JAPONCA ÜZERİNE YAPILMIŞ BİR DERLEM ÇALIŞMASI: ÇAĞDAŞ JAPONCA YAZI DİLİNİN DENGELENMİŞ DERLEMİ (BCCWJ)*

Mehmet YILDIRIM**
Oğuzhan ATİLA***

Öz: Bu çalışmada, Ulusal Japonca Araştırmaları Enstitüsü (NINJAL) bünyesindeki Derlem Geliştirme Merkezi tarafından hazırlanan derlemlerden biri olan Çağdaş Japonca Yazı Dilinin Dengelenmiş Derlemi (BCCWJ) tanıtılıp hazırlanma aşamaları ve yöntemi hakkında bilgi verilmiştir. Dünyada farklı dillerle ilgili derlem çalışmaları yapılmaktadır. Bunların yöntemleri genel olarak birbirine benzemekle birlikte derlemin hazırlanma amacına göre faklılıklar da bulunmaktadır. BCCWJ, dengelenmiş bir derlem oluşturma amacıyla hazırlandığı için, derlemin tasarımından örneklem seçimine kadar her aşamada bu amaç doğrultusunda hareket edilmiştir. Buna göre BCCWJ'nin en önemli özelliği yayın alt derlemi, kütüphane alt derlemi ve özel amaçlı alt derlem olmak üzere toplam üç alt derlemden oluşması; örneklem seçiminde ise sabit uzunluktaki metin ve değişken uzunluktaki metin şeklinde iki tür metin kullanılmasıdır. Beş yıllık bir süreçte hazırlanan ve yaklaşık 105 milyon kelimeden oluşan bu derlem, Türkçe için hazırlanacak benzer derlemler için örnek alınabilir.

Anahtar kelimeler: Derlem, Japonca derlem, yazı dili derlemi, dengelenmiş derlem, BCCWJ

* Bu çalışma, Japonya'daki Derlem Geliştirme Merkezinin İnternet sitesinden yararlanılarak hazırlanmıştır.

** Okutman, Gazi Üniversitesi TÖMER
*** Okutman, Gazi Üniversitesi TÖMER

A CORPUS STUDY IN JAPANESE LANGUAGE: THE BALANCED CORPUS OF CONTEMPORARY WRITTEN JAPANESE (BCCWJ)*

Mehmet YILDIRIM**
Oğuzhan ATİLA***

Abstract

This study describes and informs about the design process and methods of The Balanced Corpus of Contemporary Written Language (BCCWJ), one of the corpora designed by the Center for Corpus Development within National Institute for Japanese Language and Linguistics (NINJAL). There are studies on corpora in different languages around the world. Though the methods of such studies are alike in general, there are also differences depending on the purpose of the corpus design. Since BCCWJ was designed in order to create a balanced corpus, the whole process from the design of the corpus to the selection of samples was carried out in accordance with this purpose. Accordingly, the most significant feature of BCCWJ is that it consists of three sub-corpora: publication sub-corpus, library sub-corpus and special-purpose sub-corpus. Besides, it makes use of two types of sampling: fixed length samples and variable length samples. This corpus, which was prepared in a five-year period and consists of 105 million words, can be taken as an example to design similar corpora in Turkish language.

Key words: Corpus, Japanese corpus, corpus of written language, balanced corpus, BCCWJ

Giriş

Dille ilgili araştırmalarda o dili temsil edebilecek kapsamlı, doğru ve tutarlı bir kaynağa ihtiyaç duyulmaktadır. 1960'lı yıllardan itibaren kullanılmaya başlanan derlemlerin bu ihtiyacı karşılamada önemli bir yeri vardır. 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 (Sinclair, 2005).

* This study was carried out in line with the website of Centre for Corpus Development in Japan.
** Instructor, Gazi University TOMER
*** Instructor, Gazi University TOMER

Derlem temelli çalışmaların yaygınlaşmasıyla birlikte, dilbilimin bir alt dalı olarak derlem dilbilim ortaya çıkmıştır. Derlem dilbilim, doğal dil örneklemleri aracılığıyla dilin incelendiği alandır. Bir dil kuramı olmayıp dilbilimin bir dalıdır ve doğal dil verilerinin nicel ve nitel analizini yapar. Analizler genel olarak bilgisayar üzerinde özel olarak tasarlanmış yazılım programları kullanılarak yürütülür (Müller & Waibel, t.y).

Kuramsal ve bilgisayarlı dilbilim araştırmalarında derlem temelli yaklaşım en çok ihtiyaç duyulan ve en üretken teknik olarak görülmektedir (Startvik, 1992; Church ve Mercer, 1993). Bunun etkisi; konuşmayı işleme, bilgiye erişme, sözlük bilgisi, karakter tanımlama gibi doğal dil çalışmalarının hemen hemen her alanında kendini göstermektedir (Chen, Huang, Chang & Hsu, 1996). Sinclair (1987)'e göre bir derlem, belirli bir dilin örneklemi niteliğindedir ve o dille ilgili çok miktarda veri içerir. İyi bir derlemin, ait olduğu dili kapsamlı bir şekilde temsil edebilecek nitelikte ve dengelenmiş olması gerekir.

Dil araştırmalarında derlem kullanmanın en önemli gerekçesi, doğal dil verilerini sunmasıdır. Fakat Chomsky'e göre dilbilimin amacı doğal dildeki verileri sayısallaştırarak tanımlamak değil, sezgiyle açıklamaktır; çünkü doğal dillerde veriler ölçülebilir sınırlılıkta değildir, yani verilerin sayısallaştırılması bir dili tanımlamada asla yeterli olmayacaktır (McEnery & Wilson, 2001, s.11). Chomsky bu eleştirisiyle dilbilimde derlem temelli yöntemlere karşı olduğunu belirtmektedir. Fakat günümüzde bilgisayar yazılımlarındaki gelişmelerin de etkisiyle derlem dilbilimi çalışmaları hızla yaygınlaşmaktadır.

Derlemler, bir dilin özelliklerini tanıtmak, farklı dil yapılarında oluşturulmuş olan hipotezleri test etmek için kullanılabilir. Mesela dil öğrenenlerin farklı öğrenme aşamalarını (başlangıç, orta, ileri düzey) kayıt altına alan derlemler, yabancı dil öğretimiyle ilgili araştırmalarda kullanılabilir; tarihî derlemler, belirli dil özelliklerinin gelişimini takip etmeyi mümkün kılmaktadır; yine derlemler, toplum dilbilim ya da söylem çözümlemesi araştırmalarında belirli yaş gruplarının toplum dilbilimsel bazı özelliklerinin incelenmesinde kullanılabilir (Müller & Waibel, t.y).

Genel derlemlerden başka; yazı dili derlemi, konuşma dili derlemi, tarihî derlem, web derlemi, öğrenici derlemi gibi amaca göre bir dilin hemen hemen bütün kullanım alanlarında derlemler hazırlanabilmektedir. Türkçe üzerine de çeşitli derlem çalışmaları yapılmıştır. Mersin Üniversitesi akademisyenlerinden oluşan bir ekip tarafından, genel amaçlı bir derlem olan Türkçe Ulusal Derlemi (TUD) (Aksan vd., 2012); ODTÜ'de, ODTÜ Türkçe Derlem (Say, Zeyrek, Oflazer & Özge, 2002) ve Sözlü Türkçe Derlemi (STC) (Ruhi vd., 2010); Çukurova Üniversitesinde Eski Türkçe ve Karahanlı Türkçesinin Tarihsel Derlemi (Çetin & Özkan, 2011) hazırlanmıştır. Türkiye Türkçesinin Yazın Dili Derlemi (Özkan, 2011)'ni hazırlayan Bülent Özkan, Türk Çocuk Yazını Derlemi'ni de hazırlama aşamasındadır. Japonya'da, Tokyo Yabancı Diller Üniversitesinde hazırlanan Çok Dilli Sözlü Derlem'de Türkçe de yer almaktadır (Yılmaz, 2012). Bunların yanı sıra Türkçe üzerine farklı derlem çalışmaları yapılmaya devam etmektedir.

Derlemler esas olarak dilbilim araştırmalarında kullanılmakla birlikte günümüzde dille doğrudan ya da dolaylı olarak bağlantılı farklı alanlarda da derlemlerin kullanıldığı görülmektedir. Derlemlerin kullanım alanlarından bazıları Tablo 1'de görülebilir (NINJAL, 2016).

Tablo 1 *Derlemlerin Bazı Kullanım Alanları*

Alan	Kullanım Amacı
Dil Araştırmaları	Dilbilim, ana dili vb. farklı dil araştırmaları Farklı dillere ait derlemlerin kıyaslandığı karşılaştırmalı dilbilim çalışmaları
Bilgi İşleme	Ses tanıma için dil modelleme, akustik modeller oluşturma, dil işleme için model oluşturma, ses birimlerinin kontrolüne dayalı ses sentezi
Dil Eğitimi	Yabancı dil öğretiminde materyal geliştirme Ana dil öğrenimine destek sağlama
Dil Politikası	Resmî yazışma dilinde kaydedilmiş veriler
Sözlük Bilim	Örnek arama, kelimeler arasındaki bağlantıları anlama
Psikoloji	Dil üzerinde deney tasarlama, uyarıcı kontrolü sağlama

http://pj.ninjal.ac.jp/corpus_center/guidance.html sayfasından erişilmiştir.

