Journal of İstanbul Kültür University
2006/2 pp. 127-132

TÜRKÇE VE İNGİLİZCE METİNLERİN EĞİLİMİNDEN ARINDIRILMIŞ DALGALANMA ANALİZİ YÖNTEMİ İLE İNCELENMESİ

Gökhan ŞAHİN[1], Murat ERENTÜRK[1], Avadis HACINLİYAN[1]

ÖZET

Türkçe ve İngilizce metinlerde uzun sureli korelasyonları bulmak için eğilimden arındırılmış dalgalanma analizi kullanılmıştır. Bağımlı değişken olarak kelime sıklığı yerine metindeki kelimelere, bu kelimeleri oluşturan harflere dayalı birer değer verilmek suretiyle DNA rassal yürüyüşlerden esinlenen bir yaklaşım tercih edilmiş, ikisi aynı metnin İngilizce ve Türkçesi olmak üzere dört farklı metin bu yaklaşımla incelenmiştir. Sonuç olarak İngilizce metinlerin korelasyonlarının (her ne kadar metinlerin içeriği birbiri ile ilgisiz olsa da) birbirine benzediği, benzer bir davranışın Türkçe metinlerde de olduğu gözlenmiştir. Buna karşı eğilimden arındırılmış analiz detaylarında iki dili birbirinden ayıran davranış biçimleri ve her iki iki dilin kendine has korelasyonları olduğuna işaret eden özellikler gözlenmiştir. Kelime sıklığı gibi kullanılan dağarcığın özelliklerine dayanmayan bu değişkenle dilleri birbirinden ayıran özeliklerin gözlemlenmesi önerinin ilginç yanıdır.

1. Giriş

Doğal dillerde hem rassal yapılar, hem de bir sonraki adımın bir önceki ile bağımlı olması şeklindeki rasgele yürüyüşü anımsatan bir düzen vardır. Bu amaçla yapılan ilk çalışmalar değişik dilbilimsel formların istatistiksel sıklıklarını gösteren sıklık tablolarının (corpus) oluşturulması ve üzerinde istatistiksel analizlerin yapılmasıdır. Değişik metinlerden alınan örneklerde bile belirli bir istikrar gözlenmiştir. Lineer istatistik yöntemlerine ek olarak değişik araştırıcılar, doğal dillerde kendine benzer fraktal yapıların yer aldığını da ileri sürmüşlerdir. Örnek olarak Montemurro ve Purry [1] Shakespeare'in 36 oyununda bu tür bir davranış gözlemlemişlerdir. Bahm ve başkaları[2] bu durumun İngilizce dilinin mi Shakespare'in üslubuna ait bir özellik olup olmadığını araştırmak için aynı analizi hem bu oyunların Kore diline tercümeleri hem de Koreli yazarların popüler romanları üzerinde gerçekleştirmişlerdir. Öte yandan üslup determinizm ilkesinin her zaman uygulanabileceği bir özellik değildir. Örneğin kullanılan kelime sayısında İncil'de 5649, Dante'de 5860 (bunlardan 1615'i özel isim), Milton'da 8000 kadar, Shakespeare'de 15000 şeklinde önemli farklılıklar vardır.

Bu araştırıcıların kullandıkları analiz yöntemi Hurst R/S analizidir. [3] Hurst üsteli tüm Korece metinler için birbirine yakın çıkmıştır. Bu bulgu nedeniyle bu veya benzeri bir fraktal parametrenin dilleri karakterize edebileceği düşünülebilir. Analiz için Zipf'in [4] önerisine paralel sıklıklardan oluşturulan sıralamadan yararlanılmıştır. Zipf ve diğer dilbilimciler, Zipf ve bunun genellemelerini oluşturan kanunları "En az gayret ilkesi" (Least Effort Principle) şeklinde, doğal dil ile insan psikolojisi ilişkisi esasında irdelemektedir. Rassal seçim ve Markov süreci sonucunda kendi kendine düzenlenen kritiklik oluşabilmektedir. Buna karşılık Matematikçiler ise bu kanunları bir büyüme ve doğal seçim süreci olarak görürler.

Zaman serisi ile ilgili çalışmalarımızın, en azından Türk Dil Kurumunun sıklık derlemesi (corpus) için sıklıkların sıralamaya göre daha tercih edilebilir bir değişken olduğunu gösterdiği, bunun da büyük bir olasılıkla derlemenin sınırlı olmasından kaynaklandığı dikkate alınarak bu çalışmada kelimeyi oluşturan harflere dayanan bir diğer değişken önerilmiştir.

