

Türkçe Geniş Dağarcıklı Konuşma Tanıma Sistemleri için Örüünün Yeniden Deęerlendirilmesi Tabanlı Dil Modellemesi Yaklaşımları

Language Modelling Approaches for Turkish Large Vocabulary Continuous Speech Recognition Based on Lattice Rescoring

Ebru Arısoy, Murat Saraçlar

Elektrik Elektronik Mühendisliği Bölümü
Boğaziçi Üniversitesi, 34342, Bebek, İstanbul, Türkiye
arisoyeb@boun.edu.tr, murat.saraclar@boun.edu.tr

Özetçe

Bu bildiriye Türkçe için geniş daęarcıklı bir konuşma tanıma (GDKT) sistemi üzerinde deęişik dil modelleri araştırılmıştır. Türkçe'nin sondan eklemeli bir dil olması dilimizin konuşma tanıma açısından işlenmesini zorlaştırmaktadır çünkü tüm kelimelerin konuşma tanıma sözlüğüne eklenmesi imkansızdır. Bu yüzden kelimelerin konuşma tanıma birimi olduęu bir modelin yanısıra, morf adı verilen veri güdümlü, kelime parçalarından oluşan bir model kullanılmıştır. Bu model daha önce Fince, Estonyaca ve Türkçe için uygulanmış ve kelime modeline göre daha iyi başarımlara ulaşılmıştır. Kullandığımız veri tabanında temel kelime modeli ile %38.8'lik ve temel morf modeli ile %33.9'luk hata oranlarına ulaşılmıştır. Ardından kelime modeli için kök, kök-sınıf tabanlı ve morf modeli için ilk morf tabanlı yeni dil modelleri yapılmış ve bu modeller konuşma tanıyıcıdan alınan örü dil modeliyle ölçeklenmiştir. Ölçeklenmiş kök-kelime modeli %0.5'lik mutlak bir başarıml artışı sağlamıştır. Diğer iki model ise temel model kestirimlerini daha gürbüz hale getiremedięi için temel modeli iyileştirememiştir.

Abstract

In this paper, we have tried some language modelling approaches for Large Vocabulary Continuous Speech Recognition (LVCSR) of Turkish. The agglutinative nature of Turkish makes Turkish a challenging language in terms of speech recognition since it is impossible to include all possible words in the recognition lexicon. Therefore, instead of using words as recognition units, we use a data-driven sub-word approach called morphs. This method was previously applied to Finnish, Estonian and Turkish and promising recognition results were achieved compared to words as recognition units. In our database, we obtained Word Error Rates (WER) of 38.8% for the baseline word-based model and 33.9% for the baseline morph-based model. In addition, we tried some new methods. Recognition lattice outputs of each model were rescored with the root-based and root-class-based models for the word-based case and first-morph-based model for the morph-based case. The word-root composition approach achieves a 0.5% increase in the recognition performance. However, other two approaches fail due to the non-robust estimates over the baseline models.

1. Giriş

Türkçe GDKT uygulamaları için zor bir dildir. Bu uygulamalar çok büyük bir kelime haznesine ihtiyaç duyar fakat Türkçe gibi sondan eklemeli bir dilde tüm kelimelerin sözlüğe eklenmesi imkansızdır. Sözlükte bulunmayan kelimeler konuşma tanıyıcı tarafından hatalı olarak tanınacaktır. Bu yüzden de hem sözlük dışı kelimelerden kaynaklanan hataları yok edecek hem de iyi başarımlar sağlayacak yeni konuşma tanıma birimlerine ihtiyaç vardır.

Türkçe için konuşma tanıma alanında yapılan ilk çalışmalardan biri 1999 yılında kelimelerin ayrık olarak tanınmasına yöneliktir [1]. Daha sonraki çalışmalar ise geniş daęarcıklı sistemler için kelimelere alternatif yeni birimler üzerinedir. Kaynakça [2]'de Fince GDKT'da [3] uygulanan en kısa betimleme uzunluęuna dayanan bir yöntem Türkçe için denenmiştir. Kaynakça [4]'de ise kelimelerin anlamlı en kısa birimleri olan morfemler konuşma tanıma deneylerinde kullanılmıştır. Kök ve kök sonralarını konuşma tanıma birimi olarak Türkçe'ye ilk uyarlayan ise Mengüşoęlu'dur [5]. Bu çalışmalara ek olarak tüm bu önerilen birimlerin getirilerinden yararlanmak için tüm modelleri beraber kullanan birleşik model önerilmiş [6], ayrıca konuşma tanıma çıktısı Türkçe'nin dilbilgisi kurallarından yararlanılarak Aęırlıklı Sonlu Durumlu Makinelerle (ASDM) düzeltilmiştir [7].