Bilgisayar teknolojilerinde yaşanan hızlı gelişmelere bağlı olarak tüm dünyada, derlem hazırlama ve buna bağlı olarak da derlem dilbilim çalışmaları hızla artmaktadır. Derlem çalışmaları yönünden Batı'da özelikle İngiltere ön plana çıkarken Uzak Doğu ülkelerinde de önemli çalışmalar yapılmaktadır. Bu ülkelerin önde gelenlerinden biri olan Japonya, hem derlem oluşturma hem de derlem işleme yazılımlarını geliştirme yönünden önemli gelişmeler göstermiştir; hatta derlem çalışmalarını ulusal çapta yürütmek için akademik bir birim olan "Derlem Geliştirme Merkezi"ni kurmuştur. Bu durumuyla Japonya'nın derlem çalışmalarında Türkiye'den daha ileri bir konumda olduğu söylenebilir. Bunda Japonya'nın teknoloji üreten bir ülke olması, bilgisayar alanında yetişmiş insan gücüne sahip olması, ulusal projelere gerekli maddi desteği rahatlıkla sağlayabilecek imkâna sahip olması, dünyadaki gelişmeleri çok yakından takip ederek bunları ülkesi için uyarlaması gibi özelliklerin etkisinin olduğu söylenebilir.

NINJAL bünyesinde kurulan "Derlem Geliştirme Merkezi" Japoncayla ilgili çeşitli derlemler hazırlamaktadır. Merkez, şu ana kadar Çağdaş Japonca Yazı Dilinin Dengelenmiş Derlemi, Japonca Konuşma Dili Derlemi, Japonca Öğrenir Derlemi, Japonca Web Derlemi, Tarihî Japonca Derlemi, Çağdaş Japonca Derlemi gibi derlemler hazırlamıştır. Bu çalışma, Japonya'da hazırlanan derlemlerden biri olan BCCWJ'yi tanıtıp hazırlanma aşamaları ve yöntemi hakkında bilgi vermek amacıyla yapılmıştır. Böylece Türkçe derlem çalışmalarında da örnek alınabileceği düşünülmüştür.

BCCWJ'nin önemli özelliklerinden biri dengelenmiş olmasıdır. Dengelenmiş olma, iyi hazırlanmış derlemlerde bulunması gereken bir özelliktir. Dengelenmiş derlemler o dildeki farklı türlerden, bu türlerin kullanım oranlarına göre metin içermektedir. Çok sayıda metin içeren dengelenmemiş derlemlerde ise metin türlerinin ve miktarlarının önemi yoktur. (Adalı & Tantuğ, 2008).

Derlemlerin, özellikle de genel derlemlerin, dili ne derece temsil ettiği ne kadar dengelenmiş olduklarına, yani derlemde yer alan metin gruplarının çeşitliliğine ve dağılımına bağlıdır. Temsil edebilirlik açısından bir derlemin makul dengesi, kullanım amaçlarına göre belirlenir. Bu yüzden, hem yazılı hem de sözlü veriler içeren genel derlemler dengelenmiş olarak nitelenebileceği gibi; yazılı derlemler, sözlü derlemler ve özel amaçlı derlemler de kendi içinde dengelenmiş olabilir. Dengelenmiş bir derlem olarak kabul edilen İngiliz Ulusal Derlemi (BNC), birçok derlemin hazırlanmasında örnek alınmıştır. Bu derlemlerden bazıları, Amerikan Ulusal Derlemi, Kore Ulusal Derlemi, Polonya Ulusal Derlemi ve Rusya Yayın Derlemi'dir (McEnery, Xiao & Tono, 2006). Ayrıca dengelenmiş bir derlem olan Türkçe Ulusal Derlemi (TUD)'nin tasarım ilkelerinin geliştirilmesinde de BNC örnek alınmıştır (Aksan vd., 2014).

Denge, her ne kadar derlemler için olmazsa olmaz bir özellik ise de bir derlemin dengelenmiş olması, bilimsel olarak kanıtlanabilir olmasından ziyade, güçlü öngörüyle ve sezgilere güvenerek hazırlanmış olması şeklinde anlaşılmalıdır. (McEnery vd., 2006). Benzer şekilde Chen vd. de (1996) bir derlemin ne derece dengelenmiş olduğunu ve dili ne derece temsil ettiğini ölçmek için herhangi bir güvenilir ölçüt olmadığını belirtmektedir.

BCCWJ, Ulusal Japonca Araştırmaları Enstitüsü ile Eğitim, Kültür, Spor, Bilim ve Teknoloji Bakanlığının (Millî Eğitim Bakanlığı) ortak projesi olarak 2006-2010 yılları arasını kapsayan beş yıllık bir süreçte hazırlanmıştır (Maekawa, 2008).

BCCWJ'nin daha iyi anlaşılabilmesi için, Japoncanın bazı temel özelliklerini bilmek yararlı olacaktır. Japonca, Altay dil ailesi içinde yer alması ve sondan eklemeli bir dil olması yönünden Türkçe ile benzerlikler göstermektedir. Yapı bakımından eklemeli diller grubunda yer alan Japonca ile Türkçede söz dizimi benzerdir, tamlayan tamlanandan önce gelir, öznesiz cümle kurulabilir ve saygı dili kullanılır (Tekmen ve Takano, 2005). Japoncada şahıs eklerinin ve ünlü uyumunun olmaması ise Türkçe ile farklılaştıkları noktalardır.

Japonca yazı sisteminin esasını *kanji* adı verilen karakterler oluşturmaktadır. Bu karakterler Çin yazısından alınmıştır, fakat okunuşu Çinceyle farklılık arz etmektedir. Japoncada *kanji*yle birlikte *hiragana* ve *katakana* adı verilen karakterler de kullanılmaktadır. *Hiragana* ile *katakana*da yer alan sesler aynıdır fakat işaretleri farklıdır. İkisi de 46 işaretten oluşmaktadır ve yazılışları *kanji*ye göre daha sadedir. Bunların 5 tanesi sesli harf (a, i, u, e, o), 40'ı sessiz harfle kurulmuş hece, bir tanesi de Japoncadaki tek sessiz harf olan "n" sesidir. *Hiragana* eklerin ve bağlaçların, *katakana* ise yabancı kelimelerin

ve vurgulanmak istenen kelimelerin yazımında kullanılmaktadır. Bunun için Japonca bir metindeki yabancı kelimeler kolaylıkla ayırt edilebilmektedir.

1. BCCWJ'nin Tasarımı

Japonca üzerine yapılmış çeşitli derlemler bulunmakla birlikte bunların birçoğu özel amaçlı derlemlerdir. Örneğin bazı gazeteler, arşivlerindeki yazıları veri tabanlarına yükleyip ücretli olarak kullanıma açmışlardır. Bu yazılar, sadece gazetelerdeki yazı dilini yansıtmaktadır. Fakat dergi ve kitaplarda kullanılan dille gazete dili farklılık arz etmektedir (Maekawa, 2015). Yine Aozora Bunko Kütüphanesi tarafından edebî eserler sayısallaştırılarak kullanıma sunulmuştur. Fakat bunlar telif hakkı kalkmış eserler olduğundan en az elli yıl öncesinin yazı dilini yansıtmaktadır. Çağdaş Japoncanın genelini kapsayıcı nitelikte, dengelenmiş bir yazılı derlem bulunmadığından NINJAL, BCCWJ'nin tasarımını bu ihtiyacı karşılayacak şekilde yapmıştır. Bu düşünceyle BCCWJ'de sadece bir türden metinlere değil; kitap, gazete, dergi, rapor, blog gibi farklı yayınlardan alınan metinlere yer verilmiştir. Böylece farklı türlerde kullanılan yazı dilinin, bir derlemde toplanması amaçlanmıştır. Ayrıca şu ana kadar Japonca yazı diliyle ilgili toplanan verilerin büyük çoğunluğu sadece metinlerden ibaret olduğundan, bunlarla sadece Japonca karakter araması yapılabilmekte, ayrıntılı arama yapma imkânı sınırlı olmaktadır. BCCWJ'ye şekil bilgileri ve bibliyografik bilgiler gibi açıklamalar girildiği için ayrıntılı sorgulama sonuçlarına dayalı, daha derinlemesine analiz yapmak mümkündür (Yamazaki, 2015).

