# Sözcük Köklerinin Sözdizimsel Olarak Kümelenmesi

# Clustering Word Roots Syntactically

Mustafa Burak Öztürk, Burcu Can
Bilgisayar Mühendisliği Bölümü,
Hacettepe Üniversitesi,
Beytepe, Ankara, Türkiye
burakozturk.cs@gmail.com,burcucan@cs.hacettepe.edu.tr

***Özetçe*—Sözcüklerin dağılımsal gösterimleri hem sözdizimsel hem de anlambilimsel doğal dil işleme problemlerinde kullanılmaktadır. Bu bildiride sözcük köklerinin kümelenmesi için iki farklı yöntem kullanılmıştır. İlk yöntemde dağılımsal bir sözcük modeli olan word2vec [1] modeli daha önceki çalışmalardan farklı olarak Türkçedeki sözcük köklerinin kümelenmesi için kullanılmıştır. Bu amaçla, sözcük köklerinin dağılımsal olarak benzerlikleri word2vec kullanılarak modellenmiş ve kökler sözdizimsel (isim, fiil vs.) olarak birbirine benzer kategorilere ayrıştırılmıştır. Diğer yöntemde ise sözcük köklerinin kümelenmesi için bilgi teorisi ve olasılık tabanlı iki ayrı model geliştirilmiştir. Karşılıklı bilgi (mutual information) ile geliştirilen bir metrik [8] ve Jensen-Shannon ıraksama (divergence) metriği ile sözcük köklerinin benzerlikleri hesaplanmış ve bu şekilde kümeleme işlemi yapılmıştır. Sözcük köklerinin sözdizimsel olarak kümelenmesi, makine tercümesi, soru cevaplama gibi dil üretme içeren diğer doğal dil işleme uygulamalarında, özellikle sondan eklemeli diller için önemli bir yere sahiptir. Elde edilen sözcük köklerine ait kümelerin saflık değeri 0.92'ye kadar yükselmiştir.**

***Anahtar Kelimeler — Türkçe sözdizimi; kümeleme; morfoloji***

***Abstract*—Distributional representation of words is used for both syntactic and semantic tasks. In this paper two different methods are presented for clustering word roots. In the first method, the distributional model word2vec [1] is used for clustering word roots, whereas distributional approaches are generally used for words. For this purpose, the distributional similarities of roots are modeled and the roots are divided into syntactic categories (noun, verb etc.). In the other method, two different models are proposed: an information theoretical model and a probabilistic model. With a metric [8] based on mutual information and with another metric based on Jensen-Shannon divergence, similarities of word roots are calculated and clustering is performed using these metrics. Clustering word roots has a significant role in other natural language processing applications such as machine translation and question answering, and in other applications that include language generation. We obtained a purity of 0.92 from the obtained clusters.**

***Keywords — Turkish syntax; clustering; morphology***

## I. GİRİŞ

Türkçede, bütün diğer dillerde olduğu gibi isim ve eylem gibi sözdizimsel sözcük türleri bulunmaktadır [7]. Sözcüklerin otomatik olarak sözdizimsel kategorilere ayrıştırılması başta makine tercümesi olmak üzere birçok doğal dil işleme uygulamasında önemli bir yere sahiptir [10, 11, 13, 15, 18]. Türkçe gibi sondan eklemeli dillerde ise sözcük türlerinin yanında sözcük köklerinin türleri de dilin türetmeye açık olmasından ötürü önemlidir.

Sözcük sonuna gelen eklerin sırası biçim dizgesini ilgilendirir ve dilin oluşturulması sırasında bu bilgiye ihtiyaç duyulur. Türkçedeki biçim dizgesi kuralları Oflazer'de [4] farklı sözcük kökü türleri (isim, eylem, sıfat vb) için ayrı ayrı ifade edilmiştir. Bir sözcük kökünün hangi türde olduğu bilgisi doğrudan o kökün alabileceği ekleri ve o köklerden türetilebilecek sözcükleri de belirleyeceği için soru cevaplama (question answering), makine tercümesi (machine translation) gibi dilin üretme (language generation) içeren tüm doğal dil işleme uygulamalarında büyük önem taşımaktadır. Özellikle sondan eklemeli dillerde sözcük kökünün bilinmesi dilin üretilmesi aşamasında ihtiyaç duyulan ilk bilgilerden biridir.

