok daha derin araştırmalar gerçekleştirilebileceği kesindir.

Kaynakça

Aksan, Y., Aksan, M., Özel, S. A., Yılmazer, H., Demirhan, U. U., Mersinli, Ü. & Altunay, S. (2016). Web tabanlı Türkçe ulusal derlemi (TUD). In *Proceedings of the 16th academic computing conference*, 723-730.

Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh University Press.

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and linguistic computing*, *5*(4), 257-269.

Conrad, S. (2002) Corpus linguistic approaches for discourse analysis, *Annual Review of Applied Linguistics 22*(1). 75–95.

Çalışkan, N. (2016). Öğrenir derlemleri: kapsam, tasarım ve uygulamalar. *Turkophone, 1*(1), 40-53.

Çalışkan, N. (2019). Diye sözcüğü üzerine derlem temelli bir betimleme. *Türkbilig*, (*38*), 163-176.

Chambers, A. (2010) 'What is data-driven learning?', in A. O'Keeffe and M. McCarthy (eds), *The Routledge Handbook of Corpus Linguistics,* Routledge, 345–358.

Claridge, C. (2008) 'Historical corpora', in A. Lüdeling and M. Kytö (eds), *Corpus Linguistics: An International Handbook* 1, Walter de Gruyter, 242–259.

Clear, J. (1992) 'Corpus sampling', in G. Leitner (ed.) *New Directions in English Language Corpora*. Mouton de Gruyter, 21–31.

Gatto, M. (2014). *Web as corpus: theory and practice*. A&C Black.

Gries, S. T. (2009). What is corpus linguistics?. *Language and linguistics compass*, *3*(5), 1225-1241.

Greaves, C. & Warren, M. (2007) 'Concgramming: A computer-driven approach to learning the phraseology of English', *ReCALL Journal*, *17*(3), 287–306.

Gündoğdu, A. E. (2012). *Türk çocuk yazınında söz varlığı-derlem tabanlı bir uygulama*. [Yayımlanmamış yüksek lisans tezi]. Mersin Üniversitesi Eğitim Bilimleri Enstitüsü.

Francis, W. & Kucˇera, H. (1964) *Manual of ınformation to accompany a standard corpus of present-day edited American English*, for Use with Digital Computers. Brown University, Department of Linguistics.

Johns, T. (1989) 'Whence and whither classroom concordancing?', in T. Bongaerts, P. de Haan, S. Lobbe and H. Wekker (eds), *Computer Applications in Language Learning*, Foris: 9–30.

Johns, T. (1991) 'Should you be persuaded: two samples of data-driven learning materials', in T. Johns and P. King (eds), *Classroom Concordancing* (English Language Research Journal 4), ELR: 1–10.

İnce, B. & Çınar, İ. (2015). Türkçe ve Türk kültürü ders kitaplarındaki söz varlığına derlem temelli bir bakış. *IJLET*, *3*(1), 198-209.

Karadağ, Ö. (2020). Türkçe eğitimi açısından ses-harf sıklığı üzerine bir değerlendirme. *Ana Dili Eğitimi Dergisi*, *8*(1), 1-13.

Kennedy, G. (1998). *An introduction to corpus linguistics*. Longman.

Kenning, M.M. (2010) 'What are parallel and comparable corpora and how can we use them?', in A. O'Keeffe and M. McCarthy (eds), *The Routledge Handbook of Corpus Linguistics*, Routledge: 487–500.

Krishnamurthy, R. (2004) *English Collocation Studies*. Continuum (new edition of J. Sinclair, S. Jones and R. Daley (1970) English Lexical Studies).

Kurtoğlu, Ö. A. (2015). *Türkçede ad-eylem eşdizimliği için istatistiksel ve anlamsal ölçütler temelinde derlem-çıkışlı bir ulamlama*. [Yayımlanmamış doktora tezi]. Ankara Üniversitesi, Sosyal Bilimler Enstitüsü.

Kübler, N. & Aston, G. (2010). Using corpora in translation. *The Routledge handbook of corpus linguistics*, 505-515.

Leech, G. (1991) 'The State-of-the-Art in corpus linguistics', in K. Aijmer and B. Altenberg (eds) *English Corpus Linguistics*, Studies in Honour of Jan Svartvik. Longman, 8–29.