Bu bildiriye kelime ve kelime altı (morf) konuşma tanıma birimlerinin konuşma tanıma başarımları ve dil modellemesi açısından uygunlukları incelenmiştir. Kullanılan kelime altı modeller Kaynakça [2]'de Türkçe için uygulanan yöntemle elde edilmiştir. Fakat kullanılan veritabanları birbirinden farklıdır. Bunlara ek olarak, konuşma tanıyıcının örü çıktısı üzerinde kelimelerin köklerine dayanan yeni modeller denenmiştir.

Bildirinin içerięi şu şekildedir. 2. bölümde kullanılan dil modelleri ve veritabanının istatistiksel özellikleri anlatılmıştır. Konuşma tanıma deneyleri 3. bölümde verilmiştir. 4. bölümde ise örü çıktısı üzerinde yeni dil modelleri denenmekte, 5. bölüm ise bildirinin sonuçlarını içermektedir.

2. Kullanılan Dil Modelleri

Bu çalışmada dil modellemesi için iki farklı metin veritabanı kullanılmıştır. Bunlardan birincisi birçok farklı alandan

toplanmış olup 11,592,341 kelimedenden oluşur (Genel amaçlı veritabanı). İkinci veritabanı ise 15,060,045 kelimeyi kapsamakta ve sadece spor gazete haberlerinden oluşmaktadır.

2.1. Kelime Modeli

Kelime modelinde kelimeler konuşma tanıma sistemindeki sözlük elemanlarıdır. Konuşma tanıma kullanılmak üzere üçlü, dördü ve beşli dil modelleri SRILM [8] yazılımı kullanılarak oluşturulur. Daha iyi olasılık kestirimleri yapabilmek için Kneser-Ney yumuşatıcı model yöntemi uygulanmıştır. Kullanılan test kayıtları genel amaçlı metinlerden toplandığı için genel amaçlı veritabanı ve spor veritabanı için dil modelleri sınırlı bir sözlükle ayrı ayrı oluşturulmuş ve genel amaçlı modelin ağırlığı fazla olacak şekilde iki model ölçeklenerek birleştirilmiştir.

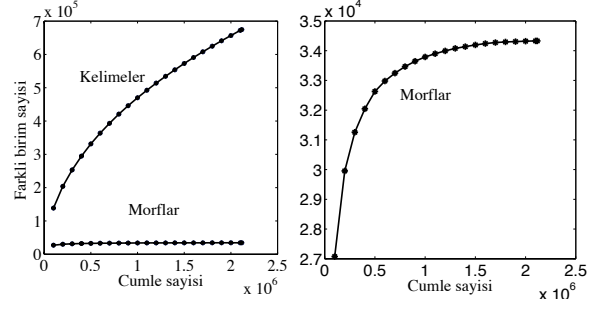
2.2. Kelime Altı (Morf) Modeli

Kelime altları morfem benzeri birimlerdir. Morfemler biçimbilim parçalayıcı kullanılarak elde edilir. Kelimelerin morflarına ayrılması ise tamamen veri güdümlüdür ve Morfessor [9],[10] adı verilen dilden bağımsız, öğreticisiz öğrenme algoritması kullanılarak çıkarılır. Bu algoritma en iyi kelime bölütlerini bulmak için en küçük betimleme uzunluğuna dayalı bir ceza fonksiyonu kullanır. Küçültülmeye çalışılan fonksiyon ise konuşma tanıma sözlük elemanlarının ve veritabanında bu sözlük elemanları ile betimlenen kelimelerin kodlama uzunluğudur. Kelime altları üzerinden bir dil modeli oluşturulabilmesi için önce metin veritabanında bir defadan fazla geçen kelimeler Morfessor algoritmasından geçirilerek morf adı verilen sözlük elemanları oluşturulur. Ardından tüm veritabanı Viterbi bölütlemesi kullanılarak kelime altlarına ayrılır ve n'li dil modelleri oluşturulur. Kelime modelinden farklı olarak bu modeller kelime sınırlarının bulunabilmesi için özel işaretlere gerek duyar. Akustik modeli olmamasına rağmen, dil modelleri bu özel işaretlerle eğitilmekte ve konuşma tanıyıcının çıktısı sonradan kelimelere dönüştürülmektedir. Böylece hata oranları kelimeler türünden hesaplanabilir.

