Patoloji Tanılarında Yazım Varyasyonları ve Hatalar

K. Hakan GÜLKESEN[a], Ahmet ERDEM[b], Beyza KAYMAKOĞLU[c], Umut ARIÖZ[d] Pınar YILDIRIM[d] Osman SAKA[a]

[a]*Akdeniz Üniversitesi, Antalya*
[b]*ODTÜ, Ankara*
[c]*Başkent Üniversitesi, Ankara*
[a]*Hacettepe Üniversitesi, Ankara*

Abstract

Typographic Errors Variants in Pathology Reports

The use of ICD-10 has been introduced to the daily practice of healthcare in Turkey. The coding of the diseases is not always easy, and some software applications have already been developed for searching disease codes. However, computers are not as "intelligent" as human beings, and when typographical errors and spelling variants occur, correct codes often cannot be found by the system. In order to design a successful search algorithm, the most frequent typographical errors and the spelling variants in the Turkish medical language must be known. The aim of this study is to obtain this knowledge by examining pathology reports. First, we developed a software program which splits the words in a text file, stores them in a database, and provides the tools for the user to link misspelled or variant words to the correct word. We used the diagnosis field of 11519 pathology reports of a private pathology centre. All the words in the reports were indexed and unnecessary words were deleted. We finally obtained 3244 different words. We found 274 different typos in this index, and observed one erroneous spelling of 204 words, two erroneous spellings of 22 words, three erroneous spellings of five words, four erroneous spellings of one word, and seven erroneous spellings of one word. Two hundred and sixty-five (97 %) of the errors were one or two letter(s) errors. Types of errors were "an extra letter in the middle" in 51 cases (19 %, gastrit-gastarit), "a shift of two letters" in 49 cases (18 %, bakteriel-bakteirel), "one letter error" in 46 cases (17 %, gögüs-göğüs), "missing a letter in the middle" in 35 cases (13 %, breakdown-beakdown), "two-letters errors" in 27 cases (10 %, eosinofilik-eosinofik). These five types of errors consist of 76 % of the total errors. A total of 118 words have variant spellings. Thirty-seven (31 %) of these were complex variants (Comissura-komissür). The remaining 69 % of the spelling variants were explainable by simple rules. The most common types of the variants were the absence of –is suffix (antrakozis- antrakoz) 25 times, the absence of –a suffix (adenoma-adenom), 21 times, and the k-c shift (akne-acne) six times.

The results of this study may help the design of better applications for searching medical terms. Search algorithms that are designed with the help of above characteristics would find most of the errors and variants in Turkish medical text related to diagnosis. We suggest that further studies in various medical centres are needed to enrich the information on this subject.

Key Words:
Information systems; ICD; Search engines

Özet

Ülkemizde ICD-10'un kullanıma girmesi ile birlikte kodların bilgi sistemleri yardımı ile taranması da gündeme gelmiştir. Ancak ne yazık ki bilgisayarlar insan kadar akıllı sistemler değildir, bir kelimenin farklı şekillerde yazılabilmesi veya bir harf hatası yapılması gibi nedenlerle aranan kod bulunamamaktadır. Taramalarda başarılı olan bir algoritmanın yazılabilmesi için Türkçedeki varyasyonlar ve yazım hatalarının bilinmesinde büyük yarar vardır.