1.1 Temel İlkeler

BCCWJ hazırlanırken daha önceki Japonca derlemlerin eksikliklerini göz önünde bulundurmak ve günümüz Japonca yazı dilini kapsamlı olarak ele almak amacıyla bazı ilkeler doğrultusunda hareket edilmiştir. Maekawa (2008) ve Yamazaki (2009) BCCWJ'nin tasarımında göz önünde bulundurulan noktaları şu şekilde sıralamışlardır:

1. Çağdaş Japoncayı Yansıtan Bir Derlem: NINJAL'in şu ana kadar yapmış olduğu kelime araştırma tekniklerinden de yararlanılarak evreni temsil edecek bir derlem tasarlanmıştır. İstatistiki yöntemler kullanılarak evrenden elde edilecek verilerin gerçek durumu yansıtmasına, yani güvenilirliğin yüksek olmasına dikkat edilmiştir.

2. Çok Amaçlı Kullanılabilir Bir Derlem: Derlemin, söz varlığı ve dil bilgisi gibi dil araştırmalarından başka; yabancı dil olarak Japonca öğretimi, ana dil öğretimi, dil politikaları, sözlükbilim, doğal dil işleme gibi alanlarda da kullanılabilmesi amaçlanmıştır. Bunun için Japoncayı farklı açılardan yansıtan bir derlem tasarlanmıştır.

3. Kullanıma Açık Bir Derlem: Derlem iki farklı ara yüzle çevrimiçi kullanılabilmektedir. Birincisi kelime sorgulamaya, ikincisi şekil bilgisi özellikleri girilerek ayrıntılı sorgulama yapmaya yöneliktir. Üçüncü kullanım şekli de DVD versiyonudur.

4. Mevcut Derlemlerle Uyum: BCCWJ, NINJAL'in şu ana kadar geliştirmiş olduğu derlemlerle uyumlu çalışabilecek şekilde yapılandırılmıştır. XML ile metin yapıları

tanımlanıp iki çeşit dil birimine (kısa birim, uzun birim) göre şekil bilgileri girilmiştir. Böylece "Taiyô Kôpasu" (Güneş Derlemi) ve "Japonca Konuşma Dili Derlemi" ile uyumlu çalışması sağlanmıştır.

1.2 BCCWJ'nin Genel Yapısı

BCCWJ; yayın alt derlemi, kütüphane alt derlemi ve özel amaçlı alt derlem olmak üzere üç alt derlemden oluşmaktadır (Şekil 1). Her alt derlem belli bir ihtiyacı karşılamak amacıyla oluşturulmuştur.

<table>
<tr><td>Yayın Alt Derlemi
35 milyon kelime
Kitap, Dergi, Gazete
2001-2005</td><td>Kütüphane Alt Derlemi
30 milyon kelime
Kitap
1986-2005</td></tr>
<tr><td colspan="2">Özel Amaçlı Alt Derlem
35 milyon kelime
Rapor, Ders Kitabı, Yerel Yönetim Bülteni, Çok Satanlar, Forum, Blog, Şiir, Hukuk Metni, Meclis Tutanağı
1976-2005 yılları arasında değişiklik göstermektedir.</td></tr>
</table>

Şekil 1. BCCWJ'nin yapısı *http://pj.ninjal.ac.jp/corpus_center/bccwj/basic-design.html* sayfasından erişilmiştir.

1.2.1 Yayın Alt Derlemi

Yaklaşık 35 milyon kelime içeren yayın alt derlemi, yazı dilinin üretim aşamasını yansıtmaktadır. Bu alt derlemde, yazılı ürünlerin satış oranına ve bilinirliğine bakılmaksızın üretilen yazılı dile odaklanılmıştır. Bunun için evrende bulunan kitap, dergi ve gazetenin örnekleme dâhil edilme olasılıkları eşittir. 2001-2005 yılları arasını kapsayan bu alt derlemin evreni, yayın kataloglarına bakılarak belirlenmiştir. Kütüphane alt derlemiyle karşılaştırıldığında kelime ve eşdizimlilik gibi dilbilimsel özellikleri ve çeşitlilikleri daha iyi yansıtacağı öngörülmüştür. Bu alt derlem yetişkinlere yönelik kitapları içermektedir. BCCWJ'yi eğitim ortamlarında kullanırken bu özelliği göz önünde bulundurmak gerekir (Yamazaki, 2015).

Yayın alt derlemi kapsamındaki metinlerin seçiminde "tabakalı örneklem" yöntemi kullanılmıştır. Konusuna göre 11 farklı türdeki kitaptan alınan metinler 55 alt konuya, 6 farklı türdeki dergiden alınan metinler 30 alt konuya, 16 farklı gazeteden alınan metinler ise 80 alt konuya ayrılmıştır (Maekawa vd., 2014).

1.2.2 Kütüphane Alt Derlemi

1986-2005 yılları arasını kapsayan ve yaklaşık 30 milyon kelimeden oluşan bu alt derlem, yazı dilinin yazar ile okuyucu arasındaki kısmını yani dağıtım kısmını yan-

sıtmaktadır. Yazılan bir kitabın toplumdaki yayılma durumunu kütüphaneler aracılığıyla tespit etmeye yöneliktir. Geniş anlamda, topluma ulaşan yazılı ürünlerdeki dil olarak da adlandırılabilir. Belli bir alana ait uzmanlık kitapları evrene dâhil edilmediği için bu alt derlem genel kelime araştırmaları için daha uygundur. Ayrıca kitaplar, geniş sayılabilecek bir zaman aralığına ait olduğu için kütüphane alt derlemi artzamanlı gözlemler yapmak için de uygundur (Yamazaki, 2015).

Kütüphane alt derlemi ile yayın alt derleminin evrenleri karşılaştırıldığında birbirine yakın sonuçlar elde edilmiştir. Buna göre yayın alt derleminin kitap evrenindeki karakter sayısı 48,54 milyar, Tokyo bölgesindeki 13 belediyenin kütüphanesinde bulunan ortak kitapların toplam karakter sayısı 47,88 milyardır (Maruyama & Yuya, 2007).

1.2.3 Özel Amaçlı Alt Derlem

Yayın yılları 1976 ile 2005 yılları arasında farklılık arz eden yazılı ürünlerden elde edilen yaklaşık 35 milyon kelimeden oluşan bir alt derlemdir. Özel amaçlı alt derlem, yayın ve kütüphane alt derlemleri ile yeterli miktarda veri toplanamayacak yazılı türler için oluşturulmuştur (Maekawa, 2008). Mesela hükûmet raporlarının analiziyle ilgili yukarıdaki iki derlemden yeteri kadar veri elde etmek zor olacağı için, raporlarla ilgili bir evren belirlenmiş ve buradan elde edilen örneklemler alt derleme kaydedilmiştir. Ders kitapları, yerel yönetim bültenleri, çok satanlar, şiirler ve meclis tutanakları da aynı şekilde kaydedilmiştir. Yine İnternet yazı dili (Yahoo Forum, Yahoo Blog) derleme kaydedilerek kâğıt tabanlı basın dili ile karşılaştırılma imkânı sağlanmıştır (Yamazaki, 2015).

Özel amaçlı alt derlemde rapor türündeki metinler, Japon Hükûmetinin yayınladığı 1006 rapordan oluşan evren içerisinden, tesadüfi örnekleme yöntemiyle seçilmiştir. Bu raporların resmî yayınlarda kullanılan Japoncayı temsil ettiği düşünülmüştür. Konusuna göre 9 farklı türdeki rapordan alınan metinler, 54 alt konuya ayrılmıştır. Yahoo Forum'dan seçilen metinlerin evreni, Ekim 2004 - Ekim 2005 tarihleri arasında foruma yazılan 3 milyondan fazla soru ve bunlara verilen cevaptan; Yahoo Blog'daki metinlerin evreni ise Nisan 2008-Nisan 2009 arasında yayınlanan yaklaşık 3,5 milyon metinden oluşmaktadır (Maekawa vd., 2014).

1.3 Çekirdek Veri

BCCWJ'de etiketlemeler bilgisayar programı aracılığıyla otomatik olarak yapılmıştır. Fakat BCCWJ'nin yaklaşık %1'ine tekabül eden 1 milyon 100 bin kelimenin etiketi, analizlerin doğruluğunu artırmak amacıyla kontrol edilerek elle düzeltilmiştir (Maekawa vd., 2014). Bu bölüm çekirdek veri olarak adlandırılmaktadır. BCCWJ'nin analiz doğruluğu %98 iken çekirdek verininki %99'un üzerindedir. Çekirdek veri, derlemi söz varlığından ziyade şekil bilgisi araştırmaları için kullanmak isteyenlere yöneliktir. Çekirdek veriyi oluşturan yazılı ürünler kitaplar, dergiler, gazeteler, raporlar, Yahoo Forum ve Yahoo Blog olmak üzere altı türdür (Yamazaki, 2015).

BCCWJ'deki çekirdek verinin içeriğiyle ilgili bilgiler Tablo 2'de görülebilir.

