Lisanî İlimler Dergisi
Journal of Linguistic Studies
Aralık / December 2025, 3 (2): 579-598

Derlem Dilbilimi ile Makine Çevirisinin Tarihsel Gelişimi: Çeviribilimsel Bir İnceleme*

The Historical Development of Corpus Linguistics and Machine Translation: A Translation Studies Perspective

Alper ÇALIK

Bartın Üniversitesi, Edebiyat Fakültesi, Mütercim ve Tercümanlık Bölümü, İngilizce Mütercim ve Tercümanlık A.B.D.
Bartin University, Faculty of Letters, Translation and Interpreting Department, Division of English Translation and Interpreting
Bartın / Türkiye
acalik@bartin.edu.tr
ORCID: 0000-0002-7261-9385

Prof. Dr. Giray FİDAN

Ankara Hacı Bayram Veli Üniversitesi, Edebiyat Fakültesi, Doğu Dilleri ve Edebiyatları Bölümü, Çin Dili ve Edebiyatı A.B.D.
Ankara Hacı Bayram Veli University, Faculty of Letters, Eastern Languages and Literature Department, Division of Chinese Language and Literature
Ankara / Türkiye
giray.fidan@hbv.edu.tr
ORCID: 0000-0002-5002-9253

Makale Bilgisi / *Article Information*

Makale Türü / *Article Types* : Araştırma Makalesi / *Research Article*
Geliş Tarihi / *Received* : 16.11.2025
Kabul Tarihi / *Accepted* : 21.12.2025
Yayın Tarihi / *Published* : 30.12.2025
Yayın Sezonu / *Pub Date Season* : Aralık / *December*
Cilt / *Volume*: 3 • **Sayı / *Issue*:** 2 • **Sayfa / *Pages*:** 579-598

Atıf / *Cite as*
ÇALIK, A., FİDAN, G. (2025). Derlem Dilbilimi ile Makine Çevirisinin Tarihsel Gelişimi: Çeviribilimsel Bir İnceleme, *Lisânî İlimler Dergisi*, 3(2), 579-598.

Doi: 10.5281/zenodo.18033927

İntihal / *Plagiarism*
Bu makale, en az iki hakem tarafından incelendi ve intihal içermediği teyit edildi.
This article has been reviewed by at least two referees and scanned via a plagiarism software.

* Bu makale Alper ÇALIK tarafından Prof. Dr. Giray FİDAN danışmanlığında hazırlanan henüz yayınlanmamış doktora tezinden üretilmiştir.

Öz: Teknolojik gelişmeler, diğer birçok alanda olduğu gibi çeviri alanında da köklü dönüşümlere yol açmıştır. Özellikle 1990'lı yıllardan itibaren bilgisayarlara erişimin yaygınlaşmasıyla birlikte bilgisayar destekli çeviri (BDÇ) araçları ve makine çevirisi sistemleri çeviri pratiğinde giderek daha görünür ve yaygın hâle gelmiştir. Günümüzde ise geleneksel çeviri anlayışı büyük ölçüde, nöral makine çevirisi (NMT) motorları ve büyük dil modelleri (Large Language Models, LLMs) tarafından üretilen çıktılar üzerinde yapılan düzenleme ve son okuma süreçlerine evrilmiştir. Bu dönüşümde özellikle Google'ın 2016 yılında nöral makine çevirisi sistemini kullanıma sunması önemli bir dönüm noktası olmuş; sonrasında yapay zekâ tabanlı sistemlerin hızla yaygınlaşmasıyla bu teknolojik etki daha da belirgin hâle gelmiştir. Bu doğrultuda, özelleştirilmiş makine çevirisi motorları ve büyük dil modelleri üzerine yürütülen çalışmalar giderek artan bir önem kazanmıştır. Söz konusu sistemlerin temelinde yer alan en kritik unsur ise veridir; özellikle paralel derlemler (çift dilli corpuslar), bu sistemlerin eğitilmesi, özelleştirilmesi ve çıktı kalitesinin artırılmasında vazgeçilmez bir bileşen hâline gelmiştir. Geçmişte daha çok betimleyici ve kuramsal dilbilim çalışmaları için oluşturulan derlemler, günümüzde teknolojik gelişmelerle bütünleşerek nöral makine çevirisi motorları ve büyük dil modelleri için sistematik biçimde hazırlanmaktadır. Bu bağlamda, bu çalışmada derlemlerin tarihsel gelişimi, makine çevirisinin tarihsel süreç içerisindeki dönüşümü ve İngilizce–Türkçe derlem çalışmaları incelenmektedir.

Anahtar Kelimeler: Derlem Dilbilimi, Makine Çevirisi, Çeviribilim, Paralel Derlemler, BDÇ Araçları

Abstract: Technological developments have led to profound transformations in the field of translation, as in many other disciplines. Especially since the 1990s, with the widespread accessibility of computers, computer-assisted translation (CAT) tools and machine translation systems have become increasingly prevalent in translation practice. Today, traditional translation has largely evolved into the post-editing and revision of outputs produced by neural machine translation (NMT) engines and large language models (LLMs). A major milestone in this transformation was the introduction of its neural machine translation system by Google in 2016, which significantly accelerated the technological impact on the field. This impact has become even more evident with the rapid proliferation of contemporary artificial intelligence systems. Accordingly, research on customized machine translation engines and large language models has gained increasing importance. The most fundamental component underlying these customized systems is data, particularly parallel corpora, which have become indispensable for training, tailoring, and improving the quality of such systems. While corpora were traditionally developed mainly for linguistic and descriptive studies, they are now increasingly integrated with technological applications and systematically constructed for neural machine translation engines and large language models. In this context, the present study examines the historical development of corpora, the evolution of machine translation throughout history, and English–Turkish corpus studies.

Keywords: Corpus Linguistics, Machine Translation, Translation Studies, Parallel Corpora, CAT Tools

Giriş

Teknolojik gelişmeler, çeviri alanında köklü dönüşümlere neden olarak hem mesleki uygulamaları hem de akademik araştırmaları yeniden şekillendirmiştir. 1990'lı yıllardan itibaren kişisel bilgisayarların yaygınlaşması, bilgisayar destekli çeviri (BDÇ) araçları ile makine çevirisi (MÇ) sistemlerinin hızla benimsenmesini sağlamış ve çeviri iş akışlarının özünü köklü biçimde değiştirmiştir (Hutchins & Somers, 1992). Son yıllarda ise özellikle nöral makine çevirisi (NMÇ) sistemlerinin geliştirilmesi gibi makine çevirisi teknolojilerindeki ilerlemeler çeviri çıktılarının kalitesinde kayda değer iyileşmeler sağlamıştır. Büyük veri setleri ile eğitilen bu sistemler, daha geniş oranda bağlamsal bilgiyi yakalayabilmekte olup hedef dilde daha akıcı çıktılar üretebilmektedir. Nitekim, ChatGPT-4o ve DeepSeek-V3 gibi büyük dil modellerinin (LLM), post-editing görevleri sırasında öğrenci çevirmenler tarafından kullanıldığında, çeviri kalitesi ve kullanıcı deneyimi açısından geleneksel NMÇ sistemlerini geride bıraktığını ifade eden yeni tarihli çalışmalar ön plana çıkmaktadır (Shao & Zhu, 2025, s. 5). Bu tür gelişmeler, yalnızca profesyonel çeviri uygulamalarını değil, aynı zamanda çevirmen eğitimindeki pedagojik yaklaşımları da dönüştürmektedir.

Nöral makine çevirisi ve büyük dil modellerindeki ilerlemelere paralel olarak, bilgisayar destekli çeviri araçlarının giderek artan teknik kapasitesi, çevirmenlerin verimliliğini, tutarlı olmasını ve terminolojik denetim yapabilmelerini önemli ölçüde artırarak çeviri iş akışlarını başka bir boyuta taşımıştır. Çeviri belleği, terminoloji yönetimi ve BDÇ araçlarına entegre gerçek zamanlı kalite kontrol mekanizmaları gibi temel işlevler, serbest çevirmenler ile kurumsal çeviri ortamlarının her ikisinde de vazgeçilmez hâle gelmiştir. Ayrıca, otomatik makine çevirisi sistemlerinin BDÇ araçlarına sorunsuz biçimde entegre edilmesi, bu araçların kullanımını profesyonel çevirmenler açısından kaçınılmaz hale getirmiş; böylece BDÇ araçları, günümüz çeviri pratiğinde insan–makine iş birliğinin merkezi platformları olarak konumlanmıştır (O'Brien ve ark., 2014).

