

ZAMAN SERİSİ ANALİZİ YÖNTEMİYLE TÜRKÇEDE OLASI KAOTİK YAPILARIN ARAŞTIRILMASI

Murat ERENTÜRK¹, Gökhan ŞAHİN¹, Avadis HACINLIYAN¹

Özet

Uzunlukları bin ile üç bin kelime arasında değişen Türkçe metinler üzerinde önceden Türk Dil Kurumu tarafından saptanan kelime sıklıkları bağımlı değişken olarak kullanılarak zaman serisi analizi yöntemiyle korrelasyon fonksiyonu, karşılıklı bilgi fonksiyonu, yanlış en yakın komşular ve Liapunov üstelleri gibi muhtemel kaotik yapılar araştırılmış ve bu tür yapıların tutarlı olarak gözlemlendiği anlaşılmıştır. Benzer bir analiz İngilizce metinler üzerinde de tekrarlanarak iki dil arasındaki muhtemel benzerlik ve farklar incelenmiştir.

1. GİRİŞ

Doğal dillerin yapısı ve içindeki hiyerarşik düzen dilbilimcilerin esas ilgi alanı olmuştur. Bu hiyerarşik yapıda en alt düzeyi oluşturan hece ve harfler seslendirme bakımından önemli olmakla birlikte anlam bakımından önem teşkil etmemektedirler. Kelime ve bunun üstündeki düzeylerde dilbilimcilerin ağırlıklı üzerinde durduğu anlam ortaya çıkar. Doğal diller zaman içinde evrim gösteren dinamik bir yapıya sahiptir. Geçirmekte olduğu evrim süreci karmaşıklık ve doğal seçim gibi iki önemli kavramla yakından ilgilidir. Bu evrim sürecine uyan yapılar seslerin ardarda gelmesiyle kısa süreli olarak evrim bakımından da çok uzun süreli olarak zamanda karmaşık ve çok ölçekli bir yapıya sahiptirler.

Son zamanlarda hızlı bir gelişme gösteren ve doğal yapıların uyduğu fraktal geometri kavramı, doğal dillerin iç yapısını tasvir etmek, çeşitli dil gurupları arasındaki farklılık derecelerini anlayabilmek üzere yararlanabileceğimiz bir araçtır. Yazılı metinler ise ard arda gelen kelimelerden oluşan bir zaman serisi haline getirirler. Uygun bir parametrisasyon bulunabilirse yazılı metinler üzerinde hece, kelime, cümle, vb. gibi değişik katmanlarda yapılacak zaman serisi analizlerinin birbirine yakın sonuçlar vererek dilin değişik ölçeklerde kendi kendine benzeyen bir karakter sergileyip sergilemeyeceği araştırılabilir.^[1] Bu nedenle, doğal dillerin yapısı sadece dilbilimcileri değil başta istatistik, matematik ve fizik olmak üzere birçok disiplini ilgilendirir. Çalışmanın amacı Zipf^[2] kanunlarına uygun olarak kelimelerin bir derlemede gözlemlenen sıklıklarını bağımsız değişken olarak kullanmak suretiyle elde edilen zaman serilerinde kaotik yapıları araştırmaktır. Aynı analiz yönteminin İngilizce metinler üzerinde de uygulanarak bu iki dil arasında muhtemel benzerlikler veya farklılıklar belirlenebilecektir. Metinler üzerinde zaman serisi analizi yapılarak fractal yapıların belirlenmesi için önemli bir başlangıç noktası çeşitli yapısal katmanlardan toplanmış sıklık verisidir. Kelime katmanında İngiliz dili için yapılmış bu tip bir çalışma mevcuttur.^[3] Benzer bir çalışma da yakın bir geçmişte Türkçe için yapılmıştır.^[4] Her iki çalışma da, incelenen metin yığımında geçen bütün kelimeleri ve bunların metin içerisinde kaçar defa geçtiklerini listelemektedir.

¹ Yeditepe Üniversitesi Fizik ve Yönetim Bilişim Sistemleri Bölümleri

2. ZAMAN SERİSİ ANALİZ YÖNTEMİ

Zaman serisini inceleme yöntemi oldukça standartlaşmıştır. Önce zamanda gözlenen tek boyutlu işaret, $x(t)$ için $(x(t), x(t+T), x(t+(k-1)T))$ gecikme vektörlerini oluştururuz^[5,6] Burada T gecikme zamanını, k ise gömme boyutunu göstermektedir. Önemli olan gecikme zamanı ve gömme boyutunu seçmektir. Gecikme zamanı çok küçük seçilirse sistemdeki periyodik ya da çok periyodikliği kaybedebiliriz. Bu da sisteme gereksiz bir determinizm kazandıracaktır. Buna karşı gecikme zamanı sistemin periyotlarından birine yakın seçilirse bu kez de söz konusu boyut sabit nokta gibi görünecek ve çekicinin boyutu yanlış örneklenecektir. Lineer bir analiz yapılacaksa öz ilinti fonksiyonunun ilk sıfır olduğu değer, lineer olmayan bir analiz yapılacaksa karşılıklı bilgi(entropi)'nin ilk minimumu gecikme zamanı olarak alınmalıdır.