Tablo 2 *Çekirdek Verinin İçeriği*

Tür	Metin Sayısı	Kelime Sayısı
Kitap	83	204.050
Dergi	86	202.268
Gazete	340	308.504
Rapor	62	197.011
Yahoo Forum	938	93.932
Yahoo Blog	471	92.746
Toplam	1980	1.098.511

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., vd. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation, 48,* 345-371. doi: 10.1007/s10579-013-9261-0

2. Örnekleme

Derlemin tasarımı sırasında, örnekleme alınacak metinlerin uzunluklarının ne kadar olacağına karar vermek gerekmektedir. Yamazaki (2015) metin uzunluklarının, derlemin bütçesi de dâhil olmak üzere birçok konuyu ilgilendirdiğini, metinler uzun olursa alınan metin sayısının az olacağını (telif haklarıyla ilgili iş yükünü azaltmak için), böylece daha az çaba gerektireceğini fakat bu durumda da derlemin, evreni tam olarak yansıtamayacağını ifade etmiştir. Ayrıca metinlerin sabit uzunlukta olmasının da başka sorunlar doğuracağını söylemiş, bu metinlerin istatistiki analizler için uygun olduğunu fakat birçok cümlenin yarım kaldığını, bu yüzden metnin bağlamına yönelik analizler için uygun olmadığını belirterek anlamsal bütünlüğe önem veriliyorsa metin büyüklüklerinin farklı olması gerektiğini vurgulamıştır.

2.1 Örneklem İçin Seçilen Metin Çeşitleri

Kullanım amacına göre metin uzunlukları farklılık arz etmektedir. BCCWJ'nin çok amaçlı kullanılabilmesi için, uzunluklarına göre iki çeşit metne yer verilmiştir. Biri nispeten kısa olan "sabit uzunluktaki metin", diğeri de bağlam anlaşılacak kadar uzun olan "değişken uzunluktaki metin"dir. Yayın alt derlemi ve kütüphane alt derlemine iki çeşit metin de alınmıştır (NINJAL, 2016).

2.1.1 Sabit Uzunluktaki Metinler

Sabit uzunluktaki metinler, tesadüfi seçilen bir noktadan itibaren 1000 karakter sayılarak elde edilmiştir. Bu sayıya noktalama işaretleri ve simgeler dâhil değildir. Bahsedilen yöntemle metin seçilirken ilk veya son cümle yarım kalabilmektedir. Derlem ara yüzünde sorgulama yapılırken bağlamın anlaşılması için cümleler tam olarak sisteme girilmiştir. Yine metindeki noktalama işaretleri ve simgeler sayıma dâhil edilmese de sisteme girilmiştir (NINJAL, 2016). Her karakterin seçilme olasılığının eşit

olabilmesi için önceden evrendeki tüm karakterlerin sayısı çıkarılmıştır. Büyüklükleri aynı olduğu için bu tür metinler, söz varlığı ve Japonca karakterlerle ilgili istatistiksel araştırmalar için uygundur (Maruyama, vd., 2015).

2.1.2 Değişken Uzunluktaki Metinler

Değişken uzunluktaki metinler, sabit uzunluktaki metinlerde olduğu gibi tesadüfi olarak seçilen bir noktadan başlamaktadır. Fakat bu tür metinlerde bağlamı korumak amacıyla metnin tamamı veya büyük bir bölümü alınmıştır. Her bölümün veya parçanın büyüklüğü aynı olmayacağı için bazı metinler kısa bazısı da uzun olabilmektedir. Çok büyük metinler, verinin yanlış yorumlanmasına sebep olabileceği için değişken uzunluktaki metinler en fazla 10.000 karakterle sınırlandırılmıştır. Bu tür metinler söylem çözümlemeleri ve metin incelemeleri için uygundur. Bu metin çeşidi yayın ve kütüphane alt derlemlerinde, özel amaçlı alt derlemde ise sadece raporlarda kullanılmıştır (Maekawa vd., 2014; Maruyama vd., 2015).

BCCJW'nin örneklem durumu Tablo 3'te ayrıntılı olarak görülebilir.

Tablo 3 *BCCJW'nin Örneklem Durumu*

Alt Derlem	Tür	Zaman Aralığı	Evren	Metin Sayısı	Kelime Sayısı (milyon)
Yayın	Kitap	2001-2005	48,5 milyar karakter	10.117	28,55
	Dergi		10,5 milyar karakter	1996	4,44
	Gazete		6,4 milyar karakter	1473	1,37
Kütüphane	Kitap	1986-2005	47,9 milyar karakter	10.551	30,38
Özel Amaçlı	Rapor	1976-2005	1006 kitap	1500	4,88
	Ders Kitabı	2005-2007	145 kitap	412	0,93
	Yerel Yönetim Bülteni	2008	100 belediye	354	3,76
	Çok Satanlar	1976-2005	951 kitap	1390	3,74
	Yahoo Forum	2004-2005	3,12 milyon soru	91.445	10,26
	Yahoo Blog	2008-2009	3,46 milyon yazı	52.680	10,19
	Şiir	1980-2005	130 kitap	252	0,25
	Hukuk Metni	1976-2005	718 yasa	346	1,08
	Meclis Tutanağı	1976-2005	32.925 toplantı	159	5,10
Toplam				172.675	104,91

Maruyama, T., Kashino, W. & Tanaka, M. (2015). Sanpuringu. *Gendai Nihongo kakikotoba kinkô kôpasu ryô no tebiki* içinde (s. 28-44). The National Institute for Japanese Language and Linguistics (NINJAL). http://pj.ninjal.ac.jp/corpus_center/bccwj/doc.html sayfasından erişilmiştir.

Tablo 3'te görüldüğü gibi üç alt derlemde yaklaşık 105 milyon kelime bulunmaktadır. Kütüphane alt derleminde 30 milyon, yayın alt derleminin "kitap" türünde yine yaklaşık 30 milyon kelime bulunmaktadır. Bunların toplam sayı içindeki oranına bakıldığında derlemin yaklaşık %60'ının kitaplardan oluştuğu görülmektedir.

BCCWJ'de, sabit uzunluktaki metinlerin ve değişken uzunluktaki metinlerin hangi yazılı ürünlerden alındığı Tablo 4'te gösterilmiştir.

Tablo 4 *Yazılı Ürünler ve Metin Çeşitleri*

Alt Derlem	Tür	Sabit Uzunluktaki Metin	Değişken Uzunluktaki Metin
Yayın	Kitap	√	√
	Dergi	√	√
	Gazete	√	√
Kütüphane	Kitap	√	√
Özel Amaçlı	Rapor	√	√
	Ders Kitabı		√
	Yerel Yönetim Bülteni		√
	Çok Satanlar		√
	Yahoo Forum		√
	Yahoo Blog		√
	Şiir		√
	Hukuk Metni		√
	Meclis Tutanağı		√

Tablo 4'te de görüldüğü gibi değişken uzunluktaki metin bütün yazılı ürünlerde kullanıldığı hâlde; sabit uzunluktaki metin, yayın alt derlemi ile kütüphane alt derleminin tamamında ve özel amaçlı alt derlemin sadece rapor türünde kullanılmıştır.

2.2 Metinlerin Örtüşmesi

Derlem için sabit uzunluktaki ve değişken uzunluktaki metinleri ayrı ayrı seçmek çok fazla iş yüküne sebep olacağı için, BCCWJ'de bir defada iki çeşit metin de alınmıştır (Yamazaki, 2015). Bunun için üç model uygulanmıştır. Yamazaki (2015) bu üç modeli şöyle açıklamıştır:

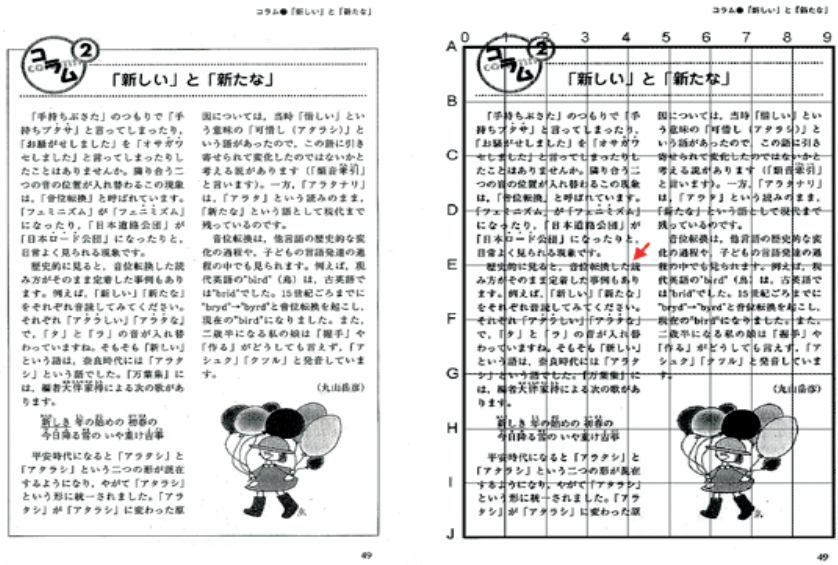
En fazla uygulanan model, sabit uzunluktaki metnin değişken uzunluktaki metnin içinde yer aldığı modeldir. Yine çok uygulanan ikinci bir model, sabit uzunluktaki metnin, değişken uzunluktaki metnin sonundan seçilmesidir. Az sayıda kullanılan üçüncü model ise sabit uzunluktaki ve değişken uzunluktaki metinlerin örtüşmediği modeldir.