BDÇ araçları ile makine çevirisi sistemlerinin çeviri iş akışları içindeki giderek artan merkezi konumu, çeviri pratiğinde veri ve derlemlerin (corpora) rolünü de daha fazla güçlendirmiştir. BDÇ araçlarında kullanılabilen çeviri bellekleri, terminoloji veri tabanları ve makine çevirisi sistemleri, çift dilli veya çok dilli derlemler ile oluşturulmaktadır. Bunun sonucunda da çeviri üretimi daha önce çevrilmiş metinlerin yeniden kullanılmasına, bu verilerin hizalanmasına veya yeniden modellenmesine daha fazla bağlı hale gelmiş olup derlem temelli bir nitelik kazanmıştır.

Tarihsel olarak derlemler, dijital ortamlarda modelleme amacıyla değil, betimleyici ve ampirik dilbilimsel çözümlemeler için geliştirilmiştir. Bu bağ-

lamda öne çıkan örneklerden biri, 1958 yılında hazırlanan Fransızca *Thesaurus* (Trésor de la Langue Française, TLF) derlemidir. Randolph Quirk tarafından başlatılan *Survey of English Usage* (SEU) derlemi ise, elektronik metin işlemenin ortaya çıkışından önce, açık biçimde dilbilgisel betimleme amacıyla derlenmiş en önemli derlemlerden olarak kabul edilmektedir (Quirk, 1968). Başlangıçta dijital olmayan yapısı nedeniyle bilgisayar temelli çözümlemelere elverişli olmasa da SEU, geleneksel olarak elle derlenen derlemler ile daha sonra geliştirilen tamamen bilgisayarlaştırılmış derlem yöntemleri arasında bir köprü işlevi görerek derlem dilbiliminin evriminde merkezi bir konum edinmiştir.

1959 yılında oluşturulan SEU, her biri yaklaşık 5.000 sözcükten oluşan ve konuşma ile yazı dilini kapsayan yaklaşık 200 metin örneğinden meydana gelen, toplamda bir milyon sözcüklük dengeli bir derlem oluşturmayı hedeflemiştir. Bu derlem, İngiliz İngilizcesini ana dili olarak konuşan eğitim almış yetişkinlerin dilbilgisi yapılarının ve kullanım örüntülerinin sistematik biçimde betimlenmesi için bir temel sunmak üzere tasarlanmıştır. Özellikle SEU'nun sözlü dil bileşeni, farklı resmiyet düzeylerini ve kişilerarası etkileşim derecelerini yansıtan geniş bir iletişim bağlamı ve söylem türü içerisinden derlenmiştir (Kennedy, 1998). Derlemin oluşturulmasındaki temel amaç dilbilimsel incelemeler yürütmek olup, derlem çalışmalarının dil kuramlarının gelişimine önemli katkılar sunduğu ve dilin doğası, yapısı ve kullanımı hakkında vazgeçilmez veriler sağladığı görülmektedir. Bununla birlikte, derlemlerin zamanla dijital ortama aktarılmaya başlanmasıyla birlikte, yürütülen çalışmaların yönü ve kapsamı da belirgin biçimde değişmeye başlamıştır.

Erken dönem dijital derlemler arasında en etkili örneklerden biri olan Brown Corpus, sistematik biçimde toplanmış özgün dil verisi sunarak ampirik dil araştırmalarında köklü bir dönüşüm yaratmıştır (Kučera & Francis, 1967). Bu derlem, sezgiye dayalı dilbilgisel betimlemeden veri temelli çözümlemeye geçişi simgelemiş ve modern derlem dilbiliminin temelini atmıştır (McEnery & Hardie, 2012). Zamanla derlemler, görece küçük ve elle derlenen veri kümelerinden, dijital ortamlar için tasarlanmış, otomatik olarak toplanan ve web ölçeğine ulaşan büyük kaynaklara evrilmiştir. Erken dönem derlem çalışmaları ağırlıklı olarak tek dilli veri kümelerine odaklanmış olup özgün dil verisindeki sıklık temelli gözlemler ile sözdizimsel eğilimlerin incelenmesini öncelemiştir (Kennedy, 1998; Sinclair, 1991).

Teknolojik gelişmelerle birlikte derlem dilbiliminin kapsamı genişlemiş; özellikle tümce ya da segment düzeyinde hizalanmış iki dilli ve çok dilli derlemler, paralel derlemler başta olmak üzere, giderek daha fazla önem kazanmıştır. Bu tarihsel arka plan dikkate alındığında, makine çevirisinin evrimi ile derlem çalışmalarının gelişiminin birbirine derinlemesine bağlı

süreçler olduğu açıkça görülmektedir. İlk derlemler öncelikle dilbilimsel betimleme için ampirik araçlar olarak tasarlanmış olsa da zaman içinde dijitalleşmeleri ve hacimlerinin büyümesi hem dilbilimsel araştırmaları hem de çeviri teknolojilerini köklü biçimde yeniden şekillendirmiştir. Kural temelli ve istatistiksel yaklaşımlardan veri odaklı nöral yapılara geçişle eş zamanlı olarak, derlemler pasif konumdan çıkarak, çağdaş makine çevirisi sistemlerinde eğitim, değerlendirme ve özelleştirme süreçlerinin temelini oluşturan etkin bileşenler hâline gelmiştir. Özellikle nöral makine çevirisi motorları ile büyük dil modellerinin büyük ölçekli paralel derlemlere bağlılığının artması, derlemlerin rolünü tamamlayıcı analitik araçlar olmaktan çıkararak, temel altyapısal kaynaklar olarak yeniden tanımlamıştır.

Bu doğrultuda, günümüz çeviri teknolojilerinin kapsamlı biçimde anlaşılabilmesi, makine çevirisi paradigmalarının evrimi ile derlem dilbiliminin gelişimini birlikte ele alan tarihsel bir bakış açısını gerekli kılmaktadır. Bu nedenle sonraki bölümlerde öncelikle makine çevirisinin tarihsel gelişimi ele alınmakta, erken dönem deneysel sistemlerden nöral ve büyük dil modeli temelli yaklaşımlara uzanan süreç ortaya konulmaktadır. Ardından, alandaki yöntemsel standartların oluşumunda belirleyici rol oynayan başlıca İngilizce derlem projelerine odaklanılarak derlem çalışmalarının tarihsel gelişimi incelenmektedir. Sonrasında, Türkçe derlem çalışmalarına ilişkin bir genel değerlendirme sunulmaktadır. Son olarak çalışma, güncel araştırma ve çeviri pratiğinde derlemlerin başlıca kullanım alanlarını tartışarak, derlemlerin modern çeviri iş akışları içinde hem analitik araçlar hem de işlevsel işletim mekanizmaları olarak nasıl konumlandığını aktarmaktadır.

Makine Çevirisinin Evrimi

Makine çevirisi (MÇ), 20. yüzyılın ortalarında ortaya çıkışından bu yana, her biri etkin dil kuramları, mevcut bilgisayar kaynakları ve veri erişilebilirliği tarafından şekillendirilen birden fazla özgün gelişim evresinden geçmiştir. Makine çevirisine yönelik ilk yaklaşımlar büyük ölçüde kural temelli olup, kaynak dil yapılarının hedef dildeki karşılıklarıyla eşleştirilmesi amacıyla titizlikle oluşturulmuş dilbilgisel kurallara ve çok dilli sözlüklere dayanmaktadır. Bu sistemler, dönemin dilbiliminde baskın olan yapısalcı ve üretici paradigmaları yansıtan sembolik ve dilsel bilgi temsilleri üzerine kurulmuştur. Ancak, önemli kuramsal gelişmelere rağmen, kural tabanlı makine çevirisi sistemleri dilsel belirsizlik, alanlar arası değişkenlik ve doğal dilin kendi içerisindeki karmaşıklığıyla başa çıkmakta zorlanmış; bu durum çeviri kalitesinin ve ölçeklenebilirliğin sınırlı kalmasına yol açmıştır (Hutchins & Somers, 1992).