Lee, D.Y.W. (2010) 'What corpora are available?', in A. O'Keeffe and M. McCarthy (eds), *The Routledge Handbook of Corpus Linguistics*, Routledge: 107–121.

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. Taylor & Francis.

McEnery, T. & Hardie, A. (2011). *Corpus linguistics: method, theory and practice*. Cambridge University Press.

Oostdijk, N. (1991) *Corpus Linguistics and the automatic analysis of English*. GA: Rodopi.

O'Sullivan, I. (2007) 'Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy', *ReCall, 19*(3), 269–286.

Özkan, B. (2013). Yöntem ve uygulama açısından "Türkiye Türkçesi söz varlığının derlem tabanlı sözlüğü". *Bilig, 66*, 149-178.

Özkan, B. (2020) Terim bilim çalışmalarında derlem kullanımı ve "kendi kendine derlem oluşturma platformu" terim bilim modülü. *Avrasya Terim Dergisi*, *8*(1), 8-17.

Sinclair, J. (2005). Meaning in the framework of corpus linguistics. *Lexicographica,* 20(2005), 20-32.

Sert, G. (2018). *Anlam alanı ve anlam ezgisi açısından eş anlamlı durum sıfatları (derlem tabanlı bir inceleme)*. [Yayımlanmamış doktora tezi], Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Enstitüsü.

Sezer, B. & Sezer, T. (2013). TS corpus: herkes için Türkçe derlem. *In Proceedings of the 27th national linguistics conference*. 217-225.

Şimşek, R. (2020). *Subliminal bir güç göstergesi: yabancı dil olarak Türkçe öğretimi ders kitapları* [Yayımlanmamış doktora tezi]. İnönü Üniversitesi, Eğitim Bilimler Enstitüsü.

Şimşek, R. & Gün, M. (2021). A corpus-based proposal for the vocabulary to be taught at a1 level in the teaching of turkish as a foreign language. *Educatıonal Polıcy Analysıs And Strategıc Research*, *16*(1), 356-373. https://dx.doi.org/10.29329/epasr

Özkan, B. (2010). Türkçenin öğretiminde sifatlarin eşdizim sözlüğü: yöntem ve uygulama. e-*Uluslararası Eğitim Araştırmaları Dergisi*, *1*(2), 51-65.

Özkan, B. (2014). Türkiye Türkçesi söz varlığında fiillerin derlem denetimi ve derlem tabanlı sözlüğü. *Bilig,* (*69*), 171-204.

Özkan, B., Tahiroğlu, B. T. & Özkan, A. E. (eds.). (2014). *Türkçe üzerine derlem dilbilim uygulamaları*. Karahan Kitabevi.

Tahiroğlu, B.T. (2010). *Bilgisayar destekli sözlük bilimi çalışmalarında derleme sözlüğü veri tabanı örneği* [Yayımlanmamış doktora tezi]. Çukurova Üniversitesi Sosyal Bilimler Enstitüsü.

Teubert, W. & Čermáková, A. (2004). Directions in corpus linguistics. *Lexicology and corpus linguistics*, 113-166.

Tognini-Bonelli, E. & Sinclair, J. (2006) 'Corpora', in K. Brown (ed), *Encyclopedia of Language and Linguistics,* Elsevier: 216–219.

Tüfekçioğlu, B. (2013). *Derlem tabanlı çevrim içi Türkçe öğrenci sözlüğü: önadlar A madde başı*. [Yayımlanmamış yüksek lisans tezi]. Mersin Üniversitesi Eğitim Bilimleri Enstitüsü.

Tüfekçioğlu, B. (2018). *Yabancı dil olarak akademik Türkçe: Sosyal bilimlerde akademik ve teknik söz varlığı.* [Yayımlanmamış doktora tezi]. Hacettepe Üniversitesi, Türkiyat Araştırmaları Enstitüsü.

Tüfekçioğlu, B. (2021). Cultural vocabulary in teaching Turkish as a foreign language. *Eurasian Journal of Applied Linguistics*, *7*(1), 338-358. https://dx.doi.org/10.32601/ejal.911449