2.3 Tesadüfi Örneklem

NINJAL, kelime araştırmalarında tesadüfi örnekleme yöntemini kullanmaktadır. BCCWJ'de de veriler, genel olarak tesadüfi örnekleme yöntemiyle toplanmıştır. Fakat çağdaş Japoncayı etraflıca görebilmek için tesadüfi olmayan örneklemeye de yer verilmiştir. Tesadüfi örneklemenin yapılabilmesi için önce evren belirlenmiştir. Kitap, gazete, dergi, ders kitabı, rapor, yerel yönetim bülteni gibi basılı ürünlerin standart yazı dilini daha iyi yansıtacağı düşünülerek evrene dâhil edilmiştir. Mektup, günlük, ambalaj, reklam tabelası gibi türler ise sayılarının tespit edilmesi zor olacağı için evrene alınmamıştır (NINJAL, 2016).

Evren belirlendikten sonra her yayın türü için tabakalar ve alt tabakalar oluşturulmuş, sonrasında da bunların evren içindeki oranı çıkarılmıştır. Tabakaların oranlarını bilmek, hangi türden ne kadar metin alınacağı konusunda yol gösterici olmaktadır. Bu derlemde kitapla ilgili tabakalar, Japon Onlu Sınıflama Sistemi (NDC) örnek alınarak oluşturulmuştur. Bu sistemde kitaplar konusuna göre; genel (0), felsefe (1), tarih (2), sosyoloji (3), doğa bilimleri (4), mühendislik (5), endüstri (6), sanat (7), dil (8), edebiyat (9), diğer (n) şeklinde sınıflandırılmıştır (Maruyama vd., 2011). Sonrasında ise belirli bir zaman aralığında yayınlanan kitapların konusuna göre karakter sayısı ayrı ayrı çıkarılmış ve bunların genel içindeki oranına bakılmıştır. Japonca *kanji* adı verilen karakterlerle yazıldığı için kelimeler bazen bir, bazen de iki veya daha çok karakter bir arada kullanılarak oluşturulmaktadır. Bunun için kitaplarda kelime sayıları yerine karakter sayıları tespit edilmiştir. Bu bilgiler yıllık yayın kataloglarından veya yayın veri tabanlarından elde edilebilmektedir.

Örneklem için metin seçilirken Şekil 2'de görüldüğü gibi önce kitabın herhangi bir sayfası seçilip on yatay ve on dikey çizgiyle bölünmüştür. Sonrasında ise çizgilerin yüz kesişim noktasından biri rastgele seçilip bu çizgiye en yakın karakterden başlamak üzere 1000 karakter sayılmıştır. Sabit uzunluktaki metinler (1000 karakter) bu şekilde seçilmiştir. Reklam, resim, şekil, grafik gibi görseller örnekleme dâhil edilmemiştir. Gazete ve dergilerden metin seçme yöntemi de aynı olmakla birlikte, bu türlerin sayfa tasarımları daha karışık olduğu için karakter sayımı da kitaba göre zor olmaktadır (Maruyama, Kashino ve Tanaka, 2015).



Şekil 2. Metin seçimi. *http://pj.ninjal.ac.jp/corpus_center/bccwj/sampling.html* sayfasından erişilmiştir.

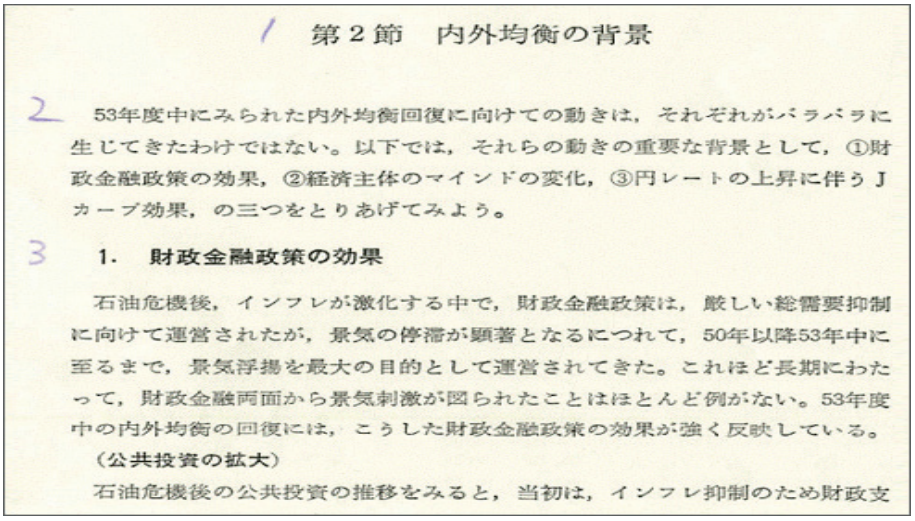
3. BCCWJ'de XML (Genişletilebilir İşaretleme Dili) ile Etiketleme

"Extensible Markup Language (XML), hem insanlar hem bilgi işlem sistemleri tarafından kolayca okunabilecek metinler oluşturmaya yarayan bir işaretleme dilidir. Veri saklamanın yanında farklı sistemler arasında veri alışverişi yapmaya yarayan bir ara format görevi de görür" (https://tr.wikipedia.org/wiki/XML).

Etiketleme genel olarak "dil dışı etiketleme" ve "dil içi etiketleme" olarak iki gruba ayrılmaktadır. Yazar adı, eser türü, yayın yılı gibi bibliyografik bilgiler dil dışı etiketleme; kelime türü gibi şekil bilgisiyle ilgili unsurlar ise dil içi etiketleme grubuna girer.

NINJAL, BCCWJ'de yer alan metinlerle ilgili ayrıntılı veri sorgulamanın mümkün olduğunu belirtmektedir. Bu veriler; bibliyografik bilgiler (başlık, yazar adı, yayın evi, yayın yılı, türü), Japonca karakter bilgileri (karakterin okunuşu, yazım hatası, Japonca standart dil kodlaması, sayıların üst ve alt simgeleri), metin yapısıyla ilgili bilgiler (metin, paragraf, cümle, başlık, alıntı, maddeler vb.), metin bilgisi (sabit uzunluktaki metinlerin kapsamıyla ilgili bilgiler) olarak gruplandırılmaktadır (Yamaguchi vd., 2011).

NINJAL'in örnek olarak gösterdiği Şekil 3'teki Japonca metin, Şekil 4'teki gibi sayısallaştırılmıştır (Maekawa vd. 2014).



第２節　内外均衡の背景

53年度中にみられた内外均衡回復に向けての動きは，それぞれがバラバラに生じてきたわけではない。以下では，それらの動きの重要な背景として，①財政金融政策の効果，②経済主体のマインドの変化，③円レートの上昇に伴うＪカーブ効果，の三つをとりあげてみよう。

1.　財政金融政策の効果

石油危機後，インフレが激化する中で，財政金融政策は，厳しい総需要抑制に向けて運営されたが，景気の停滞が顕著となるにつれて，50年以降53年中に至るまで，景気浮揚を最大の目的として運営されてきた。これほど長期にわたって，財政金融両面から景気刺激が図られたことはほとんど例がない。53年度中の内外均衡の回復には，こうした財政金融政策の効果が強く反映している。

（公共投資の拡大）

石油危機後の公共投資の推移をみると，当初は，インフレ抑制のため財政支

Şekil 3. Etiketlenecek olan Japonca metin *http://pj.ninjal.ac.jp/corpus_center/bccwj/en/XML.html* sayfasından erişilmiştir.

```
<?xml version="1.0" encoding="UTF-8"?>
<sample sampleID="OW1X_00000" version="1.0" type="variableLength">
<article articleID="OW1X_00000_V001" isWholeArticle="false">
<titleBlock>
<title><sentence type="quasi">第２節　内外均衡の背景</sentence><br
type="automatic_original"/></title>
</titleBlock>
<paragraph>
<sentence>　５３年度中にみられた内外均衡回復に向けての動きは，それぞれがバラバラに生じてきたわ
けではない。
</sentence><sentence>以下では，それらの動きの重要な背景として，…
</paragraph>
<cluster>
<titleBlock>
<title><sentence type="quasi">１．財政金融政策の効果</sentence><br
type="automatic_original"/></title>
</titleBlock>
<paragraph>
<sentence>　石油危機後，インフレが激化する中で，財政金融政策は，厳しい総需要抑制に向けて運営されたが，景気の停滞が顕著となるにつれて，５０年以降５３年中に至るまで，景気浮揚を最大の目的として運営されてきた。</sentence> …
</paragraph>
<cluster>
<titleBlock>
<title><sentence type="quasi">（公共投資の拡大）</sentence><br type="automatic_original"
/></title>
</titleBlock>
<paragraph>
<sentence>　石油危機後の公共投資の推移をみると，当初は，インフレ抑制のため財政支出が抑制され，公共事業の伸びは低いものにとどまっていた。</sentence><sentence>公的固定資本形成の実質ＧＮＰ増加寄与度は，　…
</paragraph>
```

Şekil 4. Etiketlenmiş metin *http://pj.ninjal.ac.jp/corpus_center/bccwj/en/XML.html* sayfasından erişilmiştir.

