

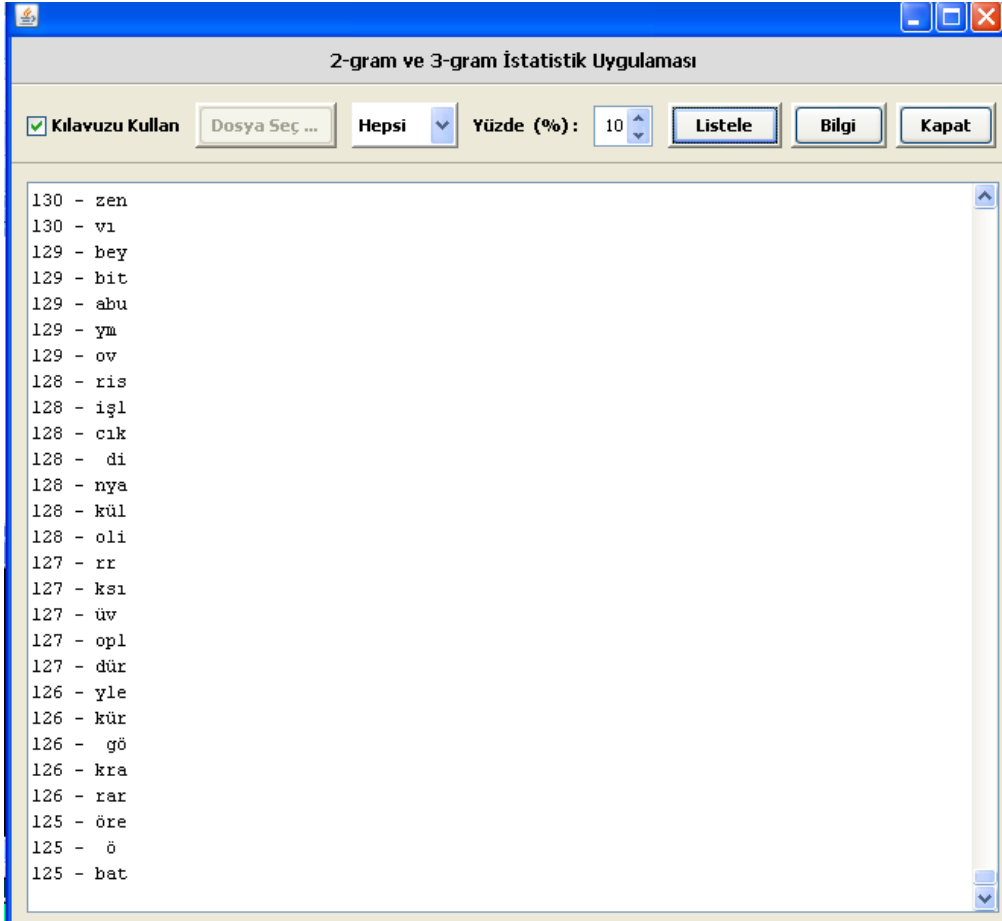


Kemik Doğal Dil İşleme Grubu

N-Gram ve Hece İstatistikleri

Yazılım temel olarak Türkçe Yazım Kılavuzundaki veya verilen herhangi bir metin dosyasındaki tüm N-Gram ve Hece istatistiklerini bulmak için tasarlanmıştır.

Yazılımın N-Gram İstatistikleri modülü ekranı Şekil 1'de görülmektedir.



Şekil 1 : N-Gram İstatistikleri Modülü

N-Gram İstatistikleri modülünde Türkçe' de var olan ikili ve / veya üçlü n-gram istatistikleri hesaplanmaktadır. "Kılavuzu Kullan" seçeneği aktif ise Türkçe Yazım Kılavuzundaki tüm kelimeler taranarak,

değilse "Dosya Seç" butonu ile kullanıcı tarafından seçilen herhangi bir Türkçe metin dosyası taranarak n-gram istatistikleri hesaplanır. Sadece ikili n-gramlar, sadece üçlü n-gramlar ve hem ikili hem de üçlü n-gramların bulunması seçeneğe bağlıdır. "Yüzde" seçeneği ile bulunan n-gramlar, girilen yüzde değerine göre en çok kullanılanlardan en az kullanılanlara doğru sıralanarak listelenir.