Kural Tabanlı Makine Çevirisi (Rule-Based Machine Translation, RBMT), tek dilli ve iki dilli metinler, çok sayıda dilbilgisel kural, bir sözlük ve bu ku-

ralları işleyen yazılım bileşenlerinden oluşan bir sistemdir. Bu tür sistemler, tanımlanan kurallar, sözlükler ve motora sağlanan metinler doğrultusunda çeviri önerileri üretmektedir. Dilsel kurallar, çözümleme (analysis), aktarım (transfer) ve üretim (generation) aşamalarında toplanmakta ve uygulanmaktadır. Sistem, kaynak ve hedef dillere ait çok sayıda dilbilgisel kural kullanılarak eğitilmekte olup makine bu kuralları işleyerek ilgili sözcükleri düzenlemekte ve bir çıktı üretmektedir. Ancak, bir dilde bulunan çok sayıda dilbilgisel kural nedeniyle, özellikle büyük hacimli veriler işlendiğinde, bu sistemlerden elde edilen çıktılar genellikle istatistiksel makine çevirisi sistemlerine kıyasla daha sınırlı bir nitelik sergilemektedir. Teorik olarak tüm bu kuralların motora öğretilmesi mümkün olmakla birlikte, dilbilgisi kurallarının çok fazla oluşu sebebiyle zaman ve maliyet açısından oldukça sorunlu bir süreçtir.

RBMT sistemlerinin geliştirilmesi, istatistiksel veya nöral makine çevirisi sistemlerine kıyasla daha az veriye ihtiyaç duymaktadır. Küçük ölçekli bir RBMT sistemi, dilbilgisel kuralların sisteme aktarılması ve sözlüklerin kullanılması yoluyla geliştirilebilse de bu yaklaşım genellikle akıcı ve doğal çıktılar elde etmek için yeterli olmamaktadır. Bu nedenle, bir RBMT sisteminin geliştirilmesi, dilsel kuralların motora doğru biçimde entegre edilebilmesi için yüksek düzeyde dil uzmanlığı gerektirmektedir.

1980'li yılların sonları ve 1990'lı yıllarda, veri işleme gücündeki artış ve dijital hale getirilmiş metinlerin giderek daha erişilebilir hâle gelmesi, veri odaklı yaklaşımlara yönelimi mümkün kılmıştır. Kural temelli sistemlerin sınırlılıklarını gidermek amacıyla geliştirilen istatistiksel makine çevirisi çeviriyi çok dilli paralel derlemlere dayanan olasılıksal bir süreç olarak kavramsallaştırmıştır. Dilsel bilginin doğrudan kodlanması yerine, istatistiksel makine çevirisi algoritmaları sözcük ve tümce öbeği hizalamalarına ilişkin olasılıkları hesaplayarak büyük veri kümelerinden çeviri örüntülerini türetmiştir.

Bu derlem temelli yaklaşım, makine çevirisinin evriminde kritik bir dönüm noktasını temsil etmektedir. Çünkü çeviri kalitesi, doğrudan doğruya kullanılan eğitim verisinin miktarına ve temsil gücüne bağlı hâle gelmiştir. Özellikle tümce öbeği temelli istatistiksel makine çevirisi sistemleri, farklı alanlarda akıcılık ve farklı alanlardan metinlerde performansını koruyabilme açısından kayda değer iyileşmeler göstermiş olup derlemler, makine çevirisi araştırma ve geliştirme süreçlerinde vazgeçilmez kaynaklar olarak konumlanmıştır (Koehn, 2010).

Bir diğer önemli dönüşüm, çeviriyi yapay sinir ağlarını kullanan uçtan uca bir öğrenme problemi olarak yeniden tanımlayan nöral makine çevirisinin ortaya çıkışıyla gerçekleşmiştir. Nöral makine çevirisi sistemleri, kaynak ve hedef dildeki sözcükleri sürekli vektör temsilleri olarak modellemek üzere, çoğunlukla dikkat mekanizmalarıyla desteklenen kodlayıcı–çözücü (enco-

der–decoder) mimarilerden yararlanmaktadır. Bu yaklaşım, sistemin bağlamsal ve anlamsal bilgiyi önceki istatistiksel modellere kıyasla daha etkili biçimde öğrenmesine olanak tanımakta ve sonuç olarak daha akıcı ve tutarlı çeviri çıktılarının üretilmesini sağlamaktadır.

Google'ın nöral makine çevirisi sistemini 2016 yılında kullanıma sunması, nöral makine çevirisi sistemlerinin yaygınlaşması açısından kritik bir dönüm noktası olmuş, çok sayıda dil çifti üzerinde tümce öbeği temelli istatistiksel makine çevirisine kıyasla belirgin kalite artışları sağlandığını ortaya koymuştur (Wu ve ark., 2016). Bu tarihten itibaren nöral makine çevirisi, hem akademik araştırmalarda hem de ticari makine çevirisi uygulamalarında baskın paradigma hâline gelmiştir.

Büyük dil modellerinin (LLM) yakın dönemde ortaya çıkışı, makine çevirisinin kapsamını ve kuramsal sınırlarını önemli ölçüde genişletmiştir. Yalnızca çeviri görevine odaklanan geleneksel nöral makine çevirisi sistemlerinin aksine, LLM'ler çok dilli ve çok kipli (multimodal) son derece büyük veri kümeleri üzerinde önceden eğitilmekte ve ardından çeviri de dâhil olmak üzere çok sayıda dilsel görev için ince ayar (fine-tuning) sürecinden geçirilmektedir. Bu modeller, bağlamı ele alma, söylem düzeyinde tutarlılığı koruma ve üslup çeşitliliği sergileme konularında dikkate değer bir yetkinlik ortaya koymakta kullanıcı odaklı değerlendirmelerde çoğu zaman geleneksel NMT sistemlerinin ötesine geçmektedir. Bu gelişmelerin bir sonucu olarak çeviri iş akışları, giderek artan biçimde post-editing ve insan–makine iş birliğine dayalı bir yapıya evrilmiş; çevirmenlerin rolü, metni sıfırdan üretmekten ziyade, makine tarafından oluşturulan çıktıları gözden geçirip iyileştirmeye yönelmiştir (O'Brien, 2012).

Derlemler başta olmak üzere tüm bu gelişim evreleri boyunca verinin; özellikle de çift dilli verinin makine çevirisinde önemi giderek daha belirleyici hâle gelmiştir. Erken dönem kural temelli sistemler ağırlıklı olarak dilbilgisel kurallara ve uzman bilgisine dayanırken, çağdaş nöral makine çevirisi sistemleri ile büyük dil modelleri özünde veri odaklıdır ve çeviri kalitesi, eğitimde kullanılan derlemlerin ölçeği, çeşitliliği ve niteliği ile doğrudan ilişkilidir. Özellikle paralel derlemler, günümüz makine çevirisi sistemlerinin temelini oluşturmakta olup hem model eğitimi hem de sistematik değerlendirme süreçlerini mümkün kılmaktadır (Tiedemann, 2011). Bu bağlamda, makine çevirisinin gelişimi ancak çeviri teknolojilerindeki ilerlemeler için ampirik bir zemin sağlayan derlem çalışmalarının eş zamanlı ilerleyişi ile birlikte tam anlamıyla kavranabilir.

Derlem Çalışmalarının Tarihsel Gelişimi

Derlem çalışmalarının tarihsel evrimi, derlem dilbiliminin bağımsız bir disiplin olarak resmen ortaya çıkışından daha önceye dayanmaktadır. Yeni-

likçi özelliklerine rağmen, dijitalleşme öncesi dönemde hazırlanan derlemler ölçek, kapsam ve analitik kesinlik açısından yapısal sınırlılıklar taşımaktaydı. Bu tür verilerin çözümlenmesi büyük ölçüde bireysel akademik takdire dayanmakta; standartlaştırılmış yöntemlerin yokluğu, bulguların titiz biçimde yeniden üretilebilmesini güçleştirmekteydi. Yirminci yüzyılın ortalarına kadar dilbilim kuramları, ağırlıklı olarak sezgiye dayalı ve yeterlilik odaklı yaklaşımlar tarafından belirlenmiş; bu durum, gerçek dil verilerinin kapsamlı biçimde kullanılmasını önemli ölçüde sınırlandırmıştır.