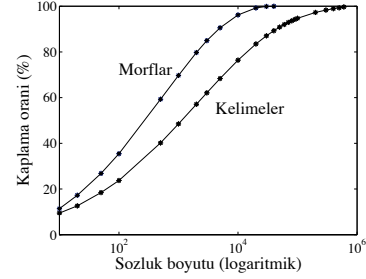
2.3. Metin Veritabanının İstatistiksel Özellikleri

Bu bölümde metin veritabanı, kelime ve morfların sözlük büyümesi ve kapsama yüzdeleri açısından karşılaştırılacaktır. Sözlük büyümesi veritabanındaki farklı birim sayısının yeni eklenen veriye göre nasıl arttığını gösterir. Kapsama yüzdesi ise sözlük boyutunun seçiminde belirleyicidir ve veritabanındaki kelimelerden kaçının doğru tanıma şansı olduğunu gösterir.

Şekil 1'de kelimeler ve morfların sözlük büyümesi açısından karşılaştırılması verilmiştir. Eklenen her cümle grubu ile farklı kelime sayısı artmaktadır. Sol taraftaki grafik morflarda farklı birim sayısının nasıl sabitlendiğini göstermektedir. Morflar veri güdümlü olduğu için bu beklenen bir sonuçtur ve aynı zamanda da istenen bir durumdur. Böylece, Şekil 2'de de görüldüğü üzere çok daha küçük bir sözlük boyutu ile %100'lük kapsama yüzdesine ulaşılabilir. Bu yüzden bu model için veritabanında geçen tüm morflar (34,328) ve kelimeler için en sık geçen 50,000 kelime sözlük elemanı olarak kullanılmıştır. Bu sözlük boyutları ile morflar için eğitim verisi üzerinden %100'lük, kelimeler için ise eğitim verisi üzerinden



Şekil 1: Kelimeler ve morflar için yeni eklenen cümlelere göre çizilmiş sözlük büyüme eğrileri



Şekil 2: Kelimeler ve morflar için logaritmik sözlük boyutuna göre değişen kapsama yüzdeleri

%90'lık kapsama yüzdesine ulaşılmıştır (Bkz. Şekil 2).

3. Konuşma Tanıma Deneyleri

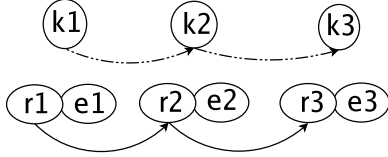
Bu çalışmada kelime ve morf modelleri için iki farklı konuşma tanıma deneyi gerçekleştirilmiştir. Her iki deneyde de kullanılan akustik modeller aynıdır. Konuşma tanıma için AT&T yazılımları [11] kullanılmıştır. Akustik modeller için önce HTK [12] yazılımıyla akustik öznelik vektörleri çıkarılmıştır. Akustik modeller 250'den fazla konuşmacıdan alınan yaklaşık 17 saatlik kayıtlarla eğitilir. Test için ise bir bayan konuşmacıya ait yaklaşık bir saatlik gazete haberlerinden okunan kayıtlar kullanılmıştır. Konuşma tanıma başarımları her iki model için Kelime Hata Oranları (KHO) ve Harf Hata Oranları (HHO) açısından değerlendirilir. Bu hata oranları örtü içerisindeki en olası yol kullanılarak hesaplanır. Sonuçlar Tablo 1'de verilmiştir. Sözlük dışı kelime sayısının etkisini ölçmek üzere kapsama yüzdesinin %95 olduğu 120,000 kelimelik bir sözlükle yapılan deneyde ise %36'lık KHO'na ulaşılmıştır.