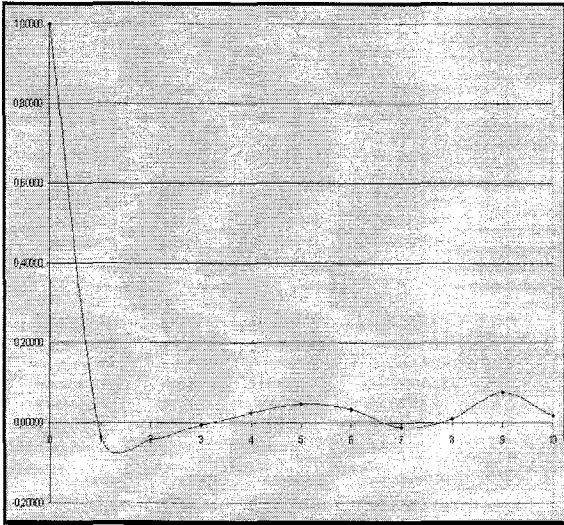
Gömme boyutu yanlış en yakın komşular yöntemi ile bulunur. Çekici olması gerekir daha küçük bir boyutlu uzaya gömülürse birbiri ile ilgili olmayan noktalar komşu görünebilir. Bu takdirde önce gömülmüş çekicide her noktanın en yakın komşusu bulunur. Gömme boyutu bir artırılarak eskiden yakın komşu olan noktalardan artık komşu olmaya bulunur. Bu şekilde komşu olmaktan çıkan noktaların toplam nokta sayısına oranı belirli eşiğin altına düşünce doğru gömmenin bulunduğu varsayılabilir.

3. ANALİZ DETAYLARI

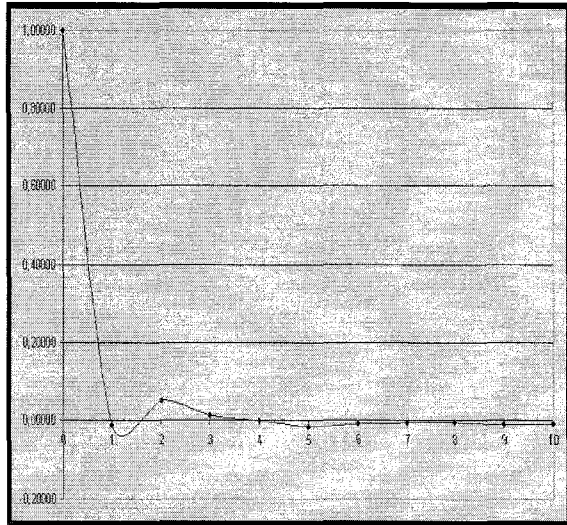
Bu çalışmada değişik metin uzunluğunun olası kaotik yapılar üzerindeki etkisi incelenmiştir. Bu amaçla 302, 604, 906 ve 1210 kelime uzunluklu, gündelik gazetelerden alınan Türkçe metinler kullanılmıştır. İki dil arasındaki olası benzerlik ve farklılıklar ortaya çıkabilmesi için aynı adımlar yaklaşık olarak aynı uzunluğuna sahip İngilizce metinler üzerine de uygulanmıştır. Önce her iki dil için öz ilinti fonksiyonu hesaplanmıştır.

ON KELİMEYE KADAR UZAKLIKLAR İÇİN “ÖZ İLİNTİ FONKSİYONU”