Bildiğimiz kadarıyla sözcük türlerinin otomatik olarak kümelenmesi üzerine daha önce bir çalışma yapılmamıştır. Ancak sözcük formlarının kümelenmesi farklı doğal dil işleme uygulamalarında ele alınmıştır. Örneğin istatistiksel makine çevirisinde kullanılan sınıf tabanlı dil modellerinin oluşturulmasında sözcük sınıfları bulunmaya çalışılmaktadır. Sınıf tabanlı n-gram modellerinde [10], sözcüklerin hepsinin ayrı ayrı ele alınması yerine benzer özellikleriyle birleştirilmiş kümeler halinde ele alınması veri seyrekliği problemini en aza indirgeyerek model için gerekli parametre sayısını azaltmaktadır [11]. Sözcüklerin kümelenmesi konuşma tanıma sistemlerinin performanslarının artırılmasında da kullanılmıştır [12, 13]. Kneser ve Ney [14] ve Martin [12] sözcük kümelemek için değişim (exchange) kümeleme algoritmalarını kullanarak maksimum olabilirlik (maximum likelihood)

yöntemi ile sözcük türlerini otomatik olarak bulmayı amaçlamışlardır.

| |
|---|
| Cal IS ma sI |
| CalIS ma sI |
| nitelik i nde ki |
| nitelik in de ki |
| iki nci |
| aSama da |
| fark lI |
| kUltUr ler e |

**Tablo 1**. Eklerine ayrılmış ve Türkçeye özgü karakterleri büyük harf allofanlarıyla (eşses) değiştirilmiş sözcükler

Jakob ve Brants [15] değişim algoritmasını geliştirmeye çalışmış ve algoritmanın dağılımsal versiyonunu geliştirerek sözcüklerin kümelenmesi işlemini daha hızlı ve doğru yapmaya çalışmışlardır. Hogenhout ve Matsumoto [16] ise sözcüklerin kümelenmesi için bir *treebank* (Wall Street Journal Treebank) kullanarak sözcüklerin sözdizimsel davranış benzerliklerinden faydalanmışlardır.

Bu çalışmada sözcük köklerini iki farklı yöntem ile kümelemeye çalıştık. Her iki çalışma kapsamında da sadece sözcüklerin Türkçe için açık kaynak, doğal dil işleme kütüphanesi olan Zemberek [2] ile eklerine ayrıştırılmış halleri kullanarak ve bunun dışında herhangi bir etiketlenmiş veri kullanmadan denetimsiz (unsupervised) olarak kök kategorilerini bulmaya çalıştık. Önerilen çalışmalardan ilki bir vektör uzay modeli olan word2vec modeli [5] ile gerçekleştirilmiş olup, ikinci yöntemde bilgi teorisi ve olasılık tabanlı bir benzerlik modeli geliştirilmiştir.

Bildiri şu bölümlerden oluşmaktadır: II. bölümde yapılan çalışmanın teorik ayrıntılarından, III. bölümde gerçekleştirilen deney ve sonuçlardan, IV. bölümde ise olası gelecek çalışmalardan bahsedilmektedir.

## II. ÇALIŞMA

### A. Veri

Çalışmada kullanılan veri 1 milyon cümle içeren Morpho Challenge 2009 [3] derleminden elde edilmiştir. Buradaki sözcükler Zemberek yardımı ile eklerine ayrıştırılmıştır. Zemberek bir sözcük için birden fazla çözümleme önerebilmektedir. Bu durumda çözümlemelerin hepsi eğitim veri kümesine dahil edilmiştir. Bu şekilde toplamda 16.746.961 eklerine ayrıştırılmış sözcük elde edilmiştir.

Çalışmada kullanılan word2vec modelinin [5] Türkçeye özgü karakterler (ü, ö, ı, ç, ş, ğ) ile uyumlu çalışmadığı gözlemlenmiştir. Türkçe dilinin fonetik (sesbilgisel) özelliklerine uygun olarak bu karakterler yerine bunların allofan (eşses) karşılıkları büyük harf ile yazılmıştır [17]. Örneğin *öğrenci* sözcüğü *OGrenci*, *çalışması* sözcüğü *CalISmasI* şeklinde kullanılmıştır (bkz. Tablo 1).