4. Konuşma Tanıma Örü Çıktısına Uygulanan Yöntemler

Bu bölümdeki yaklaşım en iyi konuşma tanıma başarımlarına sahip modellerin örü çıktıkları üzerinden ikinci defa geçmek suretiyle temel dil modellerinde bazı değişiklikler yapmaya yöneliktir. Yeni yöntemler ile oluşturulan dil modelleri temel modellerle ölçeklenerek birleştirilecek ve konuşma tanıma başarımları tekrar değerlendirilecektir. Birleştirme işlemi için

Tablo 1: Kelimeler ve morflar kullanılarak yapılan geniş dağarcıklı konuşma tanıma deneyleri hata oranları (Morflar için altılı modeller daha kötü sonuç verdiği için tablo beşli modellere kadar verilmiştir)

	Kelimeler			Morflar		
	sözlük	KHO	HHO	sözlük	KHO	HHO
3'lü	50k	38.8	15.2	34k	39.2	14.8
4'lü	50k	38.9	15.1	34k	35.0	13.1
5'li	50k	39.0	15.1	34k	33.9	12.4



Şekil 3: Köklerden kök dil modelinin oluşturulması

aşağıdaki denklem kullanılır.

$$\log P_{yeni} = ((1 - \alpha) * \log P_{LM}^m + \alpha * \log P_{LM}^t) + \frac{1}{\lambda} * \log P_{AM} \quad (1)$$

Bu denklemde $\log P_{LM}^m$ kullanılan yeni yöntemin, $\log P_{LM}^t$ temel modelin dil modelleme olasılığını vermektedir. α ise bu iki modelin birleştirilmesinde kullanılan ölçek değeridir. $\frac{1}{\lambda} * \log P_{AM}$ ise temel modelin akustik olasılığının dil modelleme katsayısı ile çarpılmış halidir. Bu şekilde hesaplanan olasılıklarla yeni modelin örü çıktısı oluşturulur.

4.1. Kök Tabanlı Model

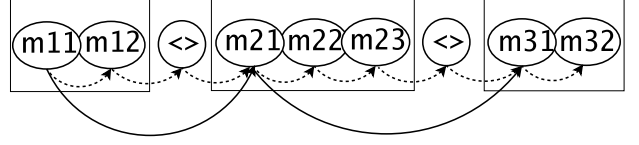
Türkçe gibi sondan eklemeli dillerde, tek bir kökten binlerce yeni kelime türetilmektedir. Bu yüzden kelimeleri kullanarak daha gürbüz n'li dil modeli kestirimleri yapabilmek için çok büyük veritabanlarına ihtiyaç vardır. Buradaki yöntemde temel düşünce kelimelerin köklerinin birbirini izleme düzenlilikleri ile kelime modeli için daha gürbüz kestirimlerin yapılabileceğidir. Türkçe için yapılan benzer bir çalışmada kökler, kelimelere göre eğitim setinden seçilen cümleler üzerinde daha iyi sonuçlar vermiştir [13]. Bizim çalışmamızda da kökler kelimelerin bir fonksiyonu olarak düşünülmüş ve kelimelerden köklerin çıkarılması için ASDM'ler kullanılmıştır. Köklerin dil modellemesinde kullanılması Şekil 3'de gösterilmiştir. Üçlü kelime dizilimlerinin olasılığı, üçlü kök dizilimlerinin olasılığı ile hesaplanmaktadır.

$$\text{Kök modeli} : P(r_3|r_2, r_1) \quad (2)$$

$$\text{Kelime modeli} : P(k_3|k_2, k_1) \quad (3)$$

Denklemde k ve r 'ler sırasıyla kelimeler ve köklere karşılık gelmektedir. Şekildeki e 'ler ise kök sonralarını göstermektedir.

Şekil 5'te ölçek değişkenine göre çizilmiş KHO eğrisi verilmiştir. α 'nın 0 olması konuşma tanımadaki dil modeli olarak sadece köklerin, 1 olması ise sadece kelimelerin kullanılması



Şekil 4: Morflardan ilk morf dil modelinin oluşturulması

anlamına gelmektedir. En iyi sonuç ise α 'nın 0.6 olduğu durum için elde edilmiş olup; kelime modeline göre mutlak %0.5'lik bir başarımların artışı sağlanmıştır.