TÜRKÇE



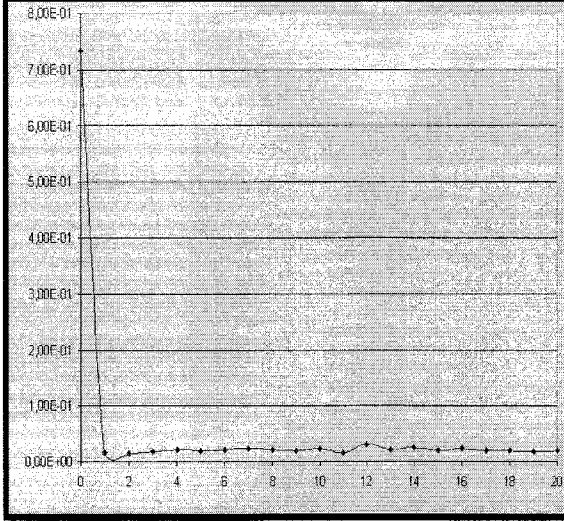
İNGİLİZCE



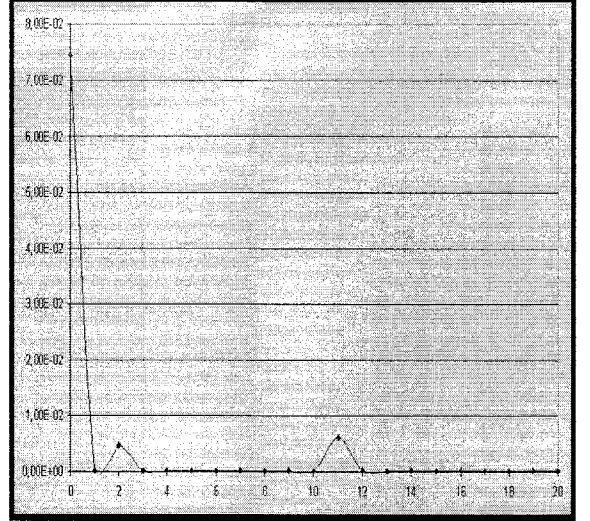
Görüldüğü gibi, her iki dilde de ardışık iki kelime (tümlenen ve tümleyen) dolayında bir sıfır ve minimum gözlenmiştir. İngilizce için 5, Türkçe için ise 7 civarında ikinci bir sıfır gözlenmiştir.

KARŞILIKLI BİLGİ FONKSİYONU

TÜRKÇE



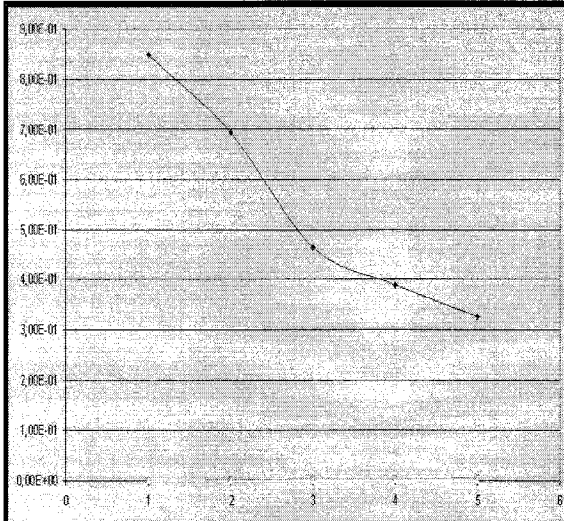
İNGİLİZCE



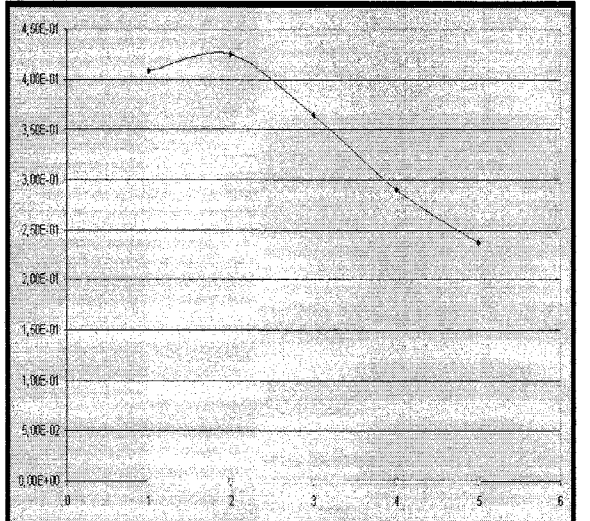
Burada karşılıklı bilgi fonksiyonundaki ilk sıfıra uyan şekilde iki kelime dolayında bir minimum gözlenmekte, ikinci bir minimum için açık bir işaret gözlenmemektedir. Gömme boyutunu saptamak için ise yanlış en yakın komşular yöntemi uygulanmış, İngilizcede 1 ila 2 boyut arasında görülen anomalinin bu dildeki isimlerin önüne gelen “a” veya “the” sözcüklerinden kaynaklandığı düşünülebilir.