1960'lı yıllarda makine tarafından okunabilir derlemlerin ortaya çıkışı, derlem araştırmalarının gelişiminde kritik bir dönüm noktasını oluşturmuştur. Brown Corpus'un hazırlanması, özgün dil verisinin bilgisayar destekli yöntemlerle sistematik biçimde incelenmesine yönelik ilk kapsamlı girişim olarak kabul edilmektedir (Francis & Kučera, 1967). Bilgisayar teknolojilerindeki ilerlemeler, metinsel verilerin depolanmasını, işlenmesini ve nicel olarak çözümlenmesini mümkün kılmış; bu sayede araştırmacılar farklı türler ve söylem alanları boyunca sıklık dağılımlarını, eşdizimsel örüntüleri ve yapısal düzenlilikleri inceleyebilmiştir.

Yirminci yüzyılın ortalarında, geçerli dilbilimsel genellemelerin soyut dil yeterliliği kavramlarından ziyade gerçek dil kullanımına ilişkin ampirik örüntülere dayanması gerektiğinin dilbilimciler tarafından giderek daha fazla kabul görmesiyle önemli bir epistemolojik dönüşüm ortaya çıkmıştır (Leech, 1991). Bu kuramsal yeniden yönelim, ampirik ve kullanım temelli yöntemler için elverişli bir düşünsel zemin hazırlamış ve çağdaş derlem temelli araştırmaların kavramsal temelini oluşturmuştur.

Bu ilk gelişmeleri takiben derlem dilbilimi, ampirik kanıtı ve sistematik veri çözümlemesini önceleyen bütüncül bir yöntemsel çerçeveye doğru kademeli olarak evrilmiştir. Yirminci yüzyılın sonlarına gelindiğinde, derlem temelli yaklaşımlar, sezgiye dayalı çözümlemelere güvenilir bir alternatif olarak kabul görmüş ve dilbilimsel araştırmalar için sağlam bir ampirik dayanak sunmuştur (Sinclair, 1991). Kennedy (1998), *derlem dilbilimi* teriminin görece yeni olmasına karşın, bu yaklaşımın dilbilimde uzun süredir var olan ampirik yöntemleri biçimselleştirdiğine dikkat çekmektedir.

Elektronik araçların ortaya çıkışından önce, tarihsel sözlükbilim ve filoloji alanlarında çalışan araştırmacılar, sözcük anlamlarını, kullanım örüntülerini ve dilbilgisel özellikleri belgelemek amacıyla dil örneklerini titizlikle toplamışlardır (Biber, Conrad & Reppen, 1998). Dijital araçlar kullanmadan derlenen metin alıntılarına dayanan bu erken dönem derlem benzeri koleksiyonlar, dilbilimsel betimlemeyi yalnızca kuramsal çıkarımlara değil, belgelenmiş gerçek dil kullanımına dayandırmaya yönelik ilk girişimleri temsil etmektedir.

1990'lı yıllardan itibaren yaşanan hızlı teknolojik gelişmeler, derlem sayısında ve tür çeşitliliğinde belirgin bir artışa yol açmıştır. Özenle tasar-

lanmış küçük ölçekli derlemler, yüz milyonlarca hatta milyarlarca sözcük içeren kapsamlı ulusal, başvuru ve web tabanlı derlemlerle desteklenmiş; pek çok durumda bu büyük ölçekli derlemler önceki yaklaşımların yerini almıştır (McEnery & Wilson, 2001). Bu süreçte, derlem ek açıklama (annotation) yöntemleri de önemli ölçüde gelişmiş olup sözcük türü etiketleme, sözdizimsel çözümleme ve giderek daha karmaşık dilbilimsel bilgi katmanlarının eklenmesi yaygınlık kazanmıştır.-Bu gelişim, karşılaştırmalı dilbilim ve diller arası araştırmaların yükselişini yansıtan iki dilli ve çok dilli derlemlere yönelik artan ilgiyle eş zamanlı olarak gerçekleşmiştir. Derlem tekniklerindeki ilerlemelerle birlikte, çok dilli ve özellikle paralel derlemler, diller arası karşılaştırmalı çözümlemeler ile çeviri odaklı araştırmalar için temel kaynaklar hâline gelmiştir (Granger, 2003).

Yirmi birinci yüzyılda derlemler, betimleyici dilbilimin yanı sıra çeviribilim, doğal dil işleme ve makine çevirisi gibi uygulamalı alanlarda da merkezi bir konum edinmiştir. Paralel ve karşılaştırılabilir derlemler, veri odaklı ve nöral dil modelleri için stratejik öneme sahip temel eğitim verileri hâline gelmiş; dil teknolojilerindeki çağdaş yöntemleri doğrudan etkilemiştir (Tiedemann, 2011). Bu doğrultuda, derlem tasarım ilkeleri günümüzde yalnızca temsil edilebilirlik ve denge gibi dilbilimsel ölçütler tarafından değil, aynı zamanda veri miktarı, hizalama kalitesi ve hesaplama verimliliği gibi teknik etmenler tarafından da belirlenmektedir (McEnery & Hardie, 2012). Bu uzun tarihsel süreç boyunca derlem çalışmaları, farklı dilbilimsel ve kültürel bağlamlarda birbirinden ayrışan gelişim yolları izlemiştir. İngilizce derlem araştırmaları güçlü kurumsal destek ve zengin dijital kaynaklardan yararlanırken, diğer dillerde derlem oluşturma çalışmaları bölgesel akademik gelenekler, teknik altyapılar ve araştırma öncelikleri doğrultusunda ilerlemiştir (McEnery & Hardie, 2012). Bu çeşitlilik, derlem evriminin belirli dilsel bağlamlar içinde ele alınmasının gerekliliğini ortaya koymaktadır. Bu çalışma, öncelikle İngilizceye ilişkin başlıca derlem çalışmalarını incelemekte; ardından Türkçe derlem araştırmalarına odaklanarak bunları kronolojik olarak yapılandırılmış bir tarihsel çerçeve içinde ele almaktadır.

İngilizce Derlem Çalışmaları

İngilizce derlem çalışmaları, büyük ölçekli veri kümelerinin oluşturulmasında öncü bir rol üstlenmiştir. Dilbilimsel araştırmalarda bilgisayarların yaygın biçimde kullanılmasından önce, Randolph Quirk'in liderliğinde 1959 yılında başlatılan *Survey of English Usage* (SEU) gibi kapsamlı veri toplama girişimleri, konuşma ve yazı diline ait özgün İngilizce verilerin dilbilgisel incelemeler için sağlanmasında kritik bir işlev görmüştür. Bununla birlikte, yöntemsel açıdan taşıdığı öneme rağmen SEU başlangıçta dijital ortamlar için tasarlanmamış olup, bu yönüyle elektronik öncesi döneme ait kapsamlı derlem çalışmalarının zirve noktasını temsil etmektedir (Greenbaum & Svartvik, 1990).

Brown Üniversitesi Günümüz Amerikan İngilizcesi Standart Korpusu (*Brown University Standard Corpus of Present-Day American English*), yaygın olarak bilinen adıyla Brown Korpusu, İngilizce derlem çalışmalarının tarihinde dönüm noktası niteliğinde bir önem taşımaktadır. 1961 yılında tasarlanan ve 1964 yılına gelindiğinde makine tarafından okunabilir biçime dönüştürülen Brown Korpusu, dilbilimsel araştırmalar için açık biçimde derlenmiş ilk elektronik derlem olma özelliğini taşımaktadır (Francis & Kučera, 1964). Bu derlemin geliştirilmesi, üretici dilbilgisinin baskın olduğu ve dil kuramlarının gözlemlenebilir dil kullanımından ziyade soyut yeterlilik kavramını merkeze aldığı bir döneme denk gelmiş; bu durum, derlem temelli ve sıklık odaklı yöntemlere yönelik kuşkucu yaklaşımların ortaya çıkmasına yol açmıştır (Leech, 1991). Buna karşın Brown Derlemi, sistematik biçimde örneklenmiş özgün metinlerin, dilbilimsel betimleme için güvenilir bir ampirik temel sunabileceğini ortaya koymuştur.