Bu çalışmanın amacı patoloji raporlarında kullanılan tanıları inceleyerek sözü edilen bilgileri elde etmektir. Bunun için önce düz metinlerde yer alan sözcükleri indeksleyen, veritabanına aktaran ve yanlış yazılmış veya varyant olan sözcükleri asıl sözcükle ilişkilendirmeye yardımcı olan bir yazılım geliştirdik. Daha sonra bir özel patoloji merkezinin veritabanındaki 11519 raporun tanı bölümleri incelenerek, tanılarda kullanılan sözcükler indekslendi, gereksiz sözcükler temizlendi. Geriye kalan 3244 farklı sözcüğün analizi yapıldığında 274 yazım hatası saptandı. 204 sözcüğün bir farklı tipte, 22 sözcüğün iki farklı tipte, beş sözcüğün üç farklı tipte, birer sözcüğün dört ve yedi farklı tipte hatalı yazımlarına rastlandı. Yapılan hataların tipi incelendiğinde 274 hatadan 265'inin (%97) bir ya da iki harflik hatalar olduğu görüldü. Bunlardan 51'inde (% 19) arada bir harfin fazla olduğu (gastrit-gastarit), 49'unda (% 18) iki harfin yer değiştirdiği (bakteriel-bakteirel), 46'sında (% 17) olması gereken harf yerine başka bir harf yazıldığı (gögüs-göğüs), 35'inde (% 13) ortada bir harfin eksik olduğu (breakdown-beakdown), 27'sinde (% 10) iki harfin kayıp olduğu (eosinofilik-eosinofik) görüldü. Bu beş hata tipinin toplamı toplam hatanın % 76(208)'sını oluşturmaktadır. Kullanılan sözcükler incelendiğinde 118 sözcüğün doğru sayılabilecek farklı yazımları olduğu görüldü. Bunlardan 37 tanesi (% 31) karmaşık varyasyonlar (Comissura-komissür) iken, kalan % 69 varyasyonun çıkarsanabilir kurallara bağlı olduğu görüldü. Bu varyasyonlardan en sık görülenlerinin, 25 kez sözcükte –is soneki olup olmaması (antrakozis- antrakoz), 21 kez –a soneki olup olmaması (adenoma-adenom), altı kez k-c harflerinin birbiri yerine kullanılması (akne-acne) olduğu görüldü.

Bu çalışmada elde edilen bilgiler, kodların taranması için kullanıldığında faydalı olabilir. Bu bilgilerle hazırlanan tarama algoritmaları ile yanlış yazım ve varyasyonların çoğunluğunda aranan terimin bulunması mümkün olacaktır. Bundan sonraki dönemde, başka merkezlerin de verileri incelenerek bu bilgi birikiminin zenginleştirilmesini öneriyoruz.

Anahtar Kelimeler:
Bilgi sistemleri; Uluslararası Hastalık Sınıflaması, Arama motorları

1. Giriş

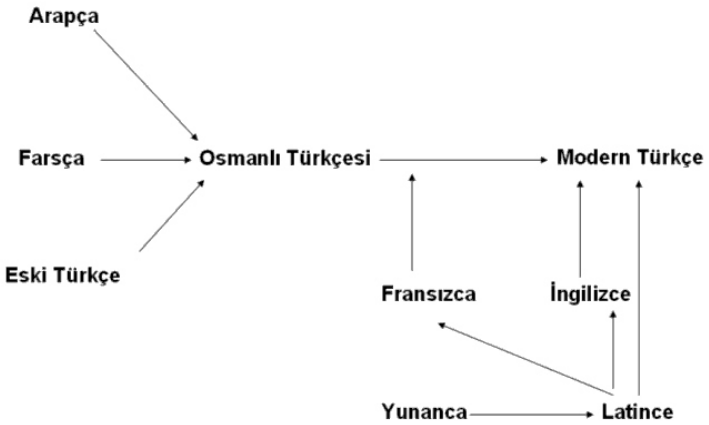
Hastalık tanılarının kodlanması, sağlık hizmetinin kalitesi, sağlık bilgisinin doru değerlendirilmesi ve sağlık yönetimi üzerine olumlu etkileri olan bir uygulamadır. Dünyada yaygın olan kodlama sistemlerinden birisi de Dünya Sağlık Örgütünce hazırlanan ICD (International Classification of Disease, Uluslararası Hastalık Sınıflaması) kodlama sistemidir. Ülkemizde de Sağlık Bakanlığı'nın kabul ettiği tanı kodlama sistemi ICD'nin onuncu sürümüdür.

Tanıların kodlanması, göründüğü kadar kolay değildir. Çok sayıda kod arasından hangisinin kullanılacağını belirlemek kimi zaman çok zaman alabilir. Doğru kodun bulunması için bilgisayar yardımına da başvurulabilir. Hazırlanacak arayüzler ile tanı içinde geçen sözcükler taranarak ilgili kodlar arasından kullanıcının seçim yapması sağlanabilir. Ancak, tarama yolu ile tanı seçmede bazı sorunlar vardır. Sistemde kodların karşılığı olan metinler vardır ve istenen kodun bulunabilmesi için tarama için kullanılan sözcük veya sözcüklerin bunlarla tam olarak eşleşmesi gerekmektedir. Bir harf bile yanlış yazıldığında istenen kod çıkmayabilir. Aynı sözcüğün farklı yazımları (varyasyonları) da olabilir. Bunlardan başka, eşanlamlı sözcükler de vardır ve bunlar da tarama için bir sorun oluştururlar.