"< >" işareti içerisindeki bölümlere "etiket" adı verilir. Bir derlemdeki etiket türleri ve bunların nasıl düzenlendiği, o derlemde sorgulama yapmada önemlidir (NINJAL, 2016). Maekawa vd.'nin (2014) açıkladığı üzere, yukarıdaki şekilde yer alan <article>, <paragraph>, <cluster>, <sentence> gibi etiketlerin hepsi, bu metnin tabaka yapısını gösterir. Ayrıca, Şekil 3'te gösterilen XML metin türüne C-XML (Karakter temelli XML) türü metin adı verilir. Diğer bir metin türü ise M-XML (Şekil [morfoloji] temelli XML)'dir (Maekawa vd., 2014).

3.1 Etiket Türleri ve Anlamları

BCCWJ'de kullanılan etiketlerden bazıları Tablo 5'te gösterilmiştir.

Tablo 5 *BCCWJ'de Kullanılan Bazı Etiket Türleri*

Etiket Türü	**Etiket Adı**	İçeriği
Örneklem	Örnek Metin (sample)	Tek bir metni işaret eder
	Örnekleme (sampling)	Metinlerin alındığı yerle ilgili bilgiler
Tabakalı Yapı (Metin Yapısı)	Metin (article)	Bir yazar tarafından yazılmış, anlamsal bütünlüğü olan metin
	Başlık (title)	İçeriği tanımlayan bir başlığı gösterir (bir bölüm veya yazının başlığı)
	Grup/Küme (cluster)	Etiketlenmiş bir başlığın altındaki metnin tamamı
	Liste (list)	Maddeler hâlinde sıralanmış metin bölümleri/ ad öbeği listeleri. Listelemeyle ilgili unsurlar.
	Paragraf (paragraph)	Paragraf ayrımlarını gösterir.
	Cümle (sentence)	Cümle ayrımlarını gösterir.
Şekil ve Tablolar (Metin Yapısı)	Şekil (figure)	Şekil, tablo, fotoğraf, resim vb.
	Başlık (caption)	Şekil ve tabloların isimleri ve açıklamaları
Alıntılar (Metin Yapısı)	Alıntı (citation)	Başka metinlerden yapılan alıntılar
	Konuşma (speech)	Konuşma ve iç konuşma
Not (Belge Yapısı)	Not (notebody)	Dipnot, son not vb. Metinle ilgili açıklayıcı ifadeler
Diğer (Belge Yapısı)	Öz (abstract)	Metin ya da başlık grubuna girmeyen özet bilgiler
	Şiir (verse)	Şiir, şarkı
Karakterler ve Yazım	Ruby (ruby)	Kanji karakterlerinin okunuşları
	Düzeltme (correction)	Metindeki düzeltilmiş Japonca karakterler
	Eksik Karakter (missing character)	Japoncada bulunmayan karakterler

http://pj.ninjal.ac.jp/corpus_center/bccwj/en/c-xml.html sayfasından erişilmiştir.

3.2 Şekil Bilgisi (Kısa Birim, Uzun Birim)

Ogura ve Fujiike (2015), ayrıntılı sorgulamaya ve analizlere uygun olması amacıyla BCCWJ'de kısa birim ve uzun birim olmak üzere iki çeşit dil birimi kullanıldığını ifade etmiş, sonrasında da bu yapıların özelliklerini belirtmiştir. Buna göre:

Kısa birimler, derlemden kullanım örnekleri bulmaya, uzun birimler ise BCCWJ'deki metinlerin dil özelliklerini tespit etmeye yönelik birimlerdir. Analiz birimlerinin, çok miktarda verinin bilgisayar ortamında işlenmesine olanak sağlayacak özellikte olması gerekir. BCCWJ'deki bütün metinler, kısa birim ve uzun birim olmak üzere iki şekilde analiz edilmiştir. Analizlerin doğruluğu %98'in üzerindedir. Analizlerin doğruluğu türlere göre çok az farklılık göstermektedir. Kısa ve uzun birimler, NINJAL'in kelime araştırmaları için geliştirdiği "Japonca Konuşma Dili Derlemi"nde de kullanılmıştır. Kısa birim, cümlenin en küçük birimlere (biçimbirim), uzun birim ise bütünlük ifade eden birimlere ayrılmış şeklidir. Örneğin 国立国語研究所は人間文化研究機構に移管される。 *"Ulusal Japonca Araştırmaları Enstitüsü, Beşeri Bilimler Enstitüsüne devrediliyor."* cümlesi / 国立 / 国語 /研究 / 所 / は / 人間 / 文化 / 研究 / 機構 / に / 移管 / さ / れる / 。/ (/Ulusal/ Japonca/ Araştırmaları/ Enstitüsü/ Beşeri/ Bilimler/ Enstitüsü/ne/ devred/il/iyor/./) şeklinde 14 kısa birime ayrılırken / 国立 国語研究所 / は / 人間文化研究機構 / に / 移管さ / れる / 。/ (/Ulusal Japonca Araştırmaları Enstitüsü/ Beşeri Bilimler Enstitüsü/ne/ devred/il/iyor/./) şeklinde 7 uzun birime ayrılmaktadır.

4. BCCWJ'nin Yayınlanması

BCCWJ, iki çevrim içi ve bir DVD versiyonu olmak üzere üç şekilde kullanıma açılmıştır. Çevrim içi kullanımı *Shonagon* ve *Chunagon* isimleri verilen iki ara yüz üzerinden sağlanmaktadır. Kelime sorgulamaya yönelik hazırlanmış olan *Shonagon*, başvuru yapılmadan kullanılabilirken biçim bilgisiyle ilgili sorgulamalar yapmaya yönelik olan *Chunagon*'u kullanmak için başvuru yapmak gerekmektedir. İki ara yüzün de kullanımı ücretsizdir. DVD versiyonu hem ücretlidir hem de kullanım şartlarının kabul edildiğini bildiren bir sözleşme imzalamak gerekmektedir. Ayrıca DVD versiyonunun akademik ve genel kullanımı için faklı ücretler uygulanmaktadır. Başvurusu iki yıl geçerli olan DVD versiyonunda sadece veriler yüklenmiştir, sorgulama yapma imkânı yoktur.

Shonagon'da farklı arama seçenekleri bulunmaktadır. Herhangi bir tür, konu ya da alt konu işaretlenmediği sürece genel arama yapılmaktadır. İstenirse 11 konuda ve her birinin alt konularında aramalar sınırlandırılabilmektedir. Ayrıca zaman aralığı sınırlaması da yapılabilmektedir. Ara yüzde (http://www.kotonoha.gr.jp/shonagon/) arama yapılabilen tür, konu ve alt konular Tablo 6'da gösterilmiştir.

Tablo 6 *Shonagon'da Arama Seçenekleri*

Tür	Konu	Alt Konu
Kitap (1971-2005)	Japon Onlu Sınıflama Sistemi (NDC)	Genel (0), felsefe (1), tarih (2), sosyoloji (3), doğa bilimleri (4), mühendislik bilimleri (5), endüstri (6), sanat (7), dil (8), edebiyat (9), diğer
Dergi (2001-2005)	Genel	Genel/medya, genel, ev/yaşam, çocuk, eğlence/sanat, uğraş/hobi, spor
	Eğitim/Sanat ve Bilim	Eğitim, öğrenme/dil, edebiyat/sanat, sosyoloji, doğa bilimleri, beşeri bilimler
	Siyaset/Ekonomi/Ticaret	Siyaset/diplomasi, ekonomi/işletme, finans/maliye, ticaret/tüketici, vatandaşlık/iş gücü
	Üretim	Tarım, orman ve balıkçılık; ulaştırma/haberleşme
	Endüstri	makine, elektrik/elektronik
	Sağlık / Tıp	Sağlık, tıp
Gazete (2001-2005)	Ulusal	4 gazete
	Bölgesel	3 gazete
	Yerel	7 gazete
Rapor (1976-2005)	Ulaştırma	Turizm, ulaştırma (ulaşım, alt yapı), Tokyo bölgesi, arazi (ülke topraklarının kullanımı)
	Diplomasi	Diplomasi (son durum), resmî kalkınma yardımı (ODA)
	Güvenlik	Polis, nükleer güvenlik, nükleer enerji, trafik güvenliği, çevre kirliliği, itfaiye, suç, savunma, afet önleme
	Eğitim	Eğitim, kültür, spor, bilim ve teknoloji (eğitim faaliyetleri)
	Çevre	Çevre, geri dönüşüm
	Sosyal Hizmetler	Sağlık ve çalışma, yaşlılar, toplum hayatı, aile planlaması, engelliler, insan hakları eğitimi ve bilinci, gençler, kadın-erkek eşitliği
	Bilim ve Teknoloji	Bilim ve teknoloji, bilişim (iletişim)
	Ekonomi	Enerji, üretim, ekonomi ve finans, kamu yararı şirketleri, tekelleşmeyi önleme, bölgesel ekonomi, kobiler, ticaret, iş gücü
	Tarım, Orman ve Balıkçılık	Gıda ve tarım, ormancılık, su ürünleri
Ders Kitapları (2005-2007)	Ana Dil	İlkokul, ortaokul, lise
	Matematik	İlkokul, ortaokul, lise
	Fen Bilgisi	İlkokul, ortaokul, lise
	Toplum	İlkokul, ortaokul, lise
	Yabancı Dil	Ortaokul, lise
	Ev Ekonomisi ve Ev İşleri	İlkokul, ortaokul, lise
	Sanat	İlkokul, ortaokul, lise
	Sağlık ve Beden Eğitimi	Ortaokul
	Bilgi	Ortaokul
	Yaşam	İlkokul