Brown Korpusu, her biri yaklaşık 2.000 sözcükten oluşan 500 metin örneğinden derlenmiş ve toplamda bir milyondan fazla sözcük içeren bir yapıdan oluşmaktadır. Derlem, 1961 yılında Amerika Birleşik Devletleri'nde yayımlanmış, farklı metin türlerini kapsayan geniş bir yelpazeden seçilmiş metinleri içermektedir. Bu derlem, modern basılı Amerikan İngilizcesinin temsiline odaklanan katmanlı rastgele örnekleme ilkeleri doğrultusunda oluşturulmuş olup üslup veya edebî nitelik açısından önceden belirlenmiş kuralcı ölçütlere dayanmamıştır (Kučera & Francis, 1967). Tür sınıflandırması, örnekleme teknikleri ve kodlama ölçütlerine ilişkin ayrıntılı belgelendirme, derlemin şeffaflığını ve uzun vadeli kullanılabilirliğini artırmış; böylece sonraki derlem hazırlama çalışmalarına yöntemsel ölçütler kazandırmıştır (Francis & Kučera, 1982).

Brown paradigmasını izleyerek geliştirilen Lancaster–Oslo/Bergen (LOB) Derlemi, İngiliz İngilizcesi için bir karşılık oluşturmak amacıyla 1970–1978 yılları arasında derlenmiştir. Brown Korpusuna benzer biçimde, 1961 yılında yayımlanmış metinlerden seçilen 500 örnek üzerinden oluşturulan ve yaklaşık bir milyon sözcük içeren bu derlem, aynı tür sınıflandırmalarını ve örnekleme yöntemlerini kullanmıştır (Johansson ve ark., 1978). İki derlem arasındaki güçlü yapısal benzerlik, Amerikan ve İngiliz İngilizcesinin sistematik biçimde karşılaştırılmasına olanak tanımış; böylece araştırmacılar yalnızca belirli dilsel yapıların varlığı ya da yokluğunu değil, aynı zamanda sözcüksel ve dilbilgisel sıklıklardaki farklılıkları da inceleyebilmiştir (Johansson, 1980).

Brown ve LOB derlemleri, Hint, Yeni Zelanda ve Avustralya İngilizcesi gibi farklı İngilizce varyantlarını temsil eden diğer birinci kuşak derlemlerin oluşturulmasında model işlevi görmüştür. Bu derlemler genellikle Brown/LOB çerçevesini izlemiş olsa da, yayımlama pratikleri ve kültürel bağlamlardaki farklılıklar belirli uyarlamaları zorunlu kılmış; bu durum, tek tip tür

sınıflandırmalarının farklı varyantlardaki dilsel gerçeklikleri yeterince temsil edebileceği varsayımının sınırlarını ortaya koymuştur (Bauer, 1993; Collins, 1991).

Yazılı derlemlerin yanı sıra, konuşma dili İngilizcesi de London–Lund Derlemi'nin (London–Lund Corpus, LLC) oluşturulmasıyla birlikte derlem çalışmalarında önemli bir araştırma alanı hâline gelmiştir. SEU'nun sözlü bileşeninden türetilen LLC, yaklaşık yarım milyon sözcükten oluşan yapısıyla, konuşma dili İngilizcesine yönelik ilk kapsamlı elektronik derlem olma özelliğini taşımaktadır (Greenbaum & Svartvik, 1990). Derlemde yer alan ayrıntılı ortografik ve prozodik ek açıklamalar, sesbilim, dilbilgisi, sözcük bilgisi ve söylem yapısı alanlarında öncü nitelikte çalışmaların yapılmasına olanak tanımıştır. Bununla birlikte, konuşmacı demografisi ve tür çeşitliliğine ilişkin sınırlılıklar, konuşma derlemlerinde tam anlamıyla temsil edilebilirliğin sağlanmasında karşılaşılan önemli güçlükleri de açıkça ortaya koymuştur.

Günümüz ölçütlerine göre görece sınırlı bir büyüklüğe sahip olmalarına rağmen, birinci kuşak İngilizce derlemler dilbilim alanı üzerinde kalıcı bir etki bırakmıştır. Bu derlemler, temsil edilebilirlik, şeffaflık ve geniş erişilebilirlik gibi temel derlem tasarım ilkelerini biçimlendirmiş ve dil araştırmalarında ampirik, veri temelli yaklaşımların uygulanabilirliğini somut biçimde ortaya koymuştur. Erken dönem İngilizce derlem çalışmaları, daha sonraki büyük ölçekli ve hesaplamalı açıdan gelişmiş derlemler için ampirik bir zemin hazırlamış; dilbilimsel araştırmalarda hâlen kullanılmalarıyla da görüldüğü üzere, çağdaş derlem temelli ve uygulamalı dilbilim çalışmalarını etkilemeye devam etmektedir (Leech, 1991; Biber, 1993).

1980'li yılların başlarından itibaren, bir milyon kelimelik derlemlerin (SEU ve Brown gibi) birçok sözcüksel ve anlamsal çözümleme için yetersiz olduğunun anlaşılmasıyla birlikte, derlem dilbilimi birinci kuşak derlemlerden çok daha büyük ölçekli derlemlere yönelmiştir (Leech, 1991). Metin yakalama ve depolama teknolojilerindeki gelişmeler, on milyonlarca hatta yüz milyonlarca sözcük içeren derlemlerin oluşturulmasını mümkün kılmıştır. Bu durum hem sözlükbilim hem de betimleyici dilbilim alanlarında köklü dönüşümler yaratmıştır.

COBUILD projesi, sözlükbilim odaklı, kapsamlı ve makine tarafından okunabilir bir derlem girişimi olarak öne çıkmaktadır. 1980 yılında Collins ile Birmingham Üniversitesi arasında bir iş birliği kapsamında başlatılan bu proje, *Birmingham Collection of English Text* temelinde geliştirilmiş ve öğrenciler, eğitimciler ile araştırmacılar açısından güncel İngilizce kullanımını yansıtmayı amaçlamıştır (Renouf, 1987). Açıkça tanımlanmış denge ilkeleri doğrultusunda oluşturulan Ana COBUILD Derlemi, verinin yaklaşık

%25'inin sözlü dilden oluşması, teknik dilden ziyade genel dile ağırlık verilmesi ve 1960 sonrası doğal olarak üretilmiş metinlerin tercih edilmesi gibi sebeplerden dolayı kısa sürede birkaç milyon sözcüğe ulaşmış; daha büyük bir *Reserve Corpus* ve bir TEFL alt derlemi ile desteklenmiştir. Bu girişim, derlem hazırlama süreçlerini derlem temelli sözlükler, dilbilgisi ve öğretim materyalleri gibi ticari ürünlerle yenilikçi bir biçimde ilişkilendirmiş; ilerleyen dönemde yüz milyonlarca sözcük içeren dinamik bir izleme derlemi (*monitor corpus*) olan *Bank of English*'e dönüşmüştür (Blackwell, 1993).

Aynı zamanda, diğer ticari-akademik iş birlikleri de kendi kapsamlı metin derlemelerini oluşturdular. Longman/Lancaster İngilizce Dil Korpusu (Longman/Lancaster English Language Corpus),, Longman Konuşma Korpusu (Longman Spoken Corpus) ve Longman Corpus of Learners' English derleminden oluşan Longman Corpus Network, sözlük oluşturma ve tanımlama için kapsamlı bir 20. yüzyıl İngilizce veri tabanı olarak hazırlandı. Bu korpus önde gelen ana dili İngilizce olan lehçelerin hem sözlü hem de yazılı biçimlerinden yaklaşık 50 milyon kelimeyi kapsamaktadır (Summers, 1991).

1990'lı yılların başındaki en iddialı derlem girişimi ise British National Corpus (BNC) olmuştur. BNC, modern İngiliz İngilizcesini temsil eden dengeli bir yapı içinde toplam 100 milyon sözcükten oluşmakta; bunun yaklaşık 10 milyon sözcüğü sözlü dil verilerinden meydana gelmektedir. Derlem, farklı alanlar, medya türleri, zaman dilimleri ve stilistik düzeyleri kullanılarak sistematik ve çok katmanlı bir örnekleme yaklaşımıyla oluşturulmuştur (Crowdy, 1993).