4.2. İlk Morf Tabanlı Model

Kökler üzerinden oluşturulan modele benzer bir çalışma da morflar için yapılmıştır. Morf modelinde en iyi sonuçlar beşli dil modeliyle elde edildiği için temel model olarak bunun örü çıktısı kullanılmıştır. Örü çıktısından ilk morfları bulabilmek için ASDM'ler kullanılır. Şekil 4 bu model için dil modellemesinin nasıl yapıldığını göstermektedir. Kesik çizgi ile gösterilen oklar temel modelin izlediği yolu, tam çizgilerle gösterilen oklar ise ilk morfa dayalı modelin izlediği yolu vermektedir. Bu durumda m_{31} için dil modelleri olasılığı şu şekilde hesaplanır:

$$\text{İlk morf modeli} : P(m_{31}|m_{21}, m_{11}) \quad (4)$$

$$\text{Morf modeli} : P(m_{31} | \langle \rangle, m_{23}, m_{22}, m_{21}) \quad (5)$$

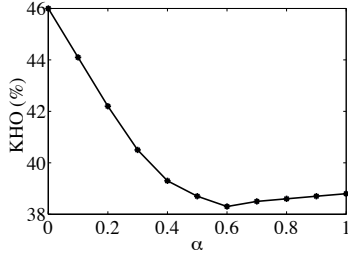
Bu iki model Denklem 1'teki hesaplamalar kullanılarak birleştirilir. İlk kelime altı modeli için ikili ve üçlü dil modellerine göre α ve KHO'nun değişim grafiği Şekil 6'te verilmiştir. Burada yeni model için daha düşük derecede modellerin kullanılması sebebi, olası ilk morf dizilimlerini sağlamak ve ilk kelime altı için daha doğru kestirimler yapabilmektir. Fakat ölçeklenmiş modellerle daha iyi başarımlara ulaşamamıştır. Bunun nedeni şu şekilde açıklanır. Kelime-kök modelinde akustik birimler kelimeler olup kök dil modelinin katkısı direkt olarak kelimelerdir. Bu yüzden kelime-kök ikilisi birlikte daha gürbüz kestirimler yaratır. Fakat ilk morf modelinde akustik birimler morflar üzerindedir ve oluşturulan yeni model sadece ilk morfu etkilemektedir. Bu da temel kestirimleri bozmaktadır.

4.3. Kök-Sınıf Tabanlı Model

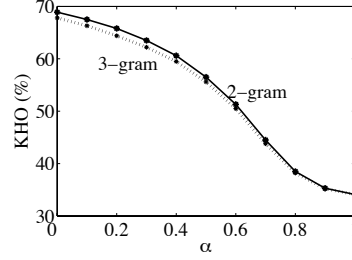
Bu yöntemdeki amaç dilbilgisel veya anlambilimsel benzerliği olan kelimeleri aynı sınıf altında toplamaktır. s 'lerin sınıfları temsil ettiği bir simgelemede kelime dizilimleri için olasılıklar şu şekilde hesaplanır.

$$P(k_1, k_2, k_3) = P(k_3|s_3) * P(s_3|s_2, s_1) \quad (6)$$

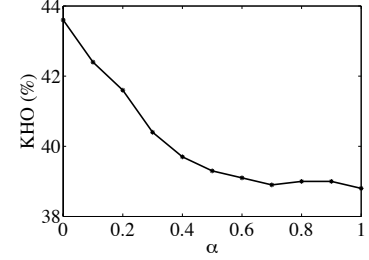
Bu modelde sınıflar kural tabanlı yada veri güdümlü olarak oluşturulur. Bu bildiride kullandığımız yöntem basit bir kural tabanlı yöntem olup; aynı köke sahip olan kelimeleri aynı sınıf olarak varsayılmaktadır. Bu yöntemin kelime modeli ile ölçek değişkenine göre birleştirilmiş hali Şekil 7'de verilmiştir. Bu modelde de ölçekleme işlemi başarımları olumsuz etkilemiştir.



Şekil 5: Kelime-kök modeli için ölçekleme değişkenine göre çizilmiş KHO eğrisi



Şekil 6: Morf-ilk morf modeli için ölçekleme değişkenine göre çizilmiş KHO eğrisi



Şekil 7: Kelime-kök-sınıf modeli için ölçekleme değişkenine göre çizilmiş KHO eğrisi

Denklem 6'dan da anlaşılacağı gibi bu yöntemle kök modeli arasındaki tek fark olasılık kestirimlerinde sınıfa ait olma olasılığının ($P(k_3|s_3)$) da kullanılmasıdır. Bu olasılığın elimizdeki veritabanından tam olarak kestirilememesi sınıf-kök modelini bozmaktadır.