YANLIŞ EN YAKIN KOMŞULAR FONKSİYONU

TÜRKÇE



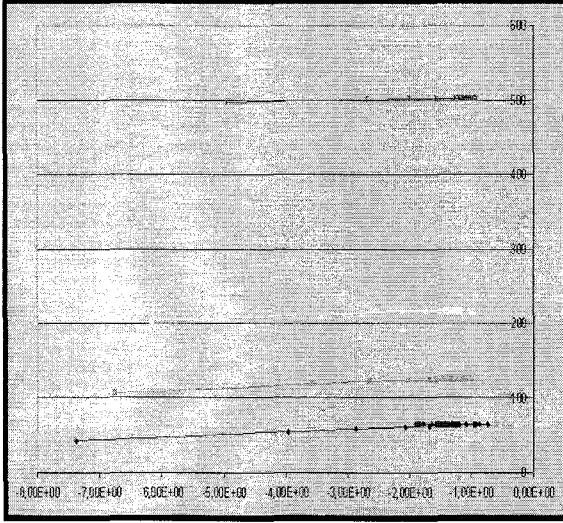
İNGİLİZCE



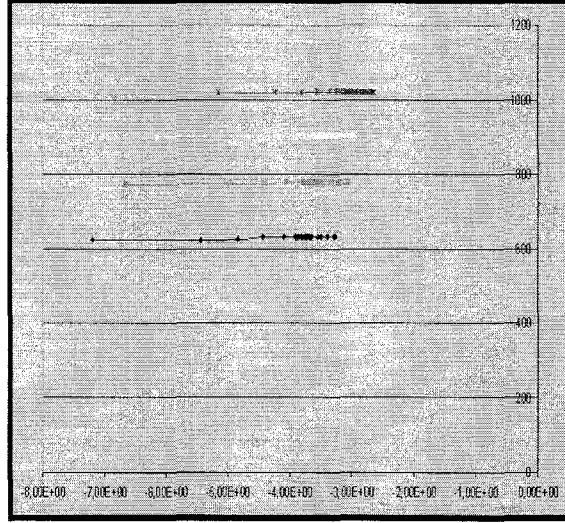
Pozitif bir Lyapunov üsteli kaotik davranış için en iyi bilinen işaret olduğundan gömme boyutu olarak 5 değeri kullanıldığında her iki dil için en büyük pozitif Lyapunov üsteline işaret eden pozitif bir eğim saptanmıştır.