[1]Yeditepe Universitesi Fizik ve Yönetim Bilişim Sistemleri Bölümleri

Eğilimden arındırılmış dalgalanma analizi (Detrended Fluctuation Analysis) durağanlaşmamış (nonstationary) zaman serilerinde uzun zamanlı korelasyonları bulmak için kullanılan önemli bir araç haline gelmiştir [5,6]. DNA [7-9], kalp atışı dinamiği [9-14] uzun zamanlı hava tahmini, ekonomik zaman serileri [16-19] ve katı hal dinamiği bu yöntemin başarıyla uygulandığı alanlar arasında bulunmaktadır. Doğal dillerde de durağan olmayan bir yapı arandığı için bu analizin R/S analizine olan üstünlüğü dikkate alınarak bu yöntem tercih edilmiştir.

Bu analizi Türkçe'ye uyarlarken DNA rasgele yürüyüşlerinden (DNA Random Walk) esinlenilmiştir[19]. Analizin kullanılmasında her dilin kendine ait bir korelasyonu olduğu ve metinlerin farklılıklarının birer eğilim gibi düşünülebileceği varsayımları etkili olmuştur.

Bu yöntemi kullanırken aşılması gereken ilk zorluk metinden zaman serisi oluşturmaktır. Bunun için metindeki her kelime rasgele yürüyüşte belirli bir adım olarak alınmıştır. Her adımın (kelimenin) uzunluğu içerdiği harfler tarafından şöyle belirlenmektedir

$$s(n) = \sum_{i=1}^{N} y(i) .$$

Burada N kelimenin uzunluğuna karşılık gelmektedir ve değeri kelimeyi oluşturan harflerin ($y(i)$) ascii koduna eşittir. Bu yöntemin tek sakıncası aynı harflerden oluşan değişik kelimelerin aynı değeri almasıdır, ancak bu durumun çok sık olmadığı varsayılmıştır.

2. Eğilimden arındırılmış dalgalanma analizi

Durağanlaşmamış (nonstationary) zaman serilerindeki korelasyon özelliklerini nitelendirilmek için kullanılan, zaman serisindeki boyutsuzlaşma üstelinin (bundan sonra α ile belirtilecek) değiştirilmiş karekök ortalama (root mean square) yöntemiyle hesaplanmasına dayanan [6] bir yöntemdir.

α 'yı $x(i)$ $[i = 1,....,N]$ ile gösterilen bir zaman serisinden hesaplamak için önce zaman serisi aşağıdaki gibi entegre edilir:

$$y(k) = \sum_{i=1}^{k} x(i) - \overline{x} .$$

Burada $\overline{x}$ zaman serisinin ortalama değeridir. k ise 1 ile N arasında değişmektedir.

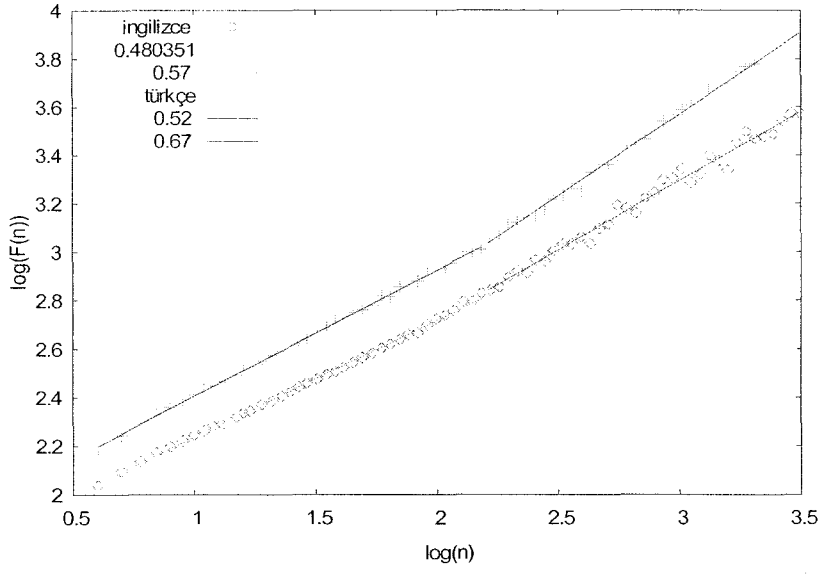
Daha sonra $y(k)$ n uzunluğunda eşit kutulara bölünür. Her kutudaki verilere en küçük kare (least squares fit) yöntemiyle bir doğru uydurulur ($y_n(k)$). Daha sonra entegre edilmiş zaman serisi yerel eğim ($y_n(k)$) çıkarılmak suretiyle eğilimlerinden arındırılır. Eğilimden arındırılmış zaman serilerinin karekök ortalama (root mean square fluctuation) dalgalanması ($F(n)$) şöyle hesaplanır:

$$F(n) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}[y(k) - y_n(k)]^2} .$$

