Uygulamalı Dilbilim Açısından Türkçenin Görünümü, Dil Derneği Yayınları:5, Ankara 1994

Bilgisayar ve Sözlükçülük Yöntemleri

Hendrik Boeschoten & Hansje Braam

Bu bildiride Tilburg Üniversitesi Dil ve Edebiyat Fakültesinde yürütmekte olduğumuz sözlükbilimsel bir çalışmayı tanıtarak uyguladığımız yöntemleri kısaca tartışmak istiyoruz. Projenin asıl amacı, Hollanda ortaöğretiminde okutulan Türkçe (ve Arapça) dersleri için esas tutulabilecek sözlük listeleri hazırlamaktır.

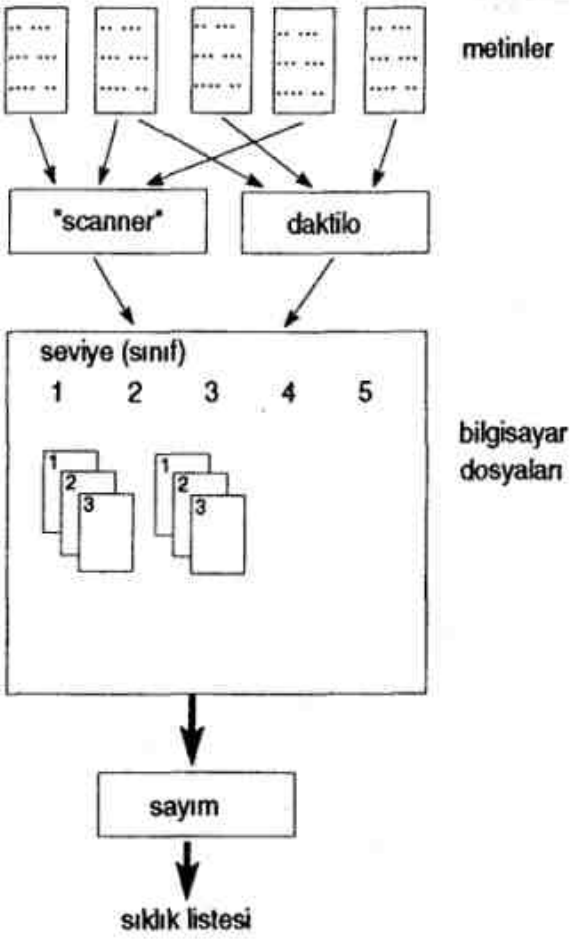
Hollanda ortaokulları ve liselerinde 3-4 yıldır Türkçe seçmeli ders olarak okutulmaktadır. Türkçe dersi herhangi bir yabancı dil dersi –Fransızca, Almanca, vb. gibi– biçiminde Hollanda'nın çeşitli kentlerinde bulunan bazı okulların programına konulmuştur. Bu arada Türkçe (ve aynı zamanda Arapça) artık bitirme sınavları için konu seçilebildiğinden dolayı eğitim sistemimizde bu dillerin önemi hayli artmış bulunuyor. Diğer yandan hemen hemen hepsi Türk asıllı olduğundan Türkçe derslerine katılan öğrencilerin herhangi bir yabancı dili sıfırdan başlayarak okuyan öğrencilere göre daha yüksek düzeye ulaşabilecekleri doğal bir beklentidir. Durum böyleyken eğitimcilere iki görev düşer. İlk olarak, öğrencilerin ortaokula girdikleri an sahip oldukları Türkçe bilgi düzeyleri saptanmalı. İkinci olarak, saptanan bu düzeyin ortalamasından yola çıkarak ortaöğretimde sunulan Türk dili derslerinin hedeflerini belirtmeli.

Hollanda Eğitim Bakanlığı geçen yıl bizi Tilburg Üniversitesi Dil ve Edebiyat Fakültesinin "Dil ve Azınlıklar" bölümü olarak ortaöğretimde verilen Türkçe derslerinde ve kullanılan metinlerde geçen sözcüklerin bir sıklık listesini çıkarmakla görevlendirmiştir. Böyle bir listenin hangi metinlere dayanacağı, kimler tarafından ne amaçla kullanılacağı konusunda bir direktif yoktu. Asıl amaç, öğretmenlerin ellerine hem ortaokulun, hem de lisenin son sınavları için birer hedef listesi *(target list)* vermektir, fakat amaçları ayrıntılarıyla proje çalışmaları sırasında saptayacağız.

Uyguladığımız yöntem şöyle: Hollanda ortöğretiminde çalışan birkaç Türk öğretmeni bize derslerinde kullandıkları metinleri yolluyorlar. Bu metinler gazeteden kesilen kısa yazılar veya Türk (çocuk) edebiyatının malı olan şiir ve öyküleri içeriyor. Yazılı metinleri bilgisayar dosyalarına koyduktan sonra sıklık sayımımızı bu dosyalardaki metinlerin üzerinde yapıyoruz. Bilgisayar çalışmalarında uygu-

ladığımız yöntemlerin salt bizim projemiz için değil, bütün çağdaş sözlükbilimsel çalışmalar için önemli olabileceğine ve bazı noktaları tartışmanın faydalı olacağına inanıyoruz.

Bizim başvurduğumuz yöntemin bir özelliği, ortaokulun ve lisenin çeşitli sınıflarında kullanılan metinleri beş düzeye göre betimleyerek aynı tutmamızdan ibarettir (bkz. Tablo 1).



Tablo 1. Metin dosyaları.

Metinleri ya özel bilgisayarda daktilo ettirerek, ya tarayıcı *(scanner)* ile okuyarak bilgisayara yüklüyoruz. Şu anda tarayıcı ile Türkçe yazılı metinleri okumakta ne denli başarıyla kullanabildiğimiz konusunda size bir fikir verebilmek için, bu işlerde bizden ileri gitmiş olan Bamberg'li arkadaşımız Semih Tezcan'ın aldığı sonuçlarından bir örneğe dikkatinizi çekmek istiyorum (bkz. Tablo 2).

basılmış bir metinden scanner ile bilgisayara alınmış metin (alıştırmasız)	*basılmış bir metinden scanner ile bilgisayara alınmış metin (alıştırmalı)*
,Şimdi bu rüzgar gec,ti buradan Ko,stum ama yetis,emedim. Nerelerde gezmis, tozmus, Ögrenemedim.	~imdi bu ruzgar geçti buradan Koştum ama yetişemedim. Nerelerde gezmiş tozmuş Ogrenemedim.
Besbelli denizden c,lhp Klyllar boyunca gitmis,tir. Tuz kokusu, katran kokusu, ter kokusu Yüreğini allak bullak etmis,tir.	Besbelli denizden cıkıp Kıyılar boyunca gitmiştir. Tuz kokusu, katran kokusu, ter kokusu Yüregini allak bullak etmiştir.
Sonra bas,laml,s, tlrmanmaya daglara dogru Bulutlarl koyun gibi gutmü~tür, Oks,saylp otlarla yaylalarda Büyütmüs,tür.	Sonra başlamış, tırmanmaya daglara dogru Bulutları koyun gibi gutmuştur, Okşayıp otları yaylalarda Büyütmüştür.
Köylere de u~radlysa eger Islak, karanllk odalarda bes,ik sallamls,tlr Günes, altında, c,alls,anlara Imdat eylemis,tir	Köylere de uğradıysa eğer Islak, karanlık odalarda beşik sallamıştır Güneş altında çalışanlara Imdat eylemiştir

Tablo 2. Tarayıcı *(scanner)* Türkçeye alıştırmadan ve Türkçeye alıştırdıktan sonra alınan sonuç (kaynak: Semih Tezcan, kişisel bildiri).

Kullanılan tarayıcı program (OMNIPAGE) aslında İngilizce, yani düz ASCI okutmak için geliştirilmiştir. Türkçeye alıştırılmadan program Türkçe metinlerden de çoğu harfleri çıkarıyor (Tablo, solda); yalnız, yumuşak /g/'yi okuyamıyor ve noktasız (ı/ harfini bilmeyip hep /l/'ye benziyor. En ağırsorun bunlardır; /ç/ ve /ş/ harflerini /c/ ve /s/ şeklinde okuması aslında sorun değildir. Program, alıştırılmadan önce de metinde geçen harflerin % 93'ünü doğru okuyor. Ama program öğretilir bir programdır. Biraz alıştırdıktan sonra bütün /ı/'ları ve bazı /ğ/'leri doğru okuyor (Tablo 2, sağda). Alıştırılmış program, metnin % 97'sini doğru okuyabiliyor. Tabii ki örnekteki sonuç ancak düzenli basılmış olan metinlerden alınıyor. Özellikle gazeteyi tarayıcıya kolay kolay okutamayız, daktilo ettirmek zorundayız.

Metinleri bilgisayar dosyalarına koyduktan sonra geçtiği metne göre her söze birkaç etiket *(label)* takılıyor.

n = düzey	*(sınıf)*
d = sözlüksel alan	*(domain)*
f = Sıklık	*(frequency)*
l = sözün türediği kök	*(lemma)*

Metinlerde geçen her söz belli bir "tip"e aittir, başka deyişle: her söz bir kökten türemiş ya da çekilmiştir. Bu söz-sözcük bağlantısını *(token-lemma matching)* ayrı bir yardımcı dosyada betimliyoruz (bkz. Tablo 3).

```
.....

&l="o^grenci                &b=leerling
 &t="o^grenc"ilere
 &t="o^grenci
 &t="o^grenciler
 &t="o^grencilerde
 &t="o^grencilere
 &t="o^grencileri
 &t="o^grencilerin
 &t="o^grencileriyle
 &t="o^grencim
 &t="o^grencinin
 &t="o^grencisi
 &t="o^grenciye
 &t="o^grencl
 &t=^grenciler
&l="o^grenim
&b=onderwijs
 &t="o^grenim
 &t="o^grenimi
 &t="o^grenimine
 &t="o^grenimini
```

```
&l="o^gret=                 &b=onderwijzen
 &t="o^gret
 &t="o^gretilir
 &t="o^gretilmeye
 &t="o^gretir
 &t="o^gretmektedir
 &t="o^gretmi,sler
 &t="o^gretti
 &t="o^grettim
 &t="o^grettin
 &t="o^grettir
 &t="o^grettiysem
&l="o^gretici               &b=leerrijk
 &t="o^gretici
 &t="o^greticidir

.....
```

Tablo 3. Sözcük dosyası (lemma list).[1]

1) Programları kullanabilmek için bütün dosyalar ASCI ile yazılıyor, yeni Türk abecesini kullanamıyoruz. Gördüğünüz gibi sözcüklerin anlamlarını Hollandaca kullanarak betimliyoruz.

Sözcüklerin sıklıklarını sonra bu yardımcı dosyayı kullanarak saptayabiliriz. Elde edilen sonuçların basit bir örneğini Tablo 4'te görebilirsiniz; bu liste aşağı yukarı 100.000 söz üzerinden elde edilmiştir. Şu var ki sonunda çıkaracağımız listelerde sıklık rakamları hem düzey, hem de kullanış alanına göre ayrı ayrı rakamlarla belirtilecektir.

```
.....
&f=00001  &l="o^g"un=          &b=pochen
 &f=00002  &l="o^g"ur=         &b=kotsen
 &f=00006  &l="o^g"ut          &b=raadgeving
 &f=00004  &l="o^g"ut=         &b=fijnmalen
 &f=00001  &l="o^g"utle=       &b=raadgeven
 &f=00013  &l="o^ge            &b=element
 &f=00006  &l="o^gle           &b=middag
 &f=00056  &l="o^gren=         &b=leren
 &f=00050  &l="o^grenci        &b=leerling
 &f=00007  &l="o^grenim        &b=onderwijs
 &f=00012  &l="o^gret=         &b=onderwijzen
 &f=00002  &l="o^gretici       &b=leerrijk
 &f=00003  &l="o^gretim        &b=onderwijs
 &f=00042  &l="o^gretmen       &b=leraar
 &f=00003  &l="oks"ur"uk       &b=hoest
 &f=00007  &l="oks"ur=         &b=hoesten
 &f=00012  &l="ok"uz           &b=os
 &f=00012  &l="ol,c"u          &b=maat
 &f=00001  &l="ol,c"u,s=       &b=
 &f=00001  &l="ol,c"ul"u       &b=in afmeting
 &f=00006  &l="ol,c=           &b=meten
 &f=00001  &l="ol,cek          &b=maatstaf
 &f=00079  &l="ol=             &b=sterven
 &f=00002  &l="old"ur"uc"u     &b=dodelijk
 &f=00012  &l="old"ur=         &b=doden
 &f=00017  &l="ol"u            &b=dode, lijk
 &f=00016  &l="ol"um           &b=dood
 &f=00005  &l="ol"uml"u        &b=sterfelijk
.....
```