LYAPUNOV ÜSTELİ

TÜRKÇE

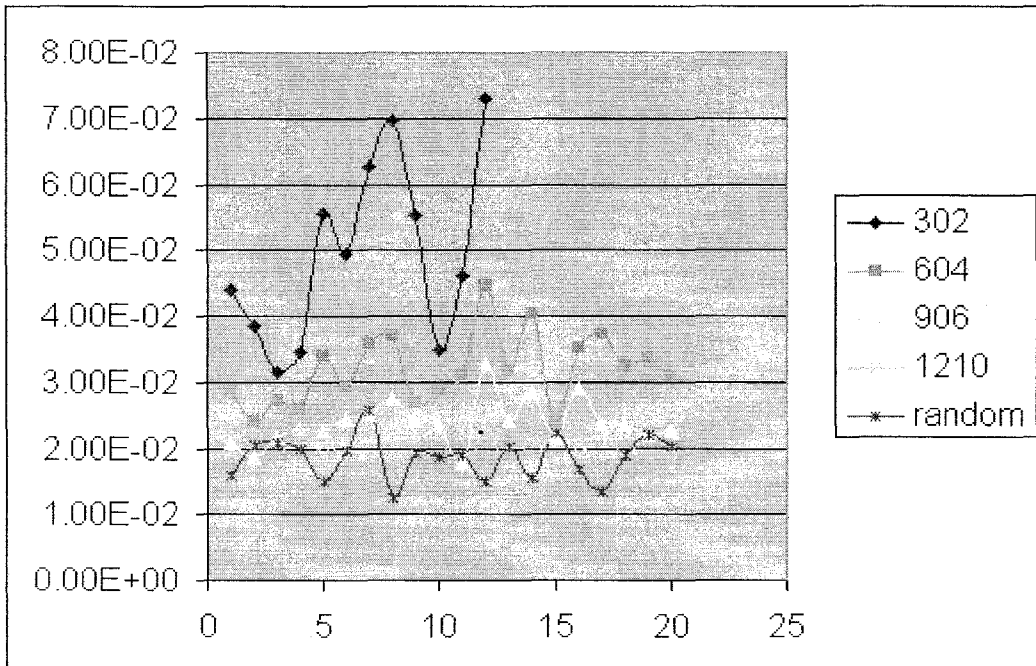


İNGİLİZCE



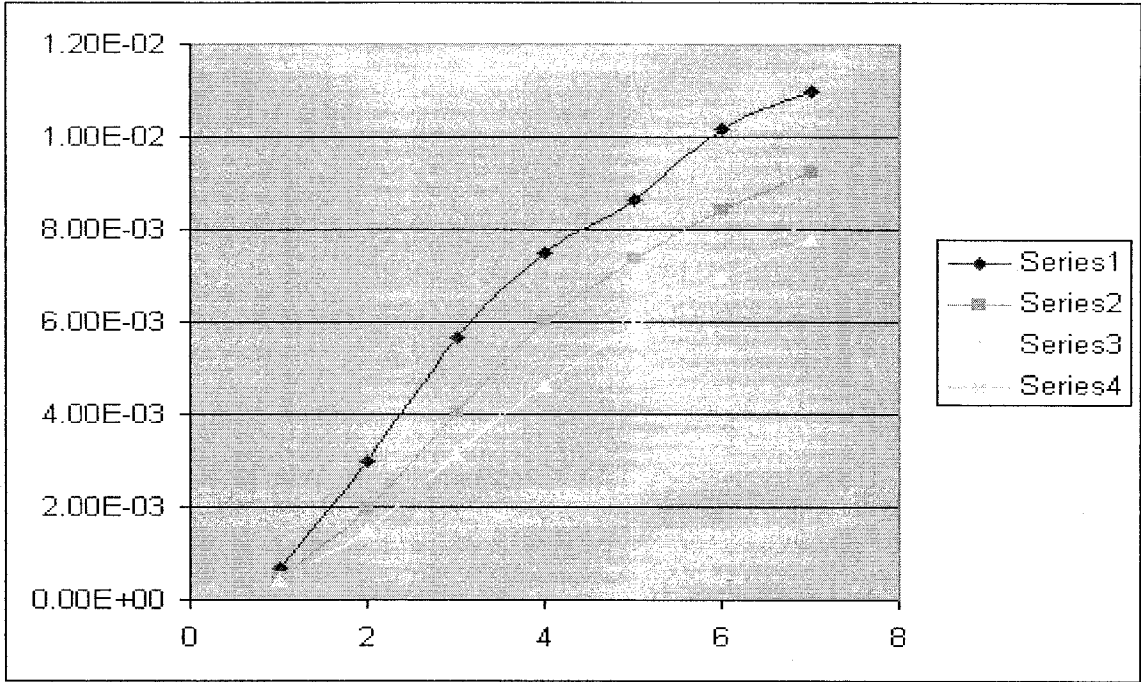
Bu arada metin uzunluğunun yanlış en yakın komşular fonksiyonu üzerindeki etkisi de dikkate değer bir davranışa sahiptir. Gömme boyutu 7 iken bir maksimum gözlemlendiği söylenebilir.

YANLIŞ EN YAKIN KOMŞULARIN METİN UZUNLUĞUNA BAĞIMLILIĞI



En yakın komşu sayısının boyutun fonksiyonu olarak artışı da lineerden azdır. Bu da bir çekicinin varlığına işaret eder.

EN YAKIN KOMŞU SAYISININ GÖMME BOYUTUNA BAĞIMLILIĞI



Pozitif bir Liapunov üsteli tutarlı olarak gözlemlenmekle birlikte metin uzunluğu ile belirli bir değişimin olduğu da çekici yapısının kullanılan parametrizasyon olan sıklığa karmaşık bir bağımlılığı olduğu, bu da parametrizasyona esin kaynağı olan Zipf kanununun ilk yaklaşım niteliğine işaret ettiği söylenebilir.

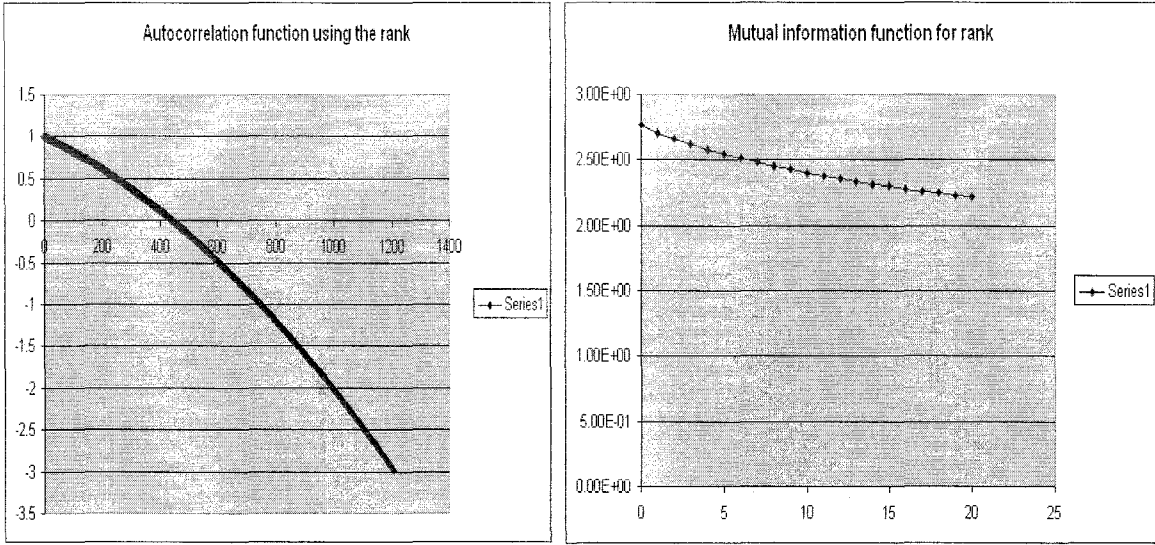
LYAPUNOV ÜSTELİNİN METİN UZUNLUĞUNA BAĞIMLILIĞI

Metin Uzunluğu	Gömme Boyutu=5 için Lyapunov Üsteli	Gömme Boyutu=7 için Lyapunov Üsteli
302	0.0126	0.0042
604	0.0146	0.0162
906	0.0144	0.0226
1210	0.0138	0.0283