International Corpus of English (ICE), İngilizcenin bölgesel varyantlarının küresel ölçekte karşılaştırmalı olarak incelenmesini mümkün kılmak amacıyla oluşturulmuştur (Greenbaum, 1991, 1996). ICE projesi, her biri yaklaşık bir milyon sözcükten oluşan ve birbirleriyle karşılaştırılabilir nitelikte tasarlanan en fazla 20 alt derlemin bir araya getirilmesini hedeflemiştir. Bu alt derlemler, en az ortaöğretim düzeyine kadar İngilizce dilinde resmî eğitim almış, 18 yaş ve üzerindeki bireylerin kullandığı İngilizceyi temsil etmeyi amaçlamaktadır.

Veriler, Birleşik Krallık, Amerika Birleşik Devletleri, Kanada, Avustralya ve Yeni Zelanda gibi İngilizcenin yaygın ya da birincil dil olduğu ülkelerin yanı sıra, Hindistan, Nijerya, Singapur ve Karayipler gibi İngilizcenin ek bir resmî dil ya da yaygın olarak kullanılan bir ikinci dil konumunda bulunduğu bölgelerden toplanmaktadır (Leech, 1991).

Türkçe Derlem Çalışmaları

Sistematik Türkçe derlem temelli çalışmalar, İngilizce ve diğer önde gelen Avrupa dillerine kıyasla görece daha geç bir dönemde gelişmiştir. Bu-

nunla birlikte, 1990'lı yıllardan itibaren Türkçe derlem çalışmaları, yöntemsel açıdan çeşitlenen ve giderek daha yapılandırılmış bir araştırma alanı hâline gelerek istikrarlı bir gelişim göstermiştir.

1990'lı yıllarda Orta Doğu Teknik Üniversitesi'nde geliştirilen METU Turkish Corpus, Türkçe üzerine dilbilimsel ve hesaplamalı araştırmalar için özel olarak tasarlanmış, modern ve bilgisayar tarafından okunabilir ilk derlem niteliğini taşımaktadır. Kemal Oflazer'in öncülüğünde oluşturulan bu derlem, Türkçe üzerine yürütülen erken dönem doğal dil işleme (NLP) araştırmalarıyla yakından ilişkili olup, özellikle eklemeli bir dil olan Türkçenin biçimbilimsel çözümlemesini ve hesaplamalı olarak modellenmesini desteklemeyi amaçlamıştır (Oflazer, 1994).

Sınırlı hacmine rağmen METU Turkish Corpus, daha sonraki Türkçe derlem projeleri için teknik ve kavramsal bir temel oluşturan derlem temelli yöntemleri ve ek açıklama (annotation) ilkelerini yerleştirmiştir. METU Turkish Corpus için bir *Corpus Workbench*'in geliştirilmesi gibi yazılım odaklı katkılar ise, derlem işleme ve çözümleme araçları sunarak bu altyapıyı daha da güçlendirmiştir (Say ve ark., 2004).

2000'li yılların başlarında, bu temel altyapı üzerine inşa edilen Turkish Treebank'in (2002–2003) oluşturulmasıyla birlikte önemli bir yöntemsel ilerleme kaydedilmiştir. Oflazer, Say ve Zeyrek tarafından geliştirilen bu kaynak, Türkçe için sistematik biçimde sözdizimsel olarak ek açıklama yapılmış ilk derlem olma özelliğini taşımaktadır (Say ve ark., 2002). Turkish Treebank, tümce yapısı ve sözdizimsel bilgiyi bütünleştirerek Türkçe sözdiziminin kapsamlı biçimde bilgisayar destekli olarak çözümlenmesine olanak sağlamış; böylece yalnızca biçimbilimsel düzeyde tasarlanmış derlemlerden yapısal olarak ek açıklamalı veri kümelerine doğru bir geçişi temsil etmiştir. Bu kaynak, ilerleyen süreçte Türkçe için bağımlılık temelli (dependency-based) sözdizimsel modellerin geliştirilmesinde önemli ölçüde etkili olmuştur.

2000'li yılların ortalarında, doğal dil işleme (NLP) odaklı deneysel ve görev-özgül derlemlerin ortaya çıkmasıyla birlikte Türkçe derlem araştırmaları önemli ölçüde genişlemiştir. Temsil edilebilirlik ya da dengeli kapsam sağlamayı hedeflemek yerine, bu veri kümeleri biçimbilimsel belirsizliğin giderilmesi, sözcük anlamı belirsizliği çözümü ve anlamsal rol etiketleme gibi belirli hesaplamalı sorunları ele almak amacıyla oluşturulmuştur. Bu dönemin önemli katkılarından biri, Sak vd. (2008) tarafından yapılan *Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus* başlıklı çalışmadır. Söz konusu çalışma, gelişmiş biçimbilimsel işleme araçlarıyla birlikte kapsamlı bir web derlemi sunmuş; bu kaynaklar özellikle makine çevirisi ve bilgi çıkarımı araştırmalarında yaygın biçimde kullanılmıştır.

2010'lu yılların başlarında daha kapsamlı ve temsil gücü yüksek derlemlerin oluşturulmasıyla birlikte önemli bir dönüm noktası yaşanmıştır. Zeyrek vd., 2012 yılında yayımladıkları *Development of a Corpus and a Treebank for Present-Day Written Turkish* başlıklı çalışmalarında, çağdaş Türkçe derlemlerin daha gelişmiş ek açıklama şemaları ve sözlü dil verileriyle zenginleştirilmesi gerekliliğini vurgulamış; böylece Türkçe derlem geliştirme çalışmalarını küresel karşılaştırmalı bir bağlam içine yerleştirmiştir.

Türkçe için derlem temelli kaynakların oluşturulması, ulusal düzeyde eşgüdümlü derlemlerin hayata geçirilmesiyle birlikte daha da ivme kazanmış ve Türkçe derlem araştırmaları daha geniş bir uluslararası araştırma bağlamında konumlanmıştır. Turkish National Corpus'un (TNC) kullanıma sunulması, Türkiye'de derlem temelli dil çalışmalarında önemli bir aşamayı temsil etmektedir. Aksan vd. tarafından geliştirilen TNC, çağdaş yazılı Türkçeyi çok sayıda tür ve konu alanı üzerinden kapsayan, dengeli ve temsil gücü yüksek bir başvuru derlemi niteliğindedir. Uluslararası kabul görmüş derlem tasarım standartlarına uygun olarak oluşturulan bu derlem, dilbilimsel, sözlükbilimsel, biçimsel ve pedagojik araştırmalar için standartlaştırılmış bir ampirik zemin sunmakta; Türkçe derlem çalışmalarını diğer dillerdeki ulusal başvuru derlemleriyle ilişkilendirmektedir (Aksan ve ark., 2012).

2015–2018 yılları arasında, Türkçeyi yabancı dil olarak öğrenen bireylerin yazılı üretimlerini derleyen Türkçe öğrenici derlemlerinde kayda değer bir genişleme yaşanmıştır. Bu derlemler, ara dil (interlanguage) çözümlemeleri, hata etiketleme çalışmaları ve derlem temelli dil öğretimi için ampirik veriler sağlamıştır. Bu dönemde Sezer'in (2017) TS Corpus Project adlı çalışması, kullanıcıların çevrim içi bir Türkçe sözlükle bağlantılı kişiselleştirilmiş derlemler oluşturmasına olanak tanıyarak, kullanıcı odaklı ve özelleştirilebilir derlem oluşturma açısından önemli bir yenilik ortaya koymuştur.

2016 yılında ITU Turkish Treebank'in, Universal Dependencies (UD) çerçevesiyle birlikte yayımlanması bir diğer önemli dönüm noktasını oluşturmuştur. Sulubacak vd. tarafından geliştirilen bu kaynak, Türkçeyi çok dilli bir ek açıklama standardı içine dâhil ederek, diller arası karşılaştırmalı çözümlemeleri ve çok dilli doğal dil işleme (NLP) araştırmalarını mümkün kılmıştır (Sulubacak ve ark., 2016; Nivre ve ark., 2016).

2016–2019 yılları arasında, Türkçe paralel derlemlerin OPUS platformuna dâhil edilmesiyle birlikte Türkçe, çeviri teknolojileri araştırmalarında artan bir önem kazanmıştır. Tiedemann tarafından derlenen bu tümce düzeyinde hizalanmış iki dilli veri kümeleri, hem istatistiksel hem de nöral makine çevirisi sistemleri için kritik eğitim kaynakları olarak işlev görmüştür (Tiedemann, 2012). BOUN Treebank'in kullanıma sunulması ise, ileri düzey doğal dil işleme araştırmalarına uygun, yüksek kaliteli ve elle gözden geçirilmiş

biçimbilimsel ve sözdizimsel ek açıklamalar sağlayarak Türkçe için mevcut sözdizimsel kaynakları önemli ölçüde zenginleştirmiştir (Türk ve ark., 2021).