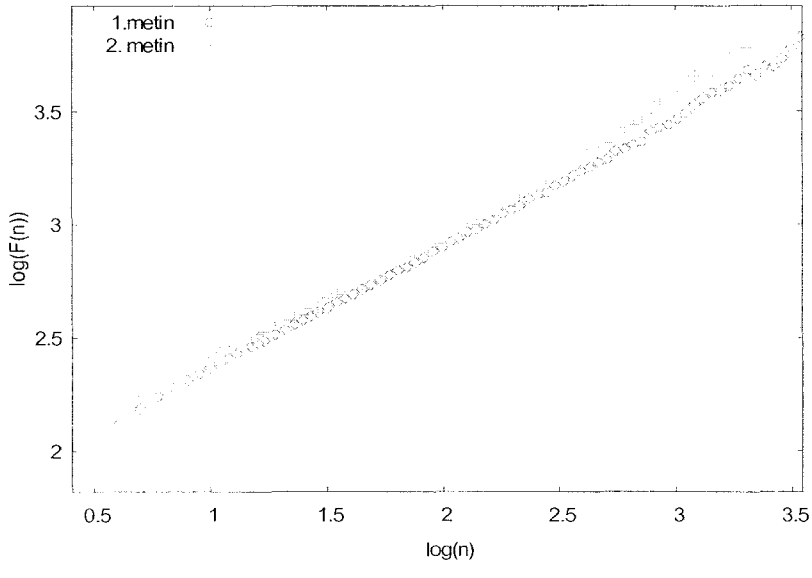
$F(n)$ tüm n değerleri için hesaplanır. $\log(F(n))$, $\log(n)$ grafiğinin eğimi α 'yı verir. α ile *1/f* eğimi arasında *1/f* eğimi= $2\alpha - 1$ şeklinde bir ilişki vardır.

3. Metinlerin incelenmesi

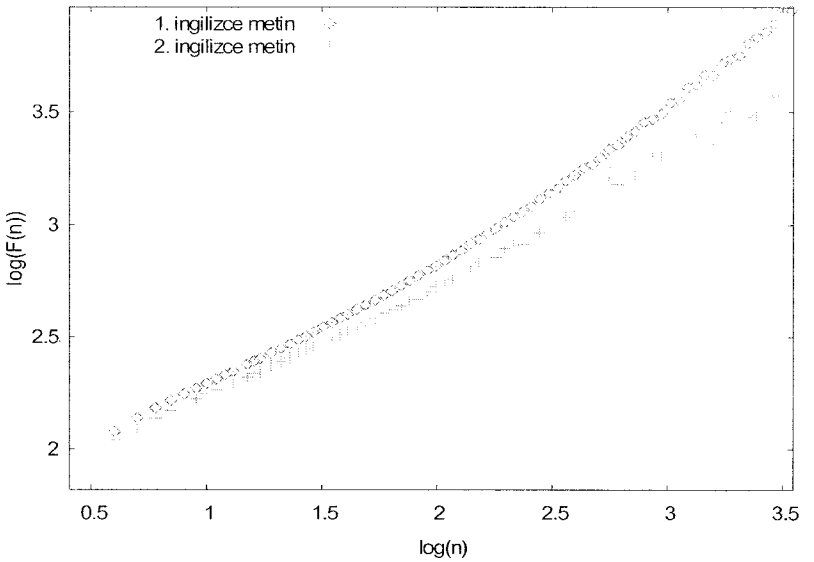
Analiz için ikisi aynı metnin İngilizce ve Türkçesi olmak üzere dört metin kullanılmıştır. Aşağıdaki grafikte birbirinin tercümesi olan iki metnin eğilimden arındırılmış dalgalanma analiz sonuçları (DFA) görülmektedir. Eğimlerdeki kırılmalar dillerin korelasyon özelliklerinde değişimi göstermektedir. Kırılmanın Türkçe'de İngilizce'ye göre daha belirgin olduğu söylenebilir. Ayrıca her iki rejimdeki eğim Türkçe için %10 mertebesinde daha fazladır. Eğimlerin değerleri grafiğin üstünde gösterilmiştir. İki dil arasındaki korelasyon farklılıkları oldukça belirgindir.