Bu sorunların çözümü için basit eşleştirmeler yerine daha akıllı algoritmalar geliştirilmelidir. Bu şekilde, sistem kullanıcı yanlış da yazsa, varyasyon veya eşanlamlı da kullansa doğru kodu verebilir. Böyle bir algoritma geliştirmek için kullanılan dille ve yapılan hatalarla ilgili bazı veriler bize yardımcı olacaktır.

Kelime işlemcilerin ortaya çıkmasından bu yana yazım hataları ilgi çeken bir çalışma alanıdır. Erken dönem çalışmalardan birinde sağlıkla ilgili olmayan İngilizce metinlerde yapılan hataların % 80'inin bir harflik hatalar olduğu bildirilmiştir [1]. İngilizce tıbbi kayıtların gözden geçirildiği bir çalışmada ise çok daha karmaşık hatalar bildirilmiştir [2]. Bu çalışmaya göre örneğin "abscess" (apse) sözcüğü tıbbi kayıtlarda 92 farklı türde yazılmakta, bu sözcük kullanıldığında % 45 olasılıkla yanlış yazılmaktadır.

İngilizce tıbbi metinlerdeki sözcüklerin yazımında yapılan hataların % 31'inde üç veya daha fazla harf hatası vardır [2]. Bildiğimiz kadarı ile Türkçe'de yapılan yazı hatalarının tipleri ve sıklığı ile ilgili bir çalışma yoktur. Ancak, Türkçe alfabenin fonetik bir alfabe olması nedeni ile İngilizce'deki kadar sık yazım hataları ile karşılaşılmayacağı düşünülebilir.
Taramalardaki bir diğer sorun da yazım varyasyonlarıdır. Acne-akne ikilisinde olduğu gibi, kelimenin birden fazla doğru yazımı vardır. Türkçe tıp dili, coğrafi ve kültürel nedenlerle başka dillerden çok sayıda sözcük transferi yapmıştır (Şekil 1). Türkçenin tıp eğitimde kullanılması Osmanlı İmparatorluğunun son dönemlerinde başlar [3]. Başlangıçta Fransızcanın etkisi belirginken, zaman içinde İngilizcenin etkisi artar. Türkçeye giren yabancı sözcükler kimi zaman olduğu gibi yazılmakta, kimi zaman fonetik transkripsiyonu kullanılmaktadır. Enfeksiyon-infeksiyon örneğinde olduğu gibi biri Fransızca diğeri İngilizce etkisinde iki farklı yazım da olabilir. Türkçenin sondan ekli bir dil olması nedeni ile ek alan sözcükler de diğer bir sorun kaynağıdır.



Şekil 1: Türkçe tıp terminolojisinin oluşumunun basitleştirilmiş şeması.

Bu çalışmada elektronik olarak kaydedilmiş patoloji raporlarında tanılar incelenerek, Türkçe tıbbi metinlerde yer alan hataların tipleri, kullanılan varyasyonlar incelenecektir. Elde edilen bilgilerin metin tabanlı tarama tasarımlarının geliştirilmesine yardımcı olacağını düşünüyoruz.

2. Gereç ve Yöntem

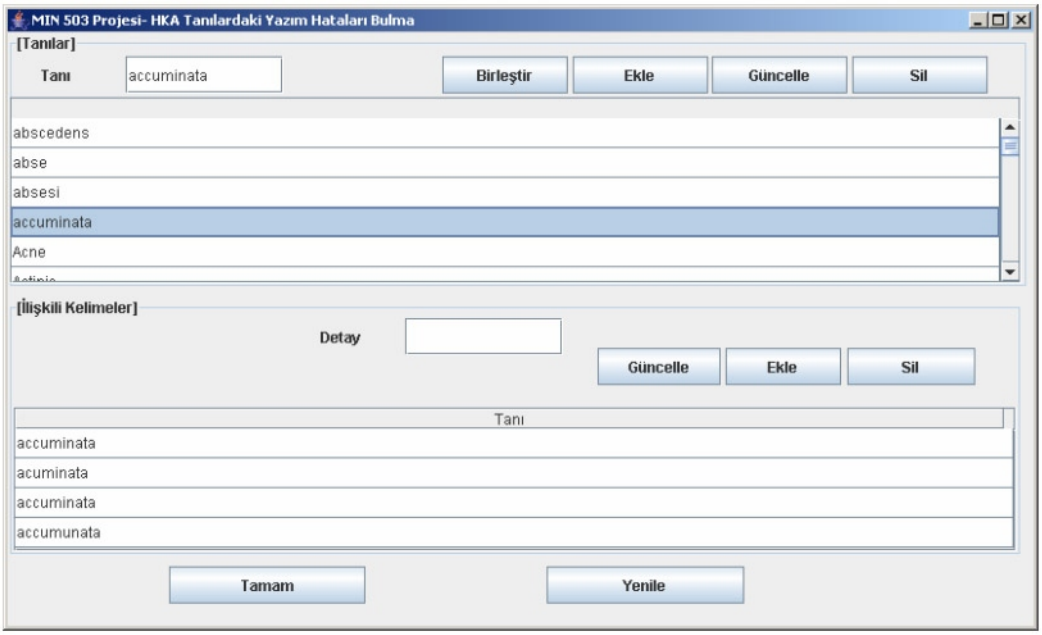
Bu çalışmada özel bir patoloji merkezinin (Çağdaş Patoloji Merkezi, Antalya) veritabanında yer alan 11519 raporun tanı kısımları incelenmiştir. Tanılar serbest metin olarak kaydedilmiş, herhangi bir kodlama kullanılmamıştır. Raporlar beş yıllık bir dönemi kapsamaktadır.