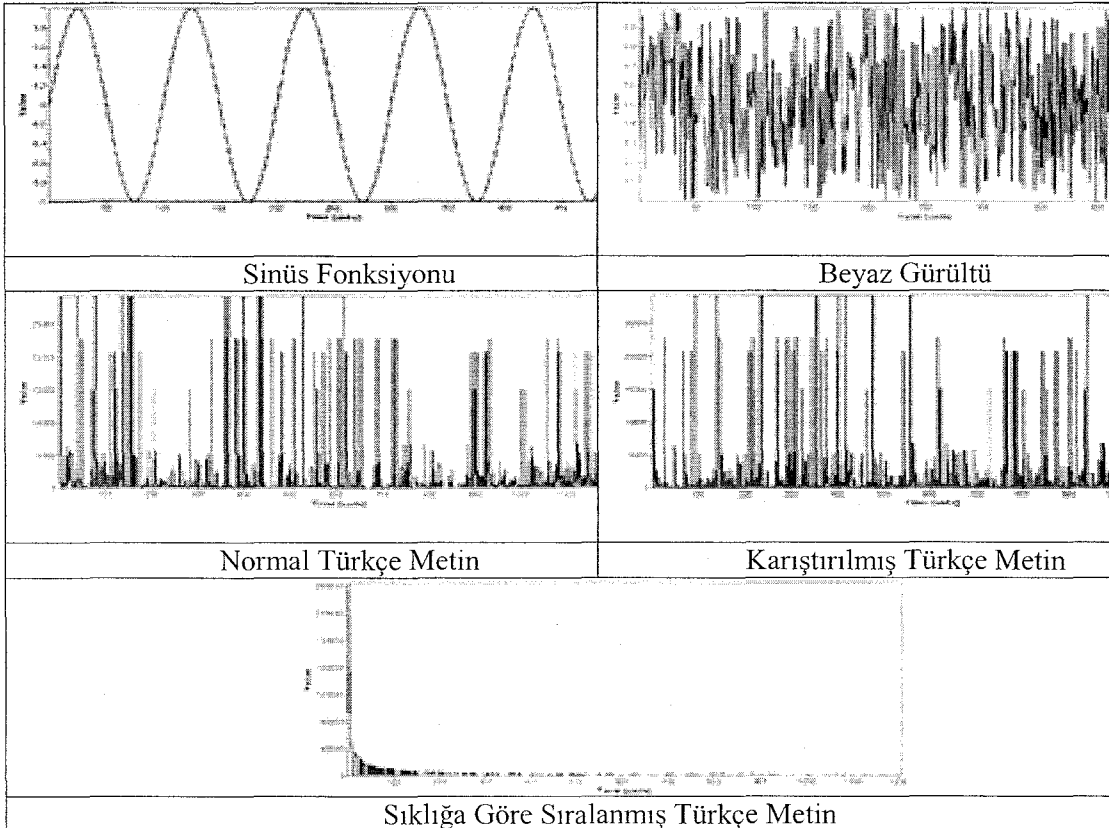
4. PARAMETRİZASYONLA İLGİLİ İNCELEMELER

Zipf kanununun bağımlı değişkeni sıklık, bağımsız değişkeni ise sıralamadaki derecedir. Ancak eğer sıklıklar yerine sıralamalar kullanılarak aynı metinden bir zaman serisi oluşturulursa, hem öz ilinti fonksiyonunun hem de karşılıklı bilgi fonksiyonunun yapılarını kaybederek bir çekiciye işaret etmedikleri gözlemlendi. Bu da sıklıkların gerçek bir bağımlı değişken olduğuna işaret eder. Grafiklerden her iki fonksiyonun derecelendirme işleminden sonra kaotik yapılarını kaybettikleri gözlenmektedir.