Yerel Yönetim Bülteni (2008)	Hokkaido	1 şehir
	Tohoku Bölgesi	6 şehir
	Kanto Bölgesi	7 şehir
	Çubu Bölgesi	9 şehir
	Kinki Bölgesi	7 şehir
	Chugoku Bölgesi	5 şehir
	Shikoku Bölgesi	4 şehir
	Kyushu-Okinawa Bölgesi	8 şehir
Yahoo Forum (2005)	Eğlence ve Hobi	Oyun, televizyon-radyo, film, müzik, şovmen-sanatçı, fal-doğaüstü, kitap-dergi-çizgi roman
	İnternet, Bilgisayar ve Elektrikli Ev Eşyaları	İnternet, bilgisayar ve malzemeleri, elektrikli ev eşyaları-ses sistemleri, cep telefonu
	İş, Ekonomi ve Para	Ev ekonomisi-birikim, hisse senedi ve ekonomi, şirket ve yönetim, sigorta-vergi-emeklilik
	Meslek ve Kariyer	Yetenek-kişisel gelişim, işe girmek-iş değişikliği, geçici iş-yarı zamanlı çalışma, çalışma sorunları-çalışma yöntemleri
	Haber, Siyaset, Uluslararası İlişkiler	Haber-olay, siyaset-sosyal sorunlar
	Spor, Açık Hava Faaliyetleri, Otomobil	Açık hava faaliyetleri, spor, bisiklet, otomobil
	Yaşam ve Yaşam Rehberi	Alışveriş, gönüllülük-çevre sorunları-uluslararası iş birliği, ev işleri-konut, kamu tesisleri-resmi daireler, sosyal hizmetler-hemşirelik, yasa-tüketici sorunları, yemek-gurme-yemek tarifleri
	Sağlık, Güzellik ve Moda	Kozmetik-güzellik, moda, ruh sağlığı, sağlık-hastalık-diyet, aşk-insan ilişkileri sorunları
	Çocuk Yetiştirme ve Okul	Çocuk yetiştirme-doğum, sınav, ilkokul-ortaokul-lise, üniversite-yurt dışı eğitim, erken çocukluk eğitimi-anaokulu-kreş
	Görgü, Törenler	Görgü, törenler, festival ve yıllık faaliyetler
	Eğitimi ve Akademi, Bilim	Genel kültür, sanat-edebiyat-tarih, dil-dil çalışması, matematik-bilim, hava-astronomi-uzay, hayvan-bitki-ev hayvanı
	Bölge, seyahat, gazi	Yurt dışı, ulaşım-harita, yurt içi
	Yahoo Japonya	Yahoo Satış, Yahoo Hizmetleri, Yahoo Forum
	Diğer	Yetişkinlere özel konular, şans oyunları

Yahoo Blog (2008)	İş ve Ekonomi	İş, finans ve yatırım, ekonomi, istihdam, meslekler
	Bilgisayar ve İnternet	İnternet, bilgisayar
	Yaşam ve Kültür	Gurme-içecek, çevre sorunları, mevsimler, felaket, olay-kaza, ulusal tatil-yıldönümü-yıllık faaliyet, kültürel etkinlikler
	Eğlence	Tema park, televizyon, film, müzik, şovmen-sanatçı, fal, doğa-üstü
	Ev	Ev hayvanı-hayvan, ev, ev aletleri, konut
	Siyaset	Uluslararası ilişkiler, siyaset ve siyasi faaliyetler,
	Sağlık ve Tıp	Sağlık ve güzellik, hastalık-belirti
	Okul ve Eğitim	Okul, eğitim
	Bilim	Doğa bilimleri, sosyal bilimler
	İlişki	Evlilik, aşk
	Bölge	Dünya, Japonya
	İlgi	Hobi ve spor
	Sanat ve İnsan	Tasarım, sanat, beşeri bilimler, sahne-tiyatro, edebiyat
	Yahoo Hizmetleri	Yahoo Avatar, Yahoo Satış, Yahoo Oyunlar, Yahoo Alışveriş, Yahoo Spor, Yahoo Blog
	Hobi ve Spor	Şans oyunları, spor, boş zaman, hobi, vasıta
Şiir (1980-2005	Tanka	
	Haiku	
	Şiir	
Hukuk Metni (1976-2005)	Konuya Göre	Anayasa, meclis, idari örgütlenme, devlet memurları, idari işlemler, yerel yönetim, yerel yönetim maliyesi, adalet, sivil işler, dedektif, polis, milli araziyi geliştirme, arazi, şehir planlaması, karayolları, afet önlemleri, inşaat ve konut, mali işler, vergi, tekelcilik ve girişimcilik, devlet tahvili, eğitim, kültür, endüstri (genel kurallar), tarım, ormancılık, deniz ürünleri, madencilik, sanayi, ticaret, finans ve sigortacılık, kara taşımacılığı, deniz taşımacılığı, havacılık, nakliye, posta, telekomünikasyon, çalışma, çevre koruma, halk sağlığı, sosyal hizmetler, savunma, dış işleri
Meclis Tutanağı (1976-2005)	Temsilciler Meclisi	Genel kurul, daimi komite, özel komite, diğer
	Senato	Genel kurul, daimi komite, özel komite, diğer

Shonagon'da, Tablo 6'da belirtilen her tür, konu ve alt konunun başında işaretlenebilecek bir kutucuk bulunmaktadır. Herhangi bir seçenek işaretlenmediği sürece bütün tür, konu ve alt konularla ilgili bütün zaman dilimlerinde sorgulama yapılmaktadır. İstenirse sorgulamada sınırlandırma yapılabilmektedir. Örneğin, "Türkiye" kelimesi (Japonca トルコ) "dergi→siyaset/ekonomi/ticaret→ siyaset/diplomasi→2003" seçeneklerinin başındaki kutucuklar işaretlenerek sorgulanabilir. Seçenekleri işaretlemede sınırlama bulunmamaktadır. Tablo 6'daki seçeneklerden istenildiği kadarı işaretlenebilir.

Shonagon'da sorgulama yapıldıktan sonra, aranan kelimeyle ilgili sonuç ekranında sırasıyla şu bilgiler gösterilmektedir: "sıra numarası", "kelimenin yer aldığı cümle-

nin baş tarafı", "kelime", "kelimenin yer aldığı cümlenin son tarafı", "yazar", "yazarın yaşı", "yazarın cinsiyeti", "tür", "başlık", "alt başlık", "cilt no", "editör/çevirmen", "yayın evi", "yayın yılı".

Sonuç

Bu çalışmada BCCWJ incelenerek tasarımı, temel ilkeleri, genel yapısı, alt derlemleri, örneklem için seçilen metin çeşitleri, etiketlenmesi ve kullanıma sunulmasıyla ilgili bilgi verilmiştir. Çalışma sonucunda BCCWJ'nin yayın alt derlemi, kütüphane alt derlemi ve özel amaçlı alt derlem olmak üzere toplam üç alt derlemden oluştuğu; örneklemde, sabit uzunluktaki metin ve değişken uzunluktaki metin şeklinde iki tür metin kullanıldığı görülmüştür.

BCCWJ'nin önemli özelliklerinden birisi dengelenmiş bir derlem olmasıdır. Dengelenmiş derlemlerin hazırlanması, dengelenmemiş derlemlere göre daha zor olmakla birlikte ait olduğu dili daha kapsamlı ve doğru şekilde yansıtabilmektedir. BCCWJ dengelenmiş yapısı ve içerdiği 105 milyonluk kelimeyle çağdaş Japonca yazı dilini kapsamlı olarak yansıtmaktadır.