Çalışmamızda kullanılmak üzere düz metni sözcüklere ayırıp bir veritabanına yerleştiren ve indeksleyen, daha sonra da bu sözcüklerden yanlış yazılmış olanları ve varyantları bir kullanıcının asıl sözcüklerle ilişkilendirmesine yardımcı olan bir program geliştirdik. Bu yazılımı geliştirirken Java programlama dili (jdk 1.5.1) ve MySQL veritabanı (v4.1) kullanıldı. Geliştirdiğimiz program aşağıdaki işlemleri yapmaktadır:

- Tanıları sözcüklere parçalayarak sözcükleri veritabanına aktarma
- Duplike(tekrarlı) sözcükleri ayıklama.
- Anahtar sözcükler saptayıp benzer sözcükleri bunlarla ilişkilendirme.
- Kullanıcının yeni anahtarlar oluşturması, varolan anahtarları silebilmesi, iki anahtarı birleştirebilmesi ve anahtarlarla ilişkilendirilmiş sözcüklerin ilişkilerini değiştirebilmesini sağlama.

Program mimarisi dört katmandan oluşmaktadır. Bunlar aşağıdaki gibidir:

▪Utility classes (yardımcı sınıflar)
▪BLO: Business Layer (iş katmanı)
▪DAO: Data Access Layer (veri erişim katmanı)
▪GUI: Graphical User Interface (grafik kullanıcı arayüzü)
▪Grafik arayüzünde kullanıcı anahtarları ve bunlarla ilişkili sözcükleri hiyerarşik olarak görebilir (Şekil 2). Kullanıcı bu arayüzden anahtarlarla ilgili aşağıdaki değişiklikleri yapabilir:
▪Birleştir: İki anahtarı tek bir terim altında birleştirir. Detaylar (anahtarla ilişkili sözcükler) da birleşir.
▪Ekle: Yeni bir anahtar ve aynı isimli yeni bir detay ekler.
▪Güncelle: Seçili anahtarı değiştirir.
▪Sil: Seçili anahtarı ve bunun detaylarını siler.
▪Detay kısmı için aşağıdaki işlemler mümkündür:
▪Güncelle: Seçili detayı değiştirir.
▪Ekle: Yeni bir detay ekler.
▪Sil: Seçili detayı siler.



Şekil 2: Hazırlanan yazılımın ekran görüntüsü.

3. Bulgular

Patoloji raporlarında kullanılan kelimeler içinden niteliksiz sıfatlar, bağlaçlar ve sayılar temizlendi. Geriye kalan 3244 sözcük yazım hataları ve varyasyonlar yönünden incelendi.
İki yüz yetmiş yazım hatası bulundu. Bunlar 232 farklı kelimenin yazımı idi. İki yüz dört sözcüğün bir farklı şekilde yanlış yazıldığı, 22 sözcüğün iki şekilde yanlış yazıldığı, beş sözcüğün üç farklı şekilde yanlış yazıldığı, bir sözcüğün dört, bir sözcüğün de yedi farklı şekilde yanlış yazıldığı görüldü. Yanlış yazımlara bazı örnekler Tablo 1'de görülmektedir. Hata tipleri incelendiğinde % 48.27'sinde yanlış harf yazımı, % 22.18'inde eksik harf, 23.27'inde fazla harf, % 7.28'inde transpozisyon (harflerin yer değiştirmesi) görüldü. Bu hataların daha ayrıntılı dökümü Tablo 2'de görülmektedir.

Tablo 1: Yanlış yazılan bazı sözcükler.

Doğru yazım	**Yanlış yazım**
Bakteriel	Bakteirel Bakterile Baktreriel
Biopsi	Biospi biopisi
değişiklikler	değişikler değişilklikler
endometrial	ednometrial endometrail
epitelial	epitelail epitetelial
epiteloid	Epiteliod epiteolid
metaplazisi	metapazisi metaplazisisi metaplazsi mteplazisi
Nonspesifik	Nonpspesifik Nonsepsifik nonsipesifik Nonspesfifik Nonspesik Nonspesipesifik Nosnpesifik
Psödoepiteliomatöz	Pseudoepiteliamatöz Psödoepiteliomatoz Psödoepiteliamatöz
Reaktif	raektif reekatif
Sitopatolojik	sitoaptolojik sitolopatolojik
Skuamöz	skumöz Squamöz Squmaöz
Süperfisyel	Süperfisye süperfiyel

Yapılan hataların tipi incelendiğinde 274 hatadan 265'inin (%97) bir ya da iki harflik hatalar olduğu görüldü. Bunlardan 51'inde (% 19) arada bir harfin fazla olduğu (gastrit-gastarit), 49'unda (% 18) iki harfin yer değiştirdiği (bakteriel-bakteirel), 46'sında (% 17) olması gereken harf yerine başka bir harf yazıldığı (gögüs-göğüs), 35'inde (% 13) ortada bir harfin eksik olduğu (breakdown-beakdown), 27'sinde (% 10) iki harfin kayıp olduğu (eosinofilik-eosinofik) görüldü. Bu beş hata tipinin toplamı toplam hatanın % 76'sını (208) oluşturmaktadır.

Yazımda varyasyonlara da bakıldı, toplam 118 sözcüğün yazımında varyasyonlara rastlandı. Bu sözcüklerden iki tanesinin dört farklı doğru yazımı, bir tanesinin üç farklı yazımı vardı, diğer 115 sözcüğün ikişer doğru yazımı vardı. Tablo 3'te varyasyonların özellikleri görülmektedir.

Varyasyonlardan 37 tanesi (% 31) karmaşık varyasyonlar iken, kalan % 69 varyasyonun çıkarsanabilir kurallara bağlı olduğu görüldü. Bu varyasyonlardan en sık görülenlerinin, 25 kez sözcükte –is soneki olup olmaması (antrakozis- antrakoz), 21 kez –a soneki olup olmaması (adenoma-adenom), altı kez k-c harflerinin birbiri yerine kullanılması (akne-acne) olduğu görüldü.

4. Tartışma

Türkçe patoloji tanılarında yapılan inceleme sonucu, yazım hatalarının % 97'sinin bir-iki harf ile ilgili hatalar olduğu görüldü. Bu bilgi, tarama arayüzlerinin tasarımı açısından önemlidir. Bir-iki harf hataya duyarlı arama algoritmalarının kullanılması durumunda yazım hatalarından kaynaklanan sorunların büyük ölçüde giderilmesi mümkün görünmektedir. Daha önce İngilizce tıp metinleri ile yapılan bir çalışmada hataların daha karmaşık yapılarda olduğu bildirilmişti [2]. Türkçe alfabenin fonetik bir alfabe olması nedeni ile yazım hatalarının daha az olduğunu düşünüyoruz.

Tablo 2: Patoloji raporlarında görülen yazım hatalarının tipleri.