5. Sonuçlar

Bu bildiriye Türkçe için geniş dağarcıklı bir konuşma tanıma sistemi tasarlanmıştır. Yapılan konuşma tanıma deneyleri sonucunda, kelime modelinde 50k kelimelik bir sözlük ve üçlü dil modeli ile %38.8'lik ve morph modelinde de 34k birimlik bir sözlük ve beşli dil modelleri ile %33.9'luk hata oranlarına ulaşılmıştır. Ardından yeni dil modelleri, kelimeler için kök ve kök-sınıf, morflar için ilk morf, yaratılmış ve temel dil modeli ile yeni dil modeli ölçeklenerek konuşma tanıyıcının akustik örü çıktısı bu modellerle tekrar değerlendirilmiştir. Kök modeli kelime modelindeki KHO'nı mutlak olarak %0.5 azaltmış, diğer iki model ise temel dil modellerine herhangi bir katkı sağlamamıştır. Bu durum ilk morf modelinin tüm akustik model birimlerini kapsamamasıyla ve kök-sınıf modelindeki $P(w_i|s_i)$ olasılıklarının tam olarak kestirilememesiyle açıklanabilir.

6. Teşekkür

Yazarlar Morfessor Programındaki yardımları için Helsinki University of Technology Konuşma Tanıma grubuna, sağladıkları akustik veritabanı için Sabancı Üniversitesine, spor haberleri metin veritabanı için ODTÜ'ye, ASDM yazılımları için AT&T - Labs Research'e teşekkür ederler. Bu çalışma TÜBİTAK-BDP ve SIMILAR NoE tarafından desteklenmektedir.

7. Kaynakça

- [1] Arslan, L. M., "Türkçe sürekli konuşma tanıma sisteminin sayı tanıma uygulaması", Proc. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU 1999), 1999, pp. 64-67.
- [2] Hacıoğlu, K., Pellom, B., Ciloglu, T., Oztürk, O., Kurimo, M., Creutz, M., "Word Splitting for Turkish", Proc. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU 2003), 2003, pp. 293-296.
- [3] Siivola, V., Hirsimäki, T., Creutz, M., Kurimo, M., "Unlimited vocabulary speech recognition based on

morphs discovered in an unsupervised manner", Proc. Eurospeech'03. Geneva, Switzerland, 2003, pp. 2293-2296.

- [4] Carkı, K., Geutner, P., Schultz, T., "Turkish LVCSR: Towards better speech recognition for agglutinative languages", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2000. Vol. 3. Istanbul, Turkey, pp. 1563-1566.
- [5] Mengusoglu, E., Deroo, O., 2001. Turkish LVCSR: Database preparation and language modeling for an agglutinative language. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2001, Student Forum. Salt-Lake City.
- [6] Arısoy, E. and Arslan L.M., "Türkçe gazete haberleri dikte sistemi", Proc. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU 2005), 2005.
- [7] Erdogan, H., Buyuk, O., Oflazer, K., "Incorporating language constraints in sub-word based speech recognition", IEEE Automatic Speech Recognition and Understanding Workshop, 2005, Cancun, Mexico.
- [8] Stolcke, A., "SRILM - An extensible language modeling toolkit", Proceedings of the International Conference on Spoken Language Processing, 2002, 901-904.
- [9] Creutz, M. and Lagus, K., "Unsupervised discovery of morphemes", Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02, 2002, 21-30.
- [10] Creutz, M. and Lagus, K., "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor", Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology, 2005, URL: <http://www.cis.hut.fi/projects/morpho/>.
- [11] Mohri M. and Riley, M. D., DCD Library - Speech Recognition Decoder Library. AT&T Labs - Research. <http://www.research.att.com/sw/tools/dcd/>.
- [12] Young, S., Ollason, D., Valtchev, V., Woodland, P., "The HTK book (for HTK version 3.2.)", March 2002.
- [13] Çiloğlu M., Çömez M. and Şahin S., "Takılı bir dil olarak Türkçe için dil modelleme", Proc. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU 2004), 2004.