Listeleme işleminden sonra "Kaydet" butonu ile sonuçlar "sonuc/nGram" klasörü altındaki "result.txt" dosyasına kaydedilir.

Yazılımın Hece İstatistikleri modülü ekranı Şekil 2'de görülmektedir.

En Çok Kullanılan 100 Hece			
Genel	Başta	Ortada	Sonda
ma 5155 %2.6435	a 2804 %3.6978	la 3030 %4.1041	mak 3991 %6.1077
mak 3991 %2.0466	ka 2876 %2.7325	le 2184 %2.862	mek 3826 %5.9117
mek 3826 %1.962	et 1324 %1.935	lan 1439 %2.0892	ma 5155 %5.8411
me 3698 %1.8963	ya 1969 %1.5715	ra 1792 %1.9722	me 3698 %4.5067
la 3030 %1.5538	e 1018 %1.4567	laş 1193 %1.8664	lık 2076 %3.0719
ka 2876 %1.4748	sa 1487 %1.364	leş 1112 %1.7399	lik 1989 %2.9668
a 2804 %1.4379	i 1150 %1.3405	len 1090 %1.6262	li 2393 %2.2753
li 2393 %1.2271	ta 1943 %1.2993	ma 5155 %1.4049	lı 1859 %2.1498
le 2184 %1.1199	ha 1298 %1.161	ti 1216 %1.3259	sı 1445 %1.4411
lık 2076 %1.0646	ba 1466 %1.1154	li 2393 %1.3196	cı 1384 %1.3235
lik 1989 %1.0199	o 873 %1.1036	ta 1943 %1.2263	si 1440 %1.1478
ya 1969 %1.0097	te 1474 %1.0388	ka 2876 %1.2137	ci 1147 %1.0255
ta 1943 %0.9964	de 1341 %0.9932	ri 1281 %1.0778	ri 1281 %0.8813
lı 1859 %0.9533	bi 860 %0.8211	re 1221 %1.0256	luk 544 %0.8107
ra 1792 %0.9189	ko 738 %0.8005	na 1128 %0.9924	ğı 675 %0.8044
sa 1487 %0.7625	ma 5155 %0.7961	te 1474 %0.9861	siz 706 %0.748
te 1474 %0.7559	ku 787 %0.7607	da 1108 %0.9735	sız 681 %0.726

Toplam Hece Sayısı	195010	Toplam Baş Hece Sayısı	67960
Toplam Orta Hece Sayısı	63278	Toplam Son Hece Sayısı	63772
Toplam Farklı Hece Sayısı	3874		

Kaydet

Şekil 2 : Hece İstatistikleri Modülü

Hece İstatistikleri modülünde Türkçe'de var hece istatistikleri hesaplanmaktadır. "Kılavuzu Kullan" seçeneği aktif ise Türkçe Yazım Kılavuzundaki tüm kelimeler taranarak, değilse "Dosya Seç" butonu ile kullanıcı tarafından seçilen herhangi bir Türkçe metin dosyası taranarak hece istatistikleri hesaplanır. "Çalıştır" butonu ile Yazım Kılavuzu veya seçilen metin dosyasındaki en çok kullanılan ilk 100 hece kullanım sırasına göre büyükten küçüğe doğru listelenir. "Genel" başlığı altında tüm heceler içerisinde, "Başta" başlığı altında kelimelerin ilk heceleri içerisinde, "Ortada" başlığı altında kelimelerin ortalarında geçen heceler içerisinde ve "Sonda" başlığı altında ise kelimelerin son heceleri içerisinde en çok kullanılan ilk 100 hece listelenir. Listeleme işleminden sonra "Kaydet" butonu ile sonuçlar "sonuc/hece" klasörü altına kaydedilir.