Hata tipi	**Örnekler**		**Olgu sayısı**
	Doğru	**Yanlış**	
Boşluk unutulmuş	bronş epitel	Bronşepitel	11
De'den önce boşluk unutulmuş	Komponentinin de	Komponentininde	3
Uygunsuz boşluk	lerden	…lerden	4
Uygunsuz kısaltma	metastaz	Met	2
Bir harf yanlış yazılmış A - e E - a O - a i - I K - c G - ğ Ğ - g i - y i - u t - d Diğer	 Adale intervertebral kardio Anjiomyoma Klaviküler pigmenti göğüs enteropati insisura irritated	 Adele intervertabral kardia Anjıomyoma Claviküler piğmenti gögüs enteropaty insusura irritaded	 3 6 4 4 5 2 2 5 2 2 11
Harf transpozisyonu	Bakteriel	Bakteirel	49
Çift transpozisyon	differansiasyon	Diffrenasiasyon	2
Sonda fazla harf 1 harf 2 harf	 Akut sebase	 Akuta Sebaseus	 2 2
Ortada fazla harf	Gastrit	Gastarit	51
Fazla hece	epitelial	Epitetelial	8
Fazla çift hece	Nonspesifik	Nonspesipesifik	1
Sonda harf kaybı 1 harf 2 harf	 Dysmenorrhea aspirasyon	 Dysmenorrhe Aspirasy	 8 3
Ortada harf kaybı 1 harf 2 harf	 Breakdown eozinofilik	 Beakdown Eozinofik	 35 6
Tek l	folliküler	Foliküler	3
Tek k	kokkobasil	Kokobasil	2
Kayıp hece	değişiklikler	Değişikler	4
Karmaşık hata 2 harf 3 harf 2 transpozisyon 1 fazla harf	 Arias aküminata perforasyonu	 Area akimülata Perforasoyuna	 27 5 1

Tablo 3: Yazım varyasyonlarının özellikleri (bir kez görülen olgular tabloda verilmemiştir).

Fark	**Örnek**	**Olgu sayısı**
K=c	Acne-akne	6
Ph=f	Atrophicus-atrofikus	3
I=y	Epithelioma-epitelyoma	2
Ch=k	Brankial=Branchial	4
Th=t	Epitelial-epithelial	3
St=z	Desmoplastik-Desmoplazik	2
U=ü	Peritubuler-peritübüler	2
X=ks	Hemitorax-hemitoraks	3
Ous=öz	Lentiginöz-lentiginous	2
-a eki	Adenoma-adenom	21
-is eki	Antrakozis- Antrakoz	25
Karmaşık fonetik değişimler	Comissura-komissür	37

Öte yandan yazım varyasyonlarının % 69'unun çıkarsanabilir basit kurallardan kaynaklandığı görüldü. Yazım varyasyonları açısından algoritma oluşturmak daha zor görünmektedir, çünkü varyasyonların % 31'i karmaşık kurallara göre olmaktadır. Türkçe tıp dilinde bir standartlaşma sorunu olduğu açıktır. Yabancı dillerden gelen sözcüklerin yazımı konusunda bir birlik yoktur. Bugünkü haliyle Türkçe tıp dili bir hayli seçmeci (eklektik) bir görüntü sergilemektedir. Sağlık personelinin yaşı, bildiği yabancı dil, mensup olduğu ekol ve çalışma alanının kullandığı terminoloji üzerinde etkili olacağı kolayca tahmin edilebilir. Bu çalışma tek bir merkezin raporlarında yapılmıştır, çokmerkezli bir çalışmada varyasyonların sayı ve tür olarak daha da fazla olacağı düşünülebilir.
Tarama açısından eşanlamlılar ayrı bir sorundur. Bu çalışmada metinler eşanlamlılık yönünden incelenmemiştir.

5. Sonuç

Elde edilen sonuçlara göre, uygun tasarımlarla taramalarda yazım hatalarının tamamına yakını, yazım varyasyonlarının üçte ikisinin yakalanması mümkündür. Daha farklı alanlarda ve coğrafi yerleşimlerde buna benzer çalışmalar yapılması Türkçe tıp dilinde sık rastlanılan hatalar ve varyasyonların incelenmesi, daha verimli tarama arayüzlerinin tasarımına yardımcı olacaktır.

6. Teşekkürler

Bu çalışmada verilerinin kullanılmasına izin veren Dr. Murat Şedele, Dr. Müjgan Yaz ve Dr. Faruk Güleç'e teşekkürlerimizi sunarız.

7. Kaynakça

[1] Damerau F. A technique for computer detection and correction of spelling errors. Communications of the ACM 1964; 7: 171-6.

[2] Shapiro AR; Centers for Disease Control and Prevention (CDC). Taming variability in free text: application to health surveillance. MMWR Morb Mortal Wkly Rep 2004; 53 Suppl: 95-100.

[3] Kaadan AN. The Ottoman Medical School of Damascus and its effect on medicine teaching in Syria. JISHIM 2002; 2: 27-29

8. Sorumlu Yazarın Adresi

Dr. K. Hakan Gülkesen, Akdeniz Üniversitesi Tıp Fakültesi Biyoistatistik ve Tıp Bilişimi AD, Antalya. e-posta: hgulkesen@akdeniz.edu.tr