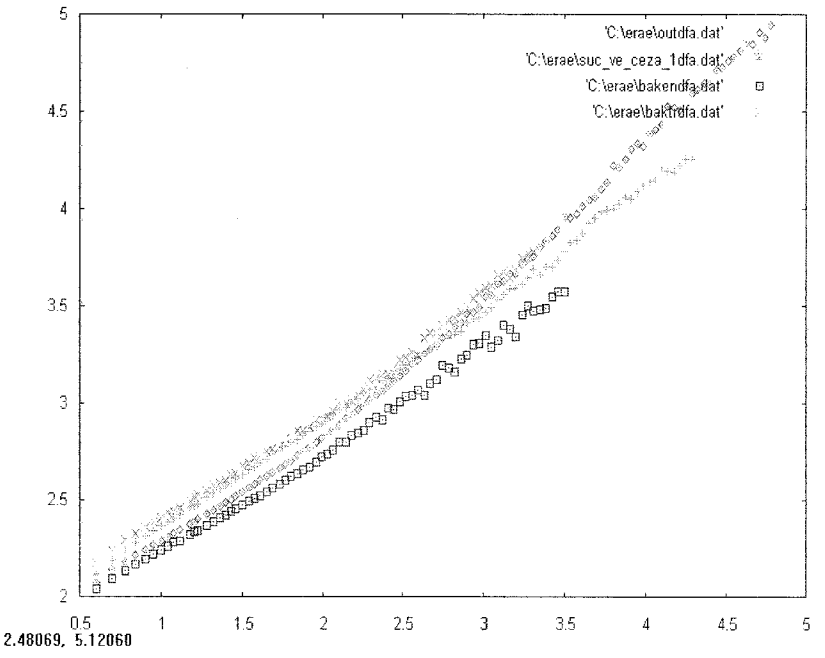
Yukarıda gözlenen korelasyonların metinlere değil de dillere ait olduğu savını güçlendirmek için aynı analiz önce içerik açısından tamamıyla farklı iki Türkçe metine uygulandı. Bu uygulama sonucunda aşağıda görülebileceği gibi iki metnin korelasyon yapısı birbirine benzer çıktı.



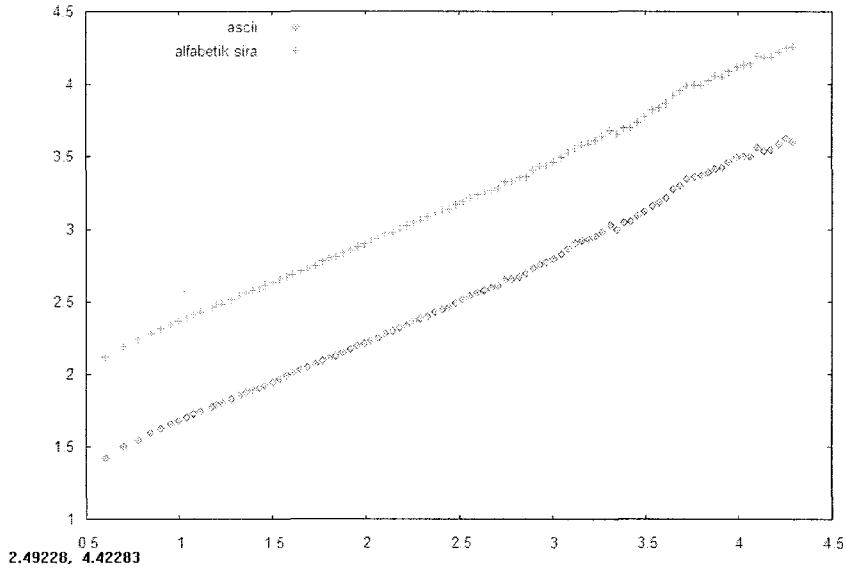
Aynı analizi içerik olarak birbirinden tamamıyla farklı iki İngilizce metine uyguladığımızda ise aşağıdaki grafikte de görüldüğü gibi yine birbirine benzer korelasyon özellikleri görmekteyiz.