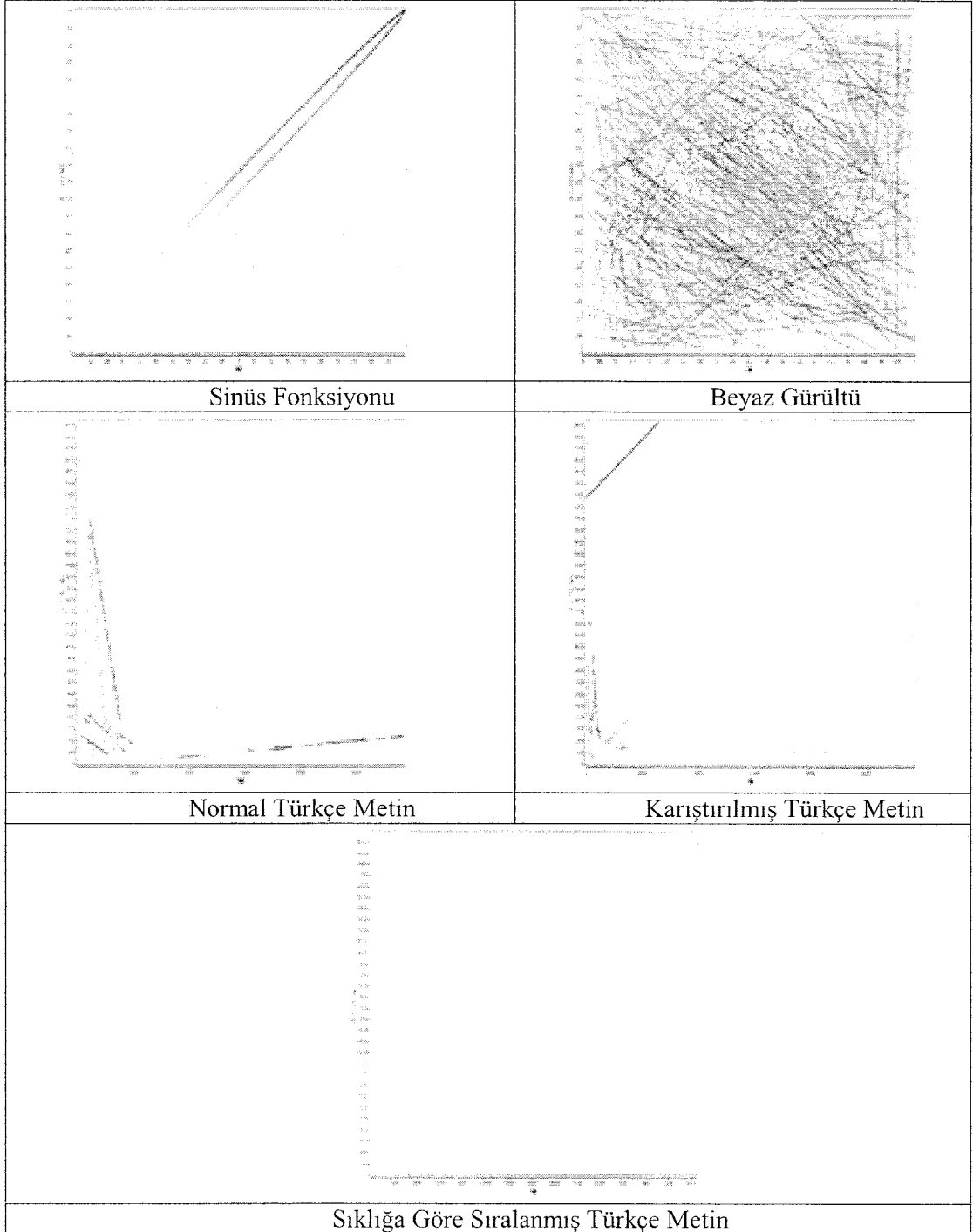
BAĞLI DEĞİŞKEN OLARAK FREKANS YERİNE DERECE