| | gerçek leş tir diğ i nin | | | | |
|---|---|---|---|---|---|
| CBOW (2-grams) | gerçek leş | leş tir | tir diğ | diğ i | … |
| skip-gram (1-skip-2grams) | gerçek tir | leş diğ | tir i | diğ nin | |

**Tablo 2**. Eklerine ayrılmış olan *gerçekleştirdiğinin* sözcüğü için, pencere boyutunun 2 olarak seçilmesi ile oluşan pencereler.

| |
|---|
| rüşvet, esaret, hasret, plebisit |
| seyyare, enfiye, turnike, muahede, beriki, fonem |
| ye, ip, ti, ki, in, m, i, dik, sin, si, de, ne |
| ağ, aydın, sağ, tutuk, kira, av, anlatı, boya, ıska, pompa |
| utan, yumuşa, damla, bağda, boşal, daral, fırla, hatırla, tanı |
| ısıt, taşın, yararlan, çoğal, azal, yansı, payla, ula, hızlan |
| diploma, parola, sigara, tanrı, yumurta, tartışma, acı, anı |
| gelin, evren, vali, evli, gebe, üste, rehber, yüksek, amir |

**Tablo 3.** word2Vec modeli ile oluşturulan kök kategorilerinden bazıları

### B. Word2Vec Modeli ile Sözcük Köklerinin Kümelenmesi

word2vec modeli birkaç modeli içermektedir. Devamlı sözcük torbası modeli (Continuous bag-of-words/CBOW) ve skip-gram modelleri [6] kullanılarak sözcükler vektörlerle ifade edilebilmektedir. Sözcük torbası modeli bir metindeki tüm sözcüklerin, sıradan bağımsız bir şekilde, sadece frekans bilgisiyle ele alınması yöntemidir. Skip gram modeli ise sözcüğün bağlamının aynı sözcüğün kendisinden önce ve sonra gelen sözcüklerle tahmin edilmesinde kullanılır. Bu çalışma kapsamında her iki model de kullanılmıştır.

Vektör olarak ifade edilen sözcükler k-means kümeleme algoritması ile kümelerine ayrıştırılmıştır.

Dağılımsal modellerdeki önemli parametrelerden biri kullanılacak pencere boyutudur. Pencere boyutu, özellik vektörüne dahil edilecek olan sözcüklerin özellik vektörü çıkarılan sözcüğün ne kadar ileri ve gerisine bakarak elde edileceğini belirleyen bir parametredir (Tablo 2). Bu çalışma kapsamında her pencerede bir kökün aldığı ekler yer almaktadır.

Veri kümesinde sözcüklerin aldığı eklerin ortalama sayısı 1,35 olarak hesaplanmıştır. Yani bir sözcüğün türünü kendisinden sonra gelen ekler ile tespit edebilmemiz için pencere boyutunun en az bu sayı kadar olması gerekir. Bu çalışma sırasında pencere boyutu olarak 4 ve 5 pencere boyutları test edilmiştir. Her kök için bu pencere boyutlarında yer alan ekler alınarak köklerin özellik vektörleri oluşturulmuştur. Türkçede önek (prefix) kullanımı fazla olmadığı için pencerelerde daha çok soneklerin (suffix) bulunduğu sözcüklerin sağ tarafı dikkate alınmıştır.

Kümeleme algoritmalarındaki önemli parametrelerden bir diğeri de oluşacak kategorilerin sayısıdır. Bu sayı k-means kümeleme algoritmasındaki k değerine denk gelmektedir. Elde edilecek kök türlerinin sayısı isim, sıfat ve fiil gibi sadece temel kategoriler düşünüldüğünde oldukça azdır. Ancak modelde ek ve kökler arasında bir ayrım

| |
|---|
| dört, üç, ün, yüküm, gönül, beş |
| yanıtla, cevapla, haşla, kucakla, üfle, tıkla |
| alkol, tırnak, davet, eşek, antijen, adalet |
| durgun, yolcu, tutuk, yolsuz, olumsuz, geniş, temiz |
| eğlen, epey, gizli, düşün, böyle, sade, dere, ön |
| ağı, sigara, genelge, itibari, obje |
| yarat, tüket, uyuştur, bat, ger, bas, koş, başar, say, yararlan |
| öğle, klavye, gücün, yem, noter, nevi, kilise, bütçe, eski, priz |

**Tablo 4.** Bilgi teorisi ve ıraksama modeli ile oluşturulan kök kategorilerinden bir kesit

yapılