

Türkçe için Verimli bir Cümle Sonu Belirleme Yöntemi

Özlem AKTAŞ

Dokuz Eylül Üniversitesi
Bilgisayar Mühendisliği Bölümü
ozlem@cs.deu.edu.tr

ÖZET

“Doğal Dil İşleme” (DDİ) çok farklı amaçlarda kullanılan bir araştırma alanıdır ve günümüzde hızla yaygınlaşmaktadır. Konuşma analizi, konuşma tanımlama, imla doğrulama DDİ uygulama alanlarından sadece birkaçıdır. Bir dilin biçimbilimsel özelliklerinin belirlenebilmesi için o dili anlatan ve üzerinde istatistiksel ve biçimbilimsel analizlerin kolayca yapılabilirdiği bir derlem oluşturulması gereklidir. Böyle bir derlem oluşturmanın ilk aşaması cümle sonu belirleme işlemidir. Bu işlem oldukça karışık ve çözülmesi zor bir işlemdir, ancak derlem oluşturmanın en önemli aşamasıdır. Bu çalışmada cümle sonu belirleme problemini çözmek için yeni bir yöntem geliştirilmiştir. Cümle sonu belirleme işlemi için kullanılan kısaltma ve kural listeleri XML yapısında kaydedilmiştir; bu dosyalar cümle sonu belirleme işleminde başarı oranının artmasını sağlamıştır. Bu yeni yöntem cümlelerin doğru ve verimli biçimde ayrıştırılmalarını sağlayarak, araştırmacılara yardım edecektir.

ABSTRACT

“Natural Language Processing” (NLP) is a research area that is used for many different purposes and it becomes more popular continuously. Speech syntheses, speech recognition, machine translation, spelling correction are some of the application of NLP. For determining a language’s morphological specialties, it is needed to generate a corpus that represents the language and make some statistical and morphological analysis on it. The first step of generating such corpus is sentence boundary detection. This process is very complicated and hard to solve, but it is the most important part of the generating corpus. In this work, new method is developed to solve sentence boundary problem. The abbreviation list and rules generated for the sentence boundary detection are stored in an XML file; these files had provided successive results in sentence boundary detection. This new method will help researchers by separating sentences correctly and efficiently, about means of time and other costs.

Anahtar Kelimeler: Doğal Dil İşleme, Türkçe derlem, biçimbilimsel analiz, cümle sonu belirleme.

1. GİRİŞ

“Doğal Dil” insanlar tarafından kullanılan dildir. 1940lardan beri araştırmacılar doğal dillerin biçimbilimsel özelliklerini belirlemek için çalışmışlardır. Shannon İngilizce dilini araştırmış ve 1940 yılında İngilizce’nin düzensizliği ve tahmin edilebilirliği hakkında ilk araştırmasını yayınlamıştır [1]. Zipf tüm istatistiksel dağılımlara uygulanabilen bir teorem önermiştir [2]. 1940 – 1950 yıllarında bilgisayar teknolojisinin henüz gelişmemiş olmasından dolayı, yeterli miktarda bilgi toplanamamış ve işlenememiştir. Bilgisayar teknolojisinin hızla gelişmesiyle birlikte daha fazla bilgi toplanmış, Shannon ve Zipf’in araştırmaları kullanılarak yeni teknolojiler geliştirilmiştir.

Doğal dilin yapısının belirlenmesi, bilgi şifreleme işlemleri, konuşma tanımlama [3], optik karakter belirleme [4], yazı doğrulama [5] gibi işlemlerde yardımcı olur. Ayrıca yazılan bir kelimeye göre bir sonraki kelimenin tahmin edilebilmesi özellikle engelli insanların haberleşmesi için çok önemlidir. Ancak bunu yaparken tahmin edilebilen kelime sayısının çok fazla olduğu unutulmamalıdır. Bu nedenle sadece yazıdaki önceden yazılan kelimelere göre gelme olasılığı çok olan kelimeler tahmin edilebilir [6].

“Doğal Dil İşleme” (DDİ), akademik araştırmalarda ve ticari amaçlarla kullanılmaktadır. DDİ, doğal dili işleyen ve anlayan bir sistemin oluşturulması olarak tanımlanabilir [7].

DDİ’de derlemi oluşturup kullanan iki çeşit analiz bulunmaktadır; Biçimbilimsel ve İstatistiksel Analiz [8]. Biçimbilimsel analiz, cümle sonu belirleme, kelime türlerini (isim, sıfat, vb.) belirleme ve kelimelerin parçalarını (kök, ek, vb.) analiz etme gibi kelimelerin biçimsel durumlarını inceler. İstatistiksel analiz iki türlü yapılabilir; harfler ve kelimeler üzerine. Sesli ve sessiz harflerin dizilimi, harflerin n-gram analizleri, harfler arasındaki ilişkiler, vb. harfler üzerinde yapılabilen analizlerdir, buna “Harf Analizi” denir. Bir kelimedeki harf sayısı, kelimelerin n-gram frekansları, kelimelerin cümle içindeki dizilimi gibi analizler kelime üzerinde yapılabilir ve “Kelime Analizi” olarak adlandırılır.

Derlemin bazı tanımları aşağıda verilmiştir:

- Derlem, dilbilimsel bilginin koleksiyonudur, yazılı yada kaydedilen konuşmalar şeklinde olabilir [9].
- Doğal olarak meydana gelen metinlerden dilin çeşitliliğini ve durumunu belirlemek amacıyla seçilen ve bir araya getirilen metinler [10].
- Doğal dil İşleme alanında kullanılmak için yazılardan oluşturulmuş özel bir veritabanıdır ve kelimeleri hızlı şekilde bulma ve işleme gibi özel işlemleri yapmaya izin verir.

Birçok doğal dil işleme işlemlerinde, cümle sonu belirleme işi ilk şarttır. Kullanılabilen doğal dil işleme araçlarının çoğu cümle sonu belirleme işini güvenilir olarak yapmaz.

Cümle sonu işaretlerinin (“.”, “!” gibi) kullanılarak cümle sonunun belirlenmesi mümkündür. Ancak bazı işaretler kısaltmalar ve bunun gibi bazı işaretleri (e-posta adresleri, numaralandırma gibi) göstermek için de kullanılabilir. Aşağıda bazı örnekler görülmektedir:

- Cumartesi akşam 5 p.m.’de geldi.
- www.cs.deu.edu.tr okulumuzun web sitesidir.
- E-posta adresi bilgi@cs.deu.edu.tr ‘dir.

Bunlar cümle sonu bulma işlemlerinde karmaşa yaratan bazı durumlardır. Tüm diğer dillerde de bu gibi durumlar mevcuttur ve cümle sonu belirleme işlemlerini zorlaştırmaktadır. Bu çalışmada Türkçe için cümle sonu belirme işlemini doğru şekilde yapabilecek yeni bir algoritma geliştirilmiştir.

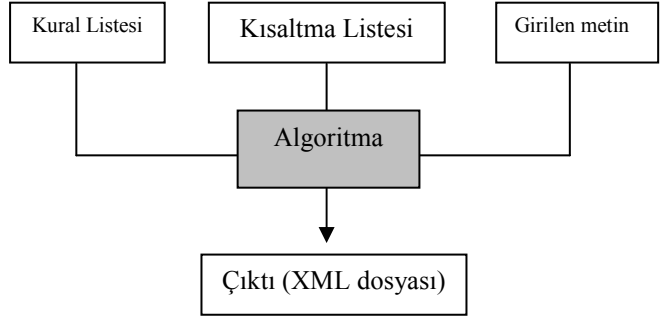
1. TÜRKÇE İÇİN GELİŞTİRİLEN CÜMLE SONU BELİRLEME YÖNTEMİ

Bir derlemi oluşturmanın ilk basamağı “cümleleri bulmak” tır. Genel olarak Türkçe cümlelerin ., ..., !, ? gibi işaretlerle bittiği bilinse de cümle sonu bulma işlemi bazı belirsizlikler sebebiyle çok karışık ve zor bir işlem haline gelmektedir. Örneğin;

- Uluslar, bu ekonomik buhran sonucunda 2. Dünya Savaşı’na yaşamıştır.
- Bu sezon kaybedilen maç sayısı 2. Dünya Kupası’na katılma şansı azalıyor.

İlk cümlede “.” Karakteri sıralama için kullanılırken ikinci cümlede cümle sonunu belirtmektedir. Ancak her iki cümlede de “.” işaretinden sonra büyük harf bulunmaktadır.

Yeni geliştirilen cümle sonu bulma algoritması aşağıdaki şekilde çalışmaktadır:



Cümle sonunu belirlemek için öncelikle XML formatında bir kural listesi oluşturulmuştur. Tablo 1’de kullanılan kurallar görülmektedir.

Tablo 1 Cümle sonu için kural listesi

Cümle Sonu	Kurallar
Doğru	L.U
Doğru	L.#
Doğru	?.'
Doğru	?."
Doğru	?.(
Doğru	?.)
Doğru	?.-
Doğru	?./
Doğru	?/
Doğru	U.L
Yanlış	L.L
Yanlış	?,.
Yanlış	#.L
Yanlış	#.'
Yanlış	#."
Yanlış	#.(
Yanlış	#.)
Yanlış	#.-
Yanlış	#.,
Yanlış	#.#
Yanlış	#.U

Cümle sonu kural listesi Tablo 1’de gösterildiği gibi üçlü grup şeklinde (“L.L” gibi) oluşturulmuştur. Ortadaki “.” (nokta) karakteri cümle sonu işaretini (“.”, “!” gibi) göstermektedir. Sol taraftaki karakter noktalama işaretinden önceki kelimenin ilk karakterinin durumunu, sağ taraftaki karakter ise işareten sonraki kelimenin ilk harfinin durumunu göstermektedir. Tablo 2’de kural listesindeki işaretlerin anlamları gösterilmiştir.

Tablo 2 Kural listesindeki işaretlerin anlamları

Karakter	Anlamı
.	Cümle sonu işaretleri (. ... ! ?)
L	Küçük harf (Lowercase)
U	Büyük harf (Uppercase)
#	Sayı
?	Herhangi karakter (ne olursa olsun)
-	-
.	.
((
))
/	/
‘	‘
“	“

Bu kuralları kullanarak cümle sonunu belirleme işleminin kolaylaşması amaçlanmıştır. Ancak kurallar oluşturulurken Türkçe dilinin özelliklerinden kaynaklanan zorluklar ortaya çıkmıştır ve çözümlenmeye çalışılmıştır.

Aşağıda Türkçe cümlelerin sonlarının bulunması sırasında belirsizlik yaratan bazı durumlar örneklendirilmiştir:

- Cumhuriyetimizin 75. yılı coşkuyla kutlandı.
- Tahta çıkan IV. Murat emirler yağdırdı.
- Olimpiyatlar için uzun zamandır çalışan Ahmet koşuda 2. Uzun atlamada ise ancak 4. olabildi.
- A. Mehmet YILDIZ size uğradı.
- Alfabenin ilk harfi A. Mehmet’e bunu öğretmeniz gerekiyor.

İlk cümlede “.” işaretinden sonra cümle bitmemektedir. “.” işareti sıralama belirlemek amacıyla kullanılmıştır. Dördüncü cümlede “A” harfi kısaltma olarak kullanılmış, bir sonraki cümlede ise tek başına bir kelime olarak kullanılmıştır.

Bunun gibi, cümlelerde belirsizlik yaratan kısaltmalar için Tablo 3’de görüldüğü gibi XML formatında bir kısaltma listesi oluşturulmuştur.

“IV. Murat” gibi roma rakamlarının kullanıldığı cümlelerde belirsizliklerin çözümlenmesi için bu kısaltma listesine roma rakamları da eklenmiştir. Kısaltma ve kural listeleri, kullanıcıların üzerlerinde değişiklikleri verimli biçimde yapabilmeleri için ana programdan bağımsız, XML formatında, ayrı birer dosya olarak düzenlenmiştir. Program bu listeleri dışardan alıp işlem yapmaktadır.

Bu listeler kullanılarak yazılar cümlelere daha verimli biçimde ayrılabilir. Ayrılan cümleler Tablo

4’te görüldüğü gibi yine XML formatında dosyalara kaydedilmektedir.

Tablo 3 XML formatındaki kısaltma listesinden örnek

```

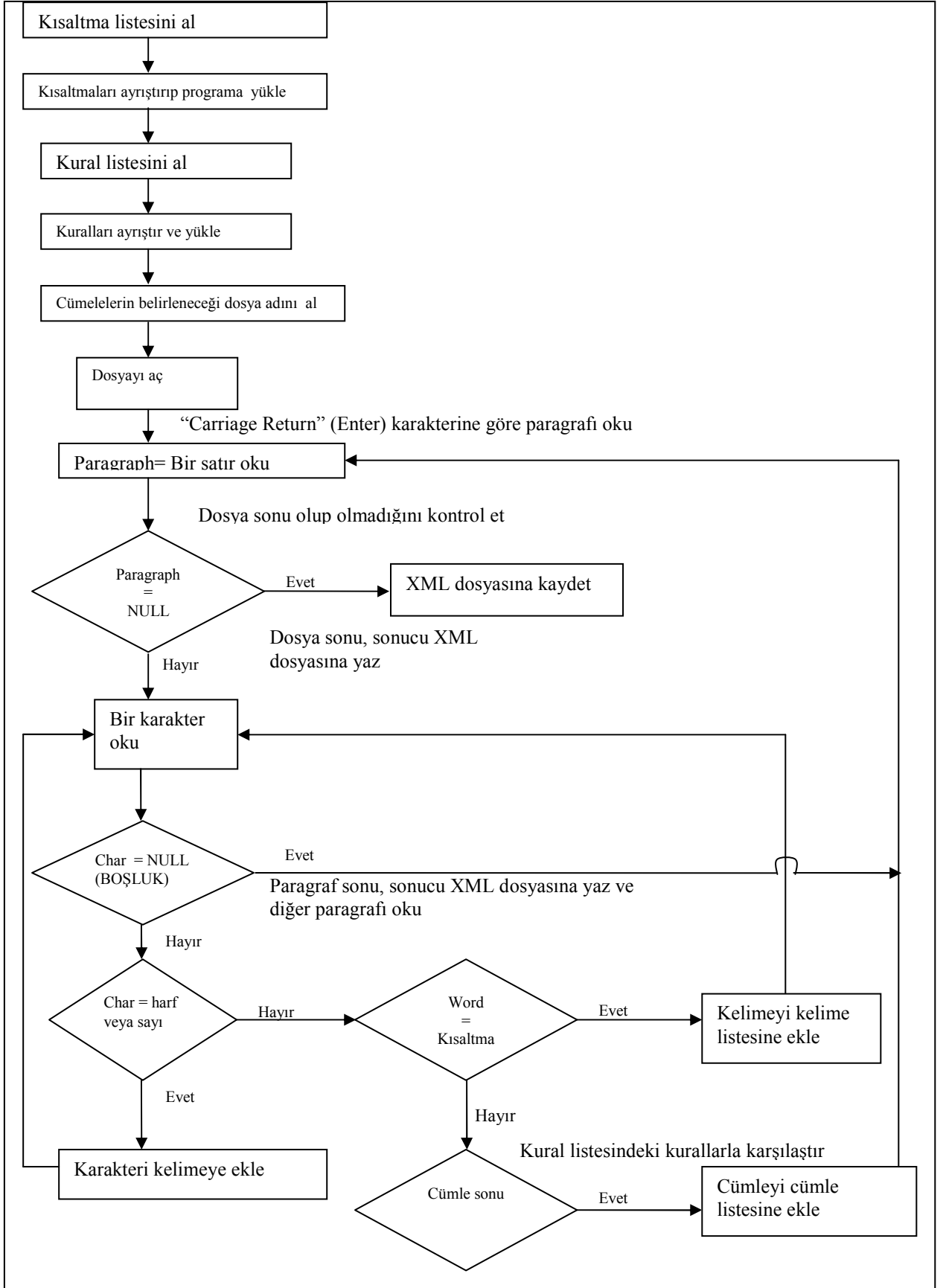
<abbreviations>
  <abbr>    A    </abbr>
  <abbr>    AA   </abbr>
  <abbr>    AAFSE </abbr>
  <abbr>    AAM   </abbr>
  <abbr>    AB    </abbr>
  <abbr>    ABD   </abbr>
  <abbr>    ABS   </abbr>
  <abbr>    ADSL  </abbr>
  <abbr>    AET   </abbr>
  <abbr>    ..... </abbr>
  <abbr>    HAVAŞ </abbr>
  <abbr>    HDD   </abbr>
  <abbr>    hek   </abbr>
  <abbr>    ..... </abbr>
  <abbr>    zf    </abbr>
  <abbr>    zm    </abbr>
  <abbr>    ZMO   </abbr>
  <abbr>    zool  </abbr>
</abbreviations>

```

Tablo 4 XML dosyasında bulunan örnek cümleler

Dosya Adı	Par.#	Cümle#	Cümle
88366.txt	0	0	Geçen hafta Radikal’de iki önemli yazı yayınlandı, anlatılanlar çok çarpıcıydı, doğrusu kamuoyunda daha geniş yankı bulmasını beklerdim.
	0	1	Nedense iddiaların üzerinde yeteri kadar durulmadı.
	1	0	İsmet Berkan ve Murat Yetkin’in yazıları, 30 Ağustos’ta emekliye ayrılan Kara Kuvvetleri ile Jandarma Komutanları Aytaç Yalman ve Şener Eruygur üzerine.
	1	1	İki komutanlığın da devir teslim törenine sınırlı sayıda gazeteci davet edildi, Radikal bu kez davete mazhar olmayan gazeteler arasında yer aldı.

Yeni geliştirilen algoritmanın çalışma şeması aşağıdaki şekilde görülmektedir:



Şekil 1 Algoritmanın çalışma şeması

3. Analiz sonuçları

Aşağıdaki paragraflar bir gazete haberinden alınmıştır:

“Uluslararası Para Fon'u (IMF) heyeti, çalışmalarını Hazine Müsteşarlığı'nda gruplar halinde sürdürdü. Edinilen bilgiye göre, kısmen Türkiye masası şefi Rıza Moghadam ve Hazine Müsteşarı İbrahim Çanakçı'nın da katıldığı toplantılara, Maliye Bakanlığı, Merkez Bankası, BDDK, Özelleştirme İdaresi Başkanlığı, Kamu Bankaları gibi kuruluşların yetkilileri de katıldı.

Bugünkü görüşmelerde, ödemeler dengesi, savunma sanayii, uluslararası rezervler, yerel yönetimler, yatırım programı, kamu mali yönetimi ve kontrol kanunu konuları tartışıldı.”

Program çalıştıktan sonra XML formatında aşağıdaki dosya oluşturulmuştur:

```
<?xml version="1.0" encoding="UTF-8"
standalone="yes" ?>
<File OriginalName="MD_ID_396980_M.txt">
  <Paragraph Index="0">
    <Sentence Index="0">Uluslararası
Para Fon'u (IMF) heyeti, çalışmalarını Hazine
Müsteşarlığı'nda gruplar halinde
sürdürdü.</Sentence>
    <Sentence Index="1">Edinilen
bilgiye göre, kısmen Türkiye masası şefi Rıza
Moghadam ve Hazine Müsteşarı İbrahim
Çanakçı'nın da katıldığı toplantılara, Maliye
Bakanlığı, Merkez Bankası, BDDK,
Özelleştirme İdaresi Başkanlığı, Kamu
Bankaları gibi kuruluşların yetkilileri de
katıldı.</Sentence>
  </Paragraph>
  <Paragraph Index="1">
    <Sentence Index="0">Bugünkü
görüşmelerde, ödemeler dengesi, savunma
sanayii, uluslararası rezervler, yerel
yönetimler, yatırım programı, kamu mali
yönetimi ve kontrol kanunu konuları
tartışıldı.</Sentence>
  </Paragraph>
</File>
```

4. Sonuçlar ve Yapılacak Çalışmalar

Bu yeni geliştirilen yöntem ile Türkçe cümlelerin sonları, önceden belirlenen kural ve kısaltma listeleri kullanılarak daha doğru ve verimli bir şekilde belirlenebilecektir.

Bu yöntem, cümle sonu belirleme alanında çalışan araştırmacılar için bir referans olabilmeyi

amaçlamıştır. Türkçe için belirsizlikler yaratan bazı durumlar (roma rakamları, kısaltmalar, sıralamalar, vb.) bu çalışma ile çözümlenmiştir. Bu çalışmada çözülemeyen belirsiz durumlar makine öğrenimi ve istatistiksel analizler kullanılarak çözümlenebilir. Yapının kolaylıkla anlaşılabilir, okunabilir ve esnek olması sebebiyle kelime türleri, kök ve ekleri bu yapıya kolaylıkla eklenip çok amaçlı bir derlem haline getirilebilir.

5. Teşekkür

Bu yazının yazılmasına vesile olan Dokuz Eylül Üniversitesi Bilgisayar Mühendisliği Bölümü öğretim elemanlarından Prof. Dr. R. Alp KUT, Doç. Dr. Yalçın ÇEBİ ve Araş. Gör. Derya BİRANT'a çok teşekkür ederim.

6. Kaynaklar

- [1] Shannon C.E. (1948): A Mathematical Theory of Communication, The Bell System Technical Journal, 27:379-423, sayfa 623-656.
- [2] Choi, S.W. (2000). Some Statistical Properties and Zipf's Law in Korean Text Corpus. Journal of Quantitative Linguistics, 7:1, sayfa 19- 30.
- [3] Nadas, A. (1984). Estimation of probabilities in the language model of the IBM speech recognition system. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32:4, sayfa 859-861
- [4] Kukich K. (1992). Technique for automatically correcting words in text. Periodical Issue Article of ACM Press, sayfa 77-439.
- [5] Church, K. & Gale, W. (1991). Probability Scoring for Spelling Correction. Statistics and Computing, sayfa 93-103.
- [6] Jurafsky, D. ve Martin, J.H. (2000). Speech and Language Processing, Prentice Hall, sayfa 193-199.
- [7] Güngördü Z. (1993). A lexical-functional grammar for Turkish. Yüksek Lisans Tezi. Bilgisayar Mühendisliği Bölümü , Bilkent Üniversitesi, Ankara.
- [8] Shannon, C.E. (1951). Prediction and Entropy of Printed English. The Bell System Technical Journal, 30:1, sayfa 50-64.
- [9] Crystal,D. (1991). A Dictionary of Linguistics and Phonetics, Blackwell, üçüncü basım.
- [10] Sinclair,J. (1991). Corpus Concordance, Collocation. OUP.