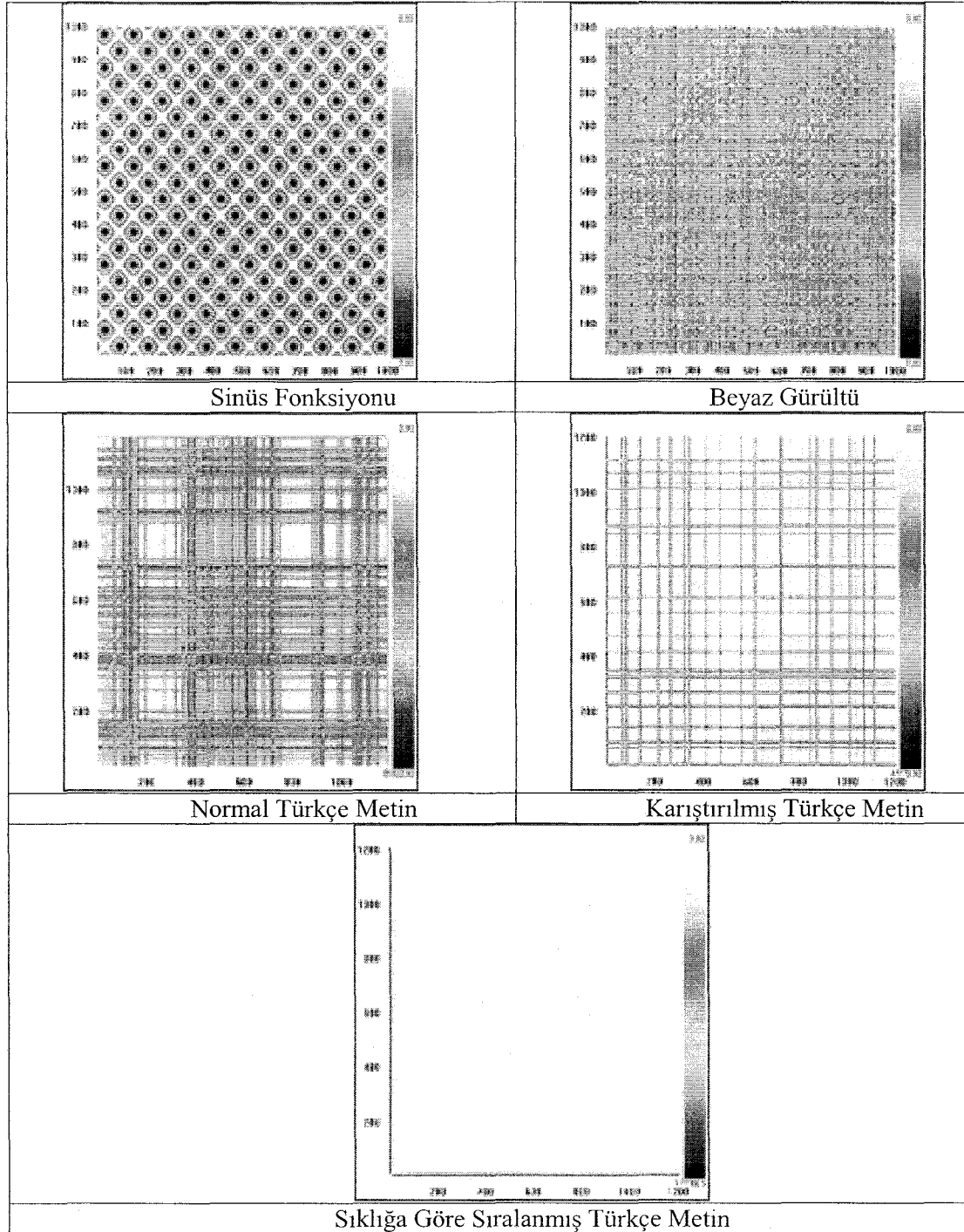
Bir diğer denemede Türkçe metindeki kelimeler rasgele karıştırılarak aynı analiz tekrarlandı. Karşılaştırmalı olarak kaotik yapısı olmayan sinüs fonksiyonu, beyaz gürültü, normal Türkçe metin, rasgele karıştırılmış Türkçe metin ve sıklığa göre sıralanmış Türkçe metin için zaman serileri aşağıda karşılaştırılmıştır.



Bu beş sistemin faz uzayı yapıları ise aşağıdadır. Normal Türkçe metnin oluşturduğu çekic yapısı açıkça görülebilmektedir.



Aynı sistemlerin tekrarlama grafikleri (recurrence plots) da aradaki farkı açıkça ortaya çıkarmaktadır.



5. SONUÇLAR

Doğal dillerin Zaman Serisi Analizinde kelime sıklığının ve kullanılan sıklık derlemesinin (corpus) önemi büyüktür. Zipf kanunu bir ilk yaklaşıklık olarak zaman serisinin oluşturulması olanağını sağlamış. Her iki dilde de kaotik yapılara rastlanmıştır. "Öz ilinti fonksiyonu" her iki dilde de bir kelimedenden sonra ilk sıfırına ulaşmaktadır. Bu nedenle yanyana gelen sözcükler arasında bir bağlantı mevcut, ancak bunun dışında kısa süreli bir bağlantı görülmemektedir. Türkçe için "Yanlış en yakın komşular" analizinden çıkan sonuca göre birbirlerine göre aralarında 5 ile 7 birimlik mesafe bulunan kelimelerin cümle yapısını içerisinde anlama katkısı büyüktür.

Her iki dilde de pozitif Liapunov üstelinin istikrarlı olarak gözlemlenmesi uzun süreli bir diğer bileşenin mevcut olduğu, bu bileşenin çekici yapısını oluşturduğu izlenimini verir.

Bu tarz çalışmaların kriptolojide kullanılan algoritmalar üzerinde yaratacağı etkiyi incelemek ilginç olabilir.

KAYNAKLAR

- [1] Hřebíček, Ludek, "Text as a Self-Similar Structure", *Text in Communication: Supra-Sentence Structures*, Universitätsverlag Dr. N. Brockmeyer, Bochum, 1992, pp. 91-96.
- [2] http://linkage.rockefeller.edu/wli/zipf/index_ru.html
- [3] Kučera, Henry; Francis, W. Nelson, *Computational Analysis of Present-Day American English*, Brown University Press, Providence, Rhode Island, 1967
- [4] Göz, İlyas, *Yazılı Türkçenin Kelime Sıklığı Sözlüğü*, Türk Dil Kurumu Yayınları, Ankara, 2003
- [5] Kantz, H. and T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge University Press, Cambridge, 1997.
- [6] Abarbanel, H.D.I., R. Brown, J. J. Sidorowich L.S. Tsimring, *The analysis of observed chaotic data*, Revs. Of Modern Phys., Vol. 65, No. 4, pp. 1331-1392, October 1993.