BCCWJ, hem DVD versiyonuyla hem de Shonagon ve Chunagon adlı iki ara yüzle kullanılabilmektedir. Shonagon kelime düzeyinde, Chunagon ise şekil bilgisi düzeyinde sorgulama yapmaya yönelik tasarlanmıştır. Özellikle akademik çalışmalar için hazırlanan DVD versiyonunda ise sorgulama yapma imkânı yoktur, sadece veriler sisteme kaydedildiği biçimiyle yer almaktadır.

Türkçeyle ilgili hem ulusal hem de uluslararası düzeyde yapılacak çalışmalarda kapsamlı veri sağlayabilecek bir derlemin hazırlanması çok önemlidir. BCCWJ'nin yapısı ve tasarlanma süreci göz önünde tutulduğunda, Türkçe için hazırlanacak derlemler için iyi bir örnek olabilir.

BCCWJ, Ulusal Japonca Araştırmaları Enstitüsü bünyesinde kurulan Derlem Geliştirme Merkezi tarafında hazırlanmıştır. Türkiye'de de Millî Eğitim Bakanlığı veya Türk Dil Kurumu gibi kurumların bünyesinde bir "Derlem Geliştirme Merkezi"nin kurulması ve bu kurumların üniversitelerle ortaklaşa çalışmaları, hem insan gücüne hem de maddi güce ihtiyaç duyan derlem çalışmalarının standartlara uygun şekilde yapılmasına büyük katkı sağlayacaktır.

Kaynakça

Adalı, E. & Tantuğ C. (2008, Ekim). *Türkiye Türkçesi derleminin geliştirilmesi.* VI. Türk Dili Kurultayı'nda sunulmuş bildiri, Ankara. http://www.adalı.net/wp-content/uploads/2012/10/Türkiye-Türkçesi-Derleminin-Geliştirilmesi.pdf sayfasında erişilmiştir.

Aksan, Y. et al. (2012). Construction of the Turkish National Corpus (TNC). In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012). İstanbul. Turkiye. http://www.lrec-conf.org/proceedings/lrec2012/papers.html

Aksan, Y., Aksan, M., Özel, S. A., Yılmazer, H., Demirhan, U. U., Mersinli, Ü., vd. (2014). Web tabanlı Türkçe ulusal derlemi (TUD). *XVI. Akademik Bilişim Konferansı bildirileri,* 723-730. http://ab.org.tr/ab14/kitap/aksan_aksan_ab14.pdf sayfasından erişilmiştir.

Center for Corpus Development, NINJAL (2016). *Sanpuringu*. http://pj.ninjal.ac.jp/corpus_center/bccwj/sampling.html sayfasından erişilmiştir.

Chen, K., Huang, C., Chang, L. & Hsu, H. (1996). Sinica Corpus: Design Methodology for Balanced Corpora. *Language, Information and Computation (PACLIC 11)*, 167-176. https://aclweb.org/anthology/Y/Y96/Y96-1018.pdf sayfasından erişilmiştir.

Church, K. W. & Mercer, R. L. (1993). Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics, 19*(1), 1-24.

Çetin, E. & Özkan, B. (2011). Tarihsel derlem kavramı, Eski Türkçe ve Karahanlı Türkçesinin tarihsel derlemi (7.-13. YY.). İ. Yazar (Ed.), *III. Uluslararası Dünya Dili Türkçe Sempozyumu* içinde (s. 282-289). İzmir.

Maekawa, K. (2008). KOTONOHA Gendai Nihongo kakikotoba kinkô kôpasu no kaihatsu [Special issue]. *Nihongo no Kenkyû*, *4*(1), 82-95. file:///C:/Users/PC/Dropbox/Japonca%20Derlemler/110006782128.pdf sayfasından erişilmiştir.

Maekawa, K. (2015). Gendai Nihongo kakikotoba kinkô kôpasu nyûmon. *Gendai Nihongo kakikotoba kinkô kôpasu ryô no tebiki* içinde (s. 1-18). The National Institute for Japanese Language and Linguistics (NINJAL). http://pj.ninjal.ac.jp/corpus_center/bccwj/doc.html sayfasından erişilmiştir.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., vd. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation, 48,* 345-371. doi: 10.1007/s10579-013-9261-0

Maruyama, T. & Yuya, A. (2007). *Gendai Nihongo kakikotoba kinkô kôpasu ni okeru sanpuru kôseihi no sanshutsuhô –gendai Nihongo kakikotoba no mojisû chôsa-* (JC-D-06-02). http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-06-02.pdf sayfasından erişilmiştir.

Maruyama, T., Yamazaki, M., Kashino, W., Sano, D., Yuya, A., Sachiko, İ., vd. (2011). *Gendai Nihongo kakikotoba kinkô kôpasu ni fukumareru sanpuru oyobi shoshijôhô no sekkei to jissô* (JC-D-10-02). http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-02.pdf sayfasından erişilmiştir.

Maruyama, T., Kashino, W. & Tanaka, M. (2015). Sanpuringu. *Gendai Nihongo kakikotoba kinkô kôpasu ryô no tebiki* içinde (s. 28-44). The National Institute for Japanese Language and Linguistics (NINJAL). http://pj.ninjal.ac.jp/corpus_center/bccwj/doc.html sayfasından erişilmiştir.

McEnery, T. & Wilson, A. (2001). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Müller, F. & Waiber, B. (t.y). *Corpus Linguistics – an introduction.* http://www.anglistik.uni-freiburg.de/seminar/abteilungen/sprachwissenschaft/ls_mair/corpus-linguistics sayfasından erişilmiştir.

Ogura, H. & Fujiike, Y. (2015). Keitairon jôhô. *Gendai Nihongo kakikotoba kinkô kôpasu ryô no tebiki* içinde (s. 58-98). The National Institute for Japanese Language and Linguistics (NINJAL). http://pj.ninjal.ac.jp/corpus_center/bccwj/doc.html sayfasından erişilmiştir.

Özkan, B. (2011) *Türkiye Türkçesinin Yazın Dili Derlemi (TD1)*, Mersin Üniversitesi, http://derlem.mersin.edu.tr/turkcederlem/. Erişim tarihi: 13.05.2016

Ruhi, Ş., Eröz-Tuğa, B., Hatipoğlu, Ç., Işık-Güler, H., Acar, M. G. C., Eryılmaz, K., vd. (2010). *Sustaining a corpus for spoken Turkish discourse: accessibility and corpus management issues.* Language Resources: From Storyboard to Sustainability and LR Lifecycle Management, LREC May 17-24, 2010, Malta, 44-48. http://www.lrec-conf.org/proceedings/lrec2010/workshops/W20.pdf sayfasından erişilmiştir.

Say, B., Zeyrek, D., Oflazer, K. & Özge, U. (2002). Development of a corpus and a treebank for present-day written Turkish. K. İmer, G. Doğan (Ed.), *Proceedings of the Eleventh International Conference of Turkish Linguistics* (s. 183-192). Cyprus: Eastern Mediterranean University.

Sinclair, J. (1987). *Looking Up —An account of the COBUILD Project in Lexical Computing*. London: Collins.

Sinclair, J. (2005). Corpus andText – Basic Principles. M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice* içinde (s. 1-16). Oxford: Oxbow Books. https://user.phil-fak.uni-duesseldorf.de/~bontcheva/SS10CTCL/CTCL-IntroNotes.pdf sayfasından erişilmiştir.

Startvik, J. (1992). Ed. Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82. *Trends in Linguistics Studies and Monographs 65.* Berlin: Mouton.

Tekmen, A. N. & Takano, A. (2005). *Japonca dilbilgisi*. Ankara: Engin.

XML (2016). https://tr.wikipedia.org/wiki/XML adresinden erişilmiştir.

Yamaguchi, M., Takada, T., Kitamura, M., Yoko, M., Oshima, H., Kobayashi, M., vd. (2011). *Gendai Nihongo kakikotoba kinkô kôpasu ni okeru denshika fômatto ver.2.2* (LR-CCG-10.04). http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-04.pdf sayfasından erişilmiştir.

Yamazaki, M. (2009). Daihyôsei o yûsuru gendai Nihongo shoseki kôpasu no kôchiku [Special issue]. *Jinkô Chinô Gakkaishi*, *24*(5), 623-631.

Yamazaki, M. (2015). Gendai Nihongo kakikotoba kinkô kôpasu no sekkei. *Gendai Nihongo kakikotoba kinkô kôpasu ryô no tebiki* içinde (s. 20-27). The National Institute for Japanese Language and Linguistics (NINJAL). http://pj.ninjal.ac.jp/corpus_center/bccwj/doc.html sayfasından erişilmiştir.

Yılmaz, S. (2012). Türkçede sözlü derlem oluşturma çalışmaları üzerine değerlendirmeler (uluslararası global COE program projesi çerçevesinde). *Kôpasu ni Motozuku Gengogaku Kyôiku Kenkyû Hôkoku*, *9*, 165-184. http://cblle.tufs.ac.jp/assets/files/publications/working_papers_09/section/165-184.pdf sayfasından erişilmiştir.