Metinlerin tümünün DFA analiz sonucu aşağıda aynı grafik üzerinde beraberce verilmiştir. İlk bakışta tüm dillerde benzer bir rassal yürüyüş yapısı olduğu söylenebilirse de ortalama değerlerdeki farklılıklar nedeniyle eğrilerin arasında bir fark olup olmadığını varyans analizi gibi bağımsız bir yöntem ile değerlendirmemiz gerekir. Yapılan değerlendirmenin sonucunda 338 e karşı 3 serbestlik dereceli F oranı 15.9 olarak bulunmuştur. İki dağılımın farklı olmama olasılığı 10^{-9} mertebesindedir.



Eğer zaman serisi oluşturulurken harflerin ascii kodu yerine alfabetik sıralaması kullanılırsa elde edilen sonuçların niteliğinin değismediği fark de gözlemlenmiştir. Aşağıdaki şekilde, aynı Türkçe metinlerden, iki farklı yöntemle oluşturulmuş zaman serisi üzerinde uygulanmış olan eğimden arındırılmış dalgalanma analizi sonuçları gösterilmektedir. Iki doğrunun da eğimleri ve eğimlerinin kırılma noktaları birbirine paraleldir.



Sonuç:

Harflere değer verilerek oluşturulan bir bağımlı değişkenin de en az Zipf kurallarına dayanan sıklık değerleri kadar uygun bir bağımsız değişken olduğu, bu değişkene dayalı DFA analizinin dillerin dinamik yapısını aydınlatabileceği gözlemlenmiştir.

Analize başlarken temel varsayımlar her dilin kendine has uzun zamanlı ve kelime komşuluğu olarak sınırlı bir pencerede korelasyonu olduğu ve bu korelasyonun yakalanması için dilbilim bakımından anlamı olan en ufak birimine bakılması gerektiğiydi. Yapılan analizler bu iki varsayımı da doğrular nitelikte sonuçlar vermiştir. Ayrıca, önerilen değişken seçiminin dilleri birbirinden ayırabileceği de anlaşılmıştır.

Kaynakça:

[1] MA Montemurro, PA Pury , "Long-range fractals correlations in literary corpora", Fractals 10, 451 2002 .

[2] J. Bhan, S. Kim, J Kim, Y Kwon, S. Yang , K. Lee, Chaos, Solitons and Fractals 29, 69 2006.

[3] H. E. Hurst, Trans Am Soc Civil Eng 116, 770 1951.

[4] G. K. Zipf, *Human Behaviour and the Principle of Least-Effort*, Addison-Wesley, Cambridge MA, 1949

[5] J.W. Kantelhardt, E. Koscielny-Bunde, H.H.A. Rego, S. Havlin, and A. Bunde Physica A 295, 441 (2001)

[6] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger, Phys. Rev. E 49 1685 (1994).

[7] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons, H.E. Stanley, Phys. Rev. E 51 5084. (1995)

[8] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, R.N. Mantegna, M. Simons, H.E. Stanley, Physica A 221 (1995) 180.

[9] S.V. Buldyrev, N.V. Dokholyan, A.L. Goldberger, S. Havlin, C.-K. Peng, H.E. Stanley, G.M. Viswanathan, Physica A 249 (1998) 430.

[10] C.-K. Peng, J. Mietus, J.M. Hausdorff, S. Havlin, H.E. Stanley, A.L. Goldberger, Phys. Rev. Lett. 70 (1993) 1343.

[11] C.-K. Peng, S. Havlin, H.E. Stanley, A.L. Goldberger, Chaos 5 (1995) 82.

[12] K.K.L. Ho, G.B. Moody, C.-K. Peng, J.E. Mietus, M.G. Larson, D. Levy, A.L. Goldberger, Circulation 96 (1997) 842.

[13] G.M. Viswanatha, C.-K. Peng, H.E. Stanley, A.L. Goldberger, Phys. Rev. E 55 (1997) 845.

[14] C.K. Peng, J.M. Hausdorff, S. Havlin, J.E. Mietus, H.E. Stanley, A.L. Goldberger, Physica A 249 (1998) 491.

[15] Y.H. Liu, P. Cizeau, M. Meyer, C.-K. Peng, H.E. Stanley, Physica A 245 (1997) 437.

[16] P. Cizeau, Y.H. Liu, M. Meyer, C.-K. Peng, H.E. Stanley, Physica A 245 (1997) 441.

[17] M. Ausloos, N. Vandewalle, P. Boveroux, A. Minguet, K. Ivanova, Physica A 274 (1999) 229.

[18] M. Ausloos, K. Ivanova, Physica A 286 (2000) 353.

[19] Alexandre Rosas, Edvaldo Nogueira, Jr., and Jose F. Fontanari, PHYSICAL REVIEW E 66, 061906 (2002).