Türkçe derlem çalışmaları, 1990'lı yıllarda teknik odaklı ve görece sınırlı veri kümeleriyle başlayan bir süreçten, ulusal, özel amaçlı, öğrenici, paralel ve web tabanlı derlemleri kapsayan kapsamlı ve çeşitlenmiş bir ekosisteme doğru evrilmiştir. Bu tarihsel gelişim çizgisi, Türkçe derlem dilbiliminin olgunlaşmasını ve kuramsal dilbilimsel incelemeler ile doğal dil işleme ve makine çevirisi gibi veri odaklı uygulamalarla olan yakın ilişkisini açık biçimde ortaya koymaktadır.

Çeviribilimde Derlem Kullanımı

Derlem dilbiliminin Çeviribilim alanına dâhil edilmesi, ampirik ve veri temelli araştırmalara yönelen önemli bir yöntemsel ve kuramsal dönüşümü beraberinde getirmiştir. 1980'li yılların sonları ve 1990'lı yılların başlarından itibaren derlemler, çeviri araştırmacılarına çevrilmiş ve çevrilmemiş özgün metinlerden oluşan geniş koleksiyonlara sistematik erişim olanağı sunmuştur. Böylece çevirinin kuramsal olarak nasıl olması gerektiğinden ziyade, pratikte nasıl gerçekleştirildiğinin gözlemlenmesini ve çözümlenmesini mümkün kılmıştır. Bu gelişme, çevirileri hedef kültür ve hedef dizge içinde tarihsel ve kültürel olarak bağlamlandırılmış ürünler olarak incelemeyi önceleyen Betimleyici Çeviri Çalışmaları (Descriptive Translation Studies, DTS) yaklaşımıyla yakından ilişkili olup, derlem temelli yöntemler bu bakış açısının uygulanması için elverişli bir ampirik çerçeve sağlamıştır (Toury, 1995).

Derlemlerin Çeviribilim alanına sağladığı temel katkılardan biri, özellikle paralel ve benzer derlemler olmak üzere farklı derlem türleri arasındaki ayrımın netleşmesidir. Paralel derlemler kaynak metin ile çevirisinin hizalanmasıyla oluşmaktadır ve araştırmacılara çeviri sürecinde ortaya çıkan çeviri farklılıklarını, eşleşmeleri ve kullanılan çeviri tekniklerini farklı dilsel düzeylerde incelemelerine olanak tanır. Buna karşılık, karşılaştırılabilir derlemler genellikle aynı dilde yer alan çevrilmiş ve özgün (çevrilmemiş) metinleri, benzer tasarım ölçütleri doğrultusunda bir araya getirmekte ya da benzer iletişimsel bağlamlarda farklı dillerde bağımsız olarak üretilmiş metinlerden oluşmaktadır (Zanettin, 2012).

Derlem Temelli Çeviri Çalışmaları (Corpus-based Translation Studies, CBTS), Baker'ın (1993, 1996) çeviri evrenselleri kavramını ortaya koymasının ardından belirgin bir önem kazanmıştır. Çeviri evrenselleri, karşılaştırılabilir derlemler içinde çevrilmiş metinlerde, çevrilmemiş metinlere kıyasla ortaya çıkması beklenen yineleyen örüntüler olarak tanımlanmaktadır. Basitleştirme (simplification), açıklık kazandırma (explicitation), standartlaşma (standardization) ve dengeleme (leveling) gibi eğilimler, mutlak ya da belirleyici özellikler olarak değil; aksine ampirik olarak sınanabilir örüntüler

olarak ele alınmıştır. Laviosa'nın (2002) araştırmaları gibi bunu izleyen derlem temelli çalışmalar bu varsayımları çevrilmiş ve çevrilmemiş metinlerin sistematik karşılaştırmaları yoluyla incelemiş ve belirli dil çiftleri, türler ve derlemler genelinde tutarlı sözcüksel ve üslupsal farklılıklar ortaya koymuştur. Evrensellik kavramı zaman içinde kapsamlı eleştirilere konu olmuş olsa da, derlem temelli yöntemler bu tür varsayımların değerlendirilmesi, geliştirilmesi ya da sorgulanması için sağlam bir ampirik zemin sunmuş; böylece çeviribilim araştırmalarının içgözleme ya da anekdot niteliğindeki kanıtlara dayanmaktan öteye taşınmasına önemli katkı sağlamıştır.

Genel örüntülerin incelenmesinin yanı sıra, derlemler çeviri üslubunun çözümlenmesinde de kritik bir rol oynamıştır. Derlem temelli çözümlemeler, bireysel çevirmenlerin dilsel seçimlerinde tutarlı ve bütünlüklü örüntüler sergilediğini; bu örüntülerin yalnızca kaynak metnin etkisiyle açıklanamayacağını ortaya koymaktadır (Baker, 2000). Paralel ve karşılaştırılabilir derlemlerin birlikte kullanılması sayesinde araştırmacılar, "çevirmen parmak izi" (*translator fingerprints*) olarak adlandırılan bu özellikleri saptayabilmekte; böylece bireysel habitus, kültürel normlar ve metinsel kısıtlamalar arasındaki etkileşime ilişkin önemli içgörüler elde edebilmektedir. Bu tür çalışmalar, çevirmeni üslupsal etkisi metinler boyunca ampirik olarak belirlenebilen bir dilsel özne olarak merkeze alarak, Çeviribilim alanının kapsamını genişletmiştir.

Derlemler, hem çeviri pratiğinde hem de çevirmen eğitiminde giderek daha merkezi bir rol üstlenmiştir. Mesleki bağlamda çeviri bellekleri, önceki çeviri bölümlerinin yeniden kullanılması ve hizalanması yoluyla zaman içinde gelişen, dinamik paralel derlemler olarak işlev görmektedir. Eğitim ortamlarında ise derlem temelli yöntemlerin, öğrenen kişilerin özerkliğini ve eleştirel farkındalığı artırmada etkili olduğu gösterilmiştir. Bu yaklaşımlar, öğrencilerin çeviri çözümlerini ve hedef dildeki kullanım uzlaşılarını kuralcı normlara dayanmak yerine ampirik verilere dayalı olarak incelemelerine olanak tanımaktadır (Zanettin ve ark., 2003). Tüm bu gelişmeler, derlemlerin yalnızca araştırma amaçlı kaynaklar olmadığını; aynı zamanda çağdaş çeviri pratiğini ve çevirmen eğitimini güçlendiren temel araçlar olarak işlev gördüğünü ortaya koymaktadır.

Sonuç

Bu çalışma, derlem dilbiliminin tarihsel gelişimi ile makine çevirisinin ilerleyişinin birbirinden bağımsız değil, aksine temelde birbiriyle iç içe geçmiş süreçler olduğunu ortaya koymuştur. Derlemler, başlangıçta betimleyici dilbilimsel çözümlemeler için dijital araçlar kullanılmadan derlenmiş koleksiyonlar iken, zamanla büyük ölçekli, dijital olarak hizalanmış paralel derlemlere evrilmiş; böylece dil verilerinin pasif depoları olmaktan çıkarak çağdaş

çeviri teknolojilerini destekleyen etkin altyapı bileşenleri hâline gelmiştir. Bu dönüşüm, özellikle nöral makine çevirisi sistemleri ve büyük dil modelleriyle belirginleşen veri odaklı paradigmaların ortaya çıkışında açıkça görülmektedir; zira bu sistemlerin etkinliği, doğası gereği, eğitimde kullanılan derlemlerin ölçeği, niteliği ve temsil gücüyle doğrudan bağlantılıdır.