Tablo 4. Sıklık listesi

Seçtiğimiz yöntem, yapılacak istatistik işlemler için de elverişlidir. Örneğin, bir sözcüğün asıl sıklığını yalnız toplu sıklıkla değil, daha nesnel bir sayıyla betimlemek istiyoruz. Çünkü herhangi bir sözcük iki üç metinde çok sık kullanılırken diğer metinlerde hiç geçmeyebilir. Böyle bir duruma normal dağılıma dayanan bir değişkenlik derecesini saptayarak yaklaşmak yanlış olur. Bunun için

esas sıklık ölçüsü olarak metinlerde ayrı ayrı elde edilen sıklıkların jeometrik ortalamasını kullanmayı düşünüyoruz:

$$jo(t) = \sqrt[N]{\prod_{i=1}^{N} f_i^t}$$

(**jo** = jeometrik ortalama; **t** = tip; **N** = metinlerin sayısı;

f_i^t = tipin i numaralı metinde sıklığı, "missing value" = 0.5)

Kullandığımız bilgisayar ile bilgisayar programlarına gelince: programlarımızı ICON adlı program dilinde yazıyoruz. Bu dil filolojik çalışmalar ve genellikle metinlere dayalı araştırmalar için pek elverişlidir. ICON, çeşitli tip bilgisayarlarda kullanılabilir. Veriler, yukarıdaki örneklerde olduğu gibi (Tablo 3, 4), dosyalarda açıkça ve düz biçimde yerleştiriliyor.[2] Kullandığımız makine, üniversitemizin merkezi sistemidir. Sistem VMS ile çalışıyor.[3] Böyle bir sistemle çalışmanın avantajları: Kullanılan standart programlar tam manasıyla "profesyonel" ve sapasağlamdır. Hem de sistem büyüktür: Projemiz seneye sona erdiğinde birkaç milyon sözcüğü makineye yerleştirmiş olacağız. Ayrıca uyguladığımız süreçler de geniş bir muhafaza hacmi gerektirmektedir.

Çalışmamıza burada dikkatinizi çekmek istememizin nedeni: Herhalde sözlüksel alanda benzeri çalışmalar başka arkadaşlar tarafından da yürütülmektedir. Bilgisayarla çalışırken karşılaşılan sorunlar aynı türden olmasa gerek. Diğer yönden, bizim projemizin çerçevesinde oldukça büyük bir metin temeli *(text-base)* geliştiriliyor ki, bu malzeme ileride başka alandaki –örneğin dilbilimsel alanda– araştırmalar için de kullanılabilecektir.

Genel olarak Türkçenin sözbilimi açısından başvurduğumuz yöntemin şu çekici yönleri vardır:

1) Araştırmanın amacı tamamen tanıtımsaldır. Yazılı dil kullanılışı olduğu gibi saptanmaktadır.

2) Böyle bir çalışmada önemli olan, esas tutulan metinlerdir. Herhangi bir metni seçmede kullanılan yöntem nasıl nesnel sayılmazsa, bizimki de sayılmaz. Yalnız, seçim kişisel de değildir, "kişiler arası" *(intersubjective)* sayılabilir.

3) Metinleri belli kullanış alanlarına göre seçtiğimizden, eksik kalan alanlara ait yeni metinler proje süresince metin temeline eklenebilir.

---

2) ICON üzerine genel bilgi için bkz.: RALPH E. GRISWOLD & MADGE T. GRISWOLD, The Icon programming language.-Englewood Cliffs N.J.: Prentice-Hall, 1983.

3) UNIX sistemi aslında daha iyi olurdu.