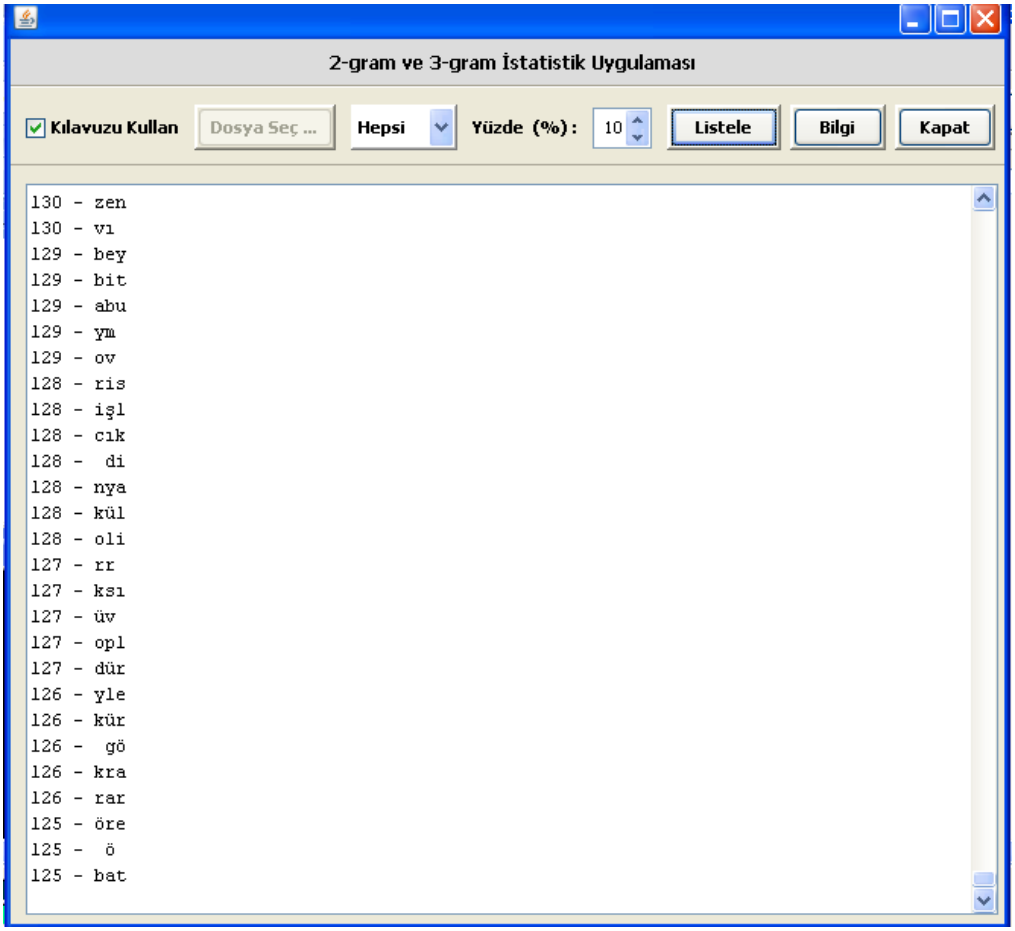
Kemik Doğal Dil İşleme Grubu

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.



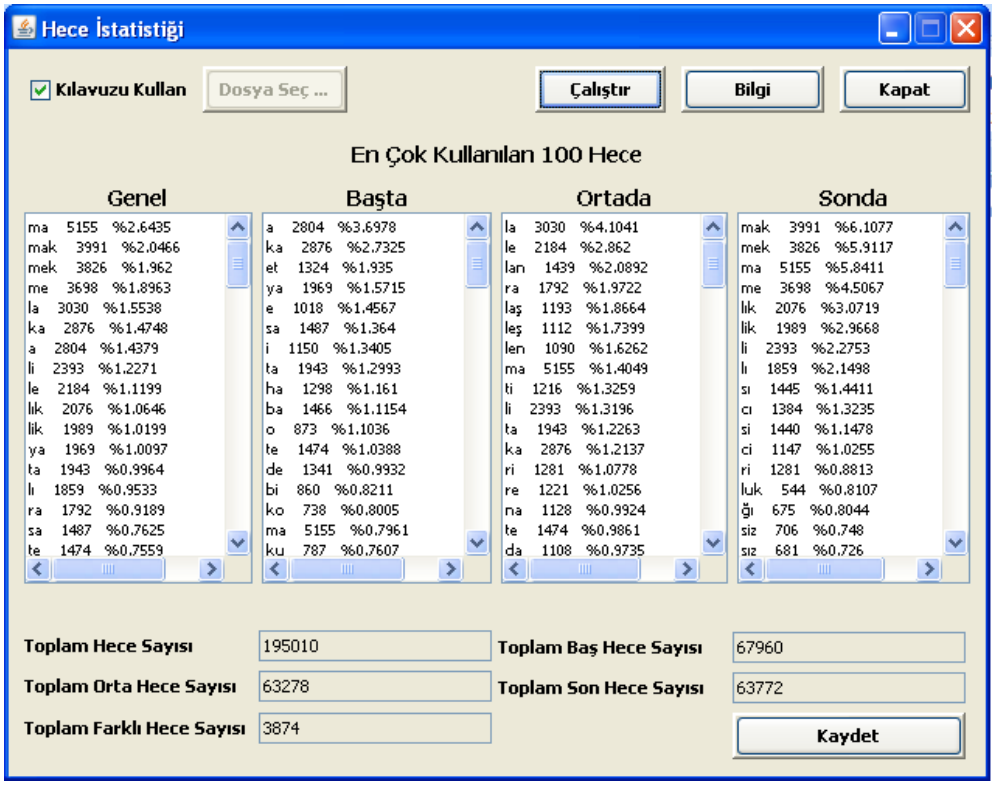
Şekil 1 : N-Gram İstatistikleri Modülü

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 kullanılandan en az kullanılana doğru sıralanarak listelenir.

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

Yazılımın Hece İstatistikleri modülü ekranı Şekil 2'de görülmektedir.



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" klasorü altına kaydedilir.