Bu çalışma, başlıca İngilizce derlem projelerinin evrimini inceleyerek ve Türkçe derlem çalışmalarını benzer bir tarihsel çerçeve içinde bağlamsallaştırarak, derlem üretiminde hem evrensel yöntemsel örüntüleri hem de dile özgü gelişim yollarını ortaya koymuştur. İngilizce derlem araştırmaları erken dönemde kurumsal destek ve teknolojik kaynaklardan yararlanmış olsa da Türkçe derlem çalışmaları hesaplamalı, dilbilimsel ve ek açıklama odaklı kaynakların kademeli birikimi yoluyla gelişmiş ve nihayetinde uluslararası standartlarla karşılaştırılabilir bir yöntemsel olgunluğa ulaşmıştır. Bu eğilim, derlem geliştirmenin teknolojik ilerlemeler, dil tipolojisi, araştırma gündemleri ve akademik ekosistemler tarafından belirlendiğini göstermektedir.

Derlem yaklaşımlarının çeviri çalışmaları alanına dâhil edilmesi, alandaki ampirik yönelimi güçlendirmiş; çeviri süreçlerinin, çeviri normlarının ve çevirmen üslubunun sistematik biçimde incelenmesini mümkün kılmıştır. Güncel çeviri uygulamalarında derlemler, yalnızca analitik araçlar olarak değil, aynı zamanda bilgisayar destekli çeviri araçları, çeviri bellekleri ve makine çevirisi sistemlerine entegre edilmiş işlevsel kaynaklar olarak da kullanılmaktadır. Çeviri süreçlerinin giderek artan biçimde post-editing ve insan–makine iş birliğine dayanmasıyla birlikte, derlemler çeviri sürecinde hem girdi hem de çıktı işlevi görmekte olup profesyonel uygulamalar yoluyla sürekli olarak genişlemektedir.

Sonuç olarak, çağdaş çeviri teknolojilerinin ve uygulamalarının anlaşılabilmesi, derlem dilbilimi ile makine çevirisinin birlikte evrimini dikkate alan tarihsel açıdan bilinçli bir bakış açısını gerekli kılmaktadır. Veri odaklı yaklaşımların hem akademik araştırmalarda hem de sektörel uygulamalarda giderek baskın hâle gelmesiyle birlikte, gelecekteki çalışmaların derlem niteliğine, etik veri toplama süreçlerine ve dile özgü kaynakların geliştirilmesine öncelik vermesi büyük önem taşımaktadır. Bu doğrultuda, çeviri teknolojilerindeki ilerlemelerin hem dilbilimsel açıdan sağlam hem de toplumsal sorumluluk bilinciyle şekillendirilmesi güvence altına alınabilecektir.

Kaynakça

Aksan, Y., Aksan, M., Koltuksuz, A., Sezer, T., Mersinli, Ü., & Ufuk, U. (2012). *Construction of the Turkish National Corpus (TNC).*

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). John Benjamins.

Baker, M. (1996). *Corpus-based translation studies: The challenges that lie ahead.* John Benjamins.

Baker, M. (2000). Towards a methodology for investigating the style of a literary translator. *Target, 12*(2), 241–266.

Bauer, L. (1993). *Manual of information to accompany the Wellington Corpus of Written New Zealand English.* Victoria University of Wellington.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243–257. https://doi.org/10.1093/llc/8.4.243

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge University Press.

Blackwell, S. (1993). From dirty data to clean language. In J. Aarts et al. (Eds.), *English corpus linguistics* (pp. 97–106).

Collins, P. (1991). *Cleft and pseudo-cleft constructions in English.* Routledge.

Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing, 8*(4), 259–265.

Francis, W. N., & Kučera, H. (1964). *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers.* Brown University.

Francis, W. N., & Kučera, H. (1967). *Computational analysis of present-day American English.* Brown University Press.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Houghton Mifflin.

Granger, S. (2003). The corpus approach: A common way forward for contrastive linguistics and translation studies. In S. Granger, J. Lerot, & S. Petch-Tyson (Eds.), *Corpus-based approaches to contrastive linguistics and translation studies* (pp. 17–29). Rodopi.

Greenbaum, S. (1991). The development of the International Corpus of English. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 83–91). Longman.

Greenbaum, S. (Ed.). (1996). *Comparing English worldwide: The International Corpus of English.* Clarendon Press.

Greenbaum, S., & Svartvik, J. (1990). *The London–Lund Corpus of Spoken English.* Lund University Press.

Hutchins, W. J., & Somers, H. L. (1992). *An introduction to machine translation.* Academic Press.

Johansson, S. (1980). Some aspects of the vocabulary of learned and scientific English. In S. Greenbaum, G. Leech, & J. Svartvik (Eds.), *Studies in English linguistics for Randolph Quirk* (pp. 123–147). Longman.

Johansson, S., Leech, G., & Goodluck, H. (1978). *Manual of information to accompany the Lancaster–Oslo/Bergen Corpus of British English.* Norwegian Computing Centre for the Humanities.

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

Koehn, P. (2010). *Statistical machine translation.* Cambridge University Press. https://doi.org/10.1017/CBO9780511815829

Laviosa, S. (2002). *Corpus-based translation studies: Theory, findings, applications.* Rodopi.

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 8–29). Longman.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice.* Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction* (2nd ed.). Edinburgh University Press.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., … Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. *Proceedings of LREC 2016*, 1659–1666.

O'Brien, S. (2012). Translation as human–computer interaction. *Translation Spaces, 1*(1), 101–122. https://doi.org/10.1075/ts.1.06obr

O'Brien, S., Balling, L. W., Carl, M., Simard, M., & Specia, L. (2014). *Post-editing of machine translation: Processes and applications.* Cambridge Scholars Publishing.

Oflazer, K. (1994). Two-level description of Turkish morphology. *Literary and Linguistic Computing, 9*(2), 137–148. https://doi.org/10.1093/llc/9.2.137

Renouf, A. (1987). Corpus development. In J. Sinclair (Ed.), *Looking up* (pp. 1–40). Collins ELT.

Quirk, R. (1968). The Survey of English Usage. In *Essays on the English language: Medieval and modern* (Chap. 7). Longman.

Sak, H., Güngör, T., & Saraçlar, M. (2008). Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *Proceedings of the International Conference on Natural Language Processing* (pp. 417–427). Springer.

Say, B., Zeyrek, D., Oflazer, K., & Özge, U. (2002). Development of a corpus and a treebank for present-day written Turkish. In *Proceedings of the 11th International Conference of Turkish Linguistics* (pp. 183–192).

Say, B., Zeyrek Bozşahin, D., Oflazer, K., & Özge, U. (2004). *Development of a corpus and a treebank for present-day written Turkish*. https://hdl.handle.net/11511/81526

Sezer, T. (2017). TS Corpus Project: An online Turkish dictionary and TS DIY corpus. *European Journal of Language and Literature, 9*, 18–24. https://doi.org/10.26417/ejls.v9i1.18

Shao, X., & Zhu, Y. (2025). The assistance role of LLMs and NMT in student translators' Chinese–English post-editing. *Journal of China Computer-Assisted Language Learning.*

Sinclair, J. M. (1991). *Corpus, concordance, collocation.* Oxford University Press.

Summers, D. (1991). *Longman/Lancaster English Language Corpus: Criteria and design.* Longman.

Sulubacak, U., Gökırmak, M., Tyers, F., Çöltekin, Ç., Nivre, J., & Eryiğit, G. (2016). Universal dependencies for Turkish. In *Proceedings of COLING 2016* (pp. 3444–3454).

Tiedemann, J. (2011). *Bitext alignment.* Morgan & Claypool. https://doi.org/10.2200/S00320ED1V01Y201102HLT014

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *Proceedings of LREC 2012*, 2214–2218.

Toury, G. (1995). *Descriptive translation studies and beyond.* John Benjamins.

Türk, U., Atmaca, F., Özateş, Ş. B., Berk, G., Bedir, S. T., Köksal, A., Öztürk, B., Güngör, T., & Özgür, A. (2021). Resources for Turkish dependency parsing: Introducing the BOUN Treebank and the BoAT Annotation Tool. *Language Resources and Evaluation*, 1–49.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., … Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv*. https://arxiv.org/abs/1609.08144

Zanettin, F. (2012). *Translation-driven corpora: Corpus resources in descriptive and applied translation studies.* St. Jerome.

Zanettin, F., Bernardini, S., & Stewart, D. (Eds.). (2003). *Corpora in translator education.* St. Jerome.

Zeyrek, D., Demirşahin, I., Sevdik-Çallı, A., Bayezit, E., & Yücebaş, E. (2012). Development of a corpus and a treebank for present-day written Turkish. *Proceedings of LREC 2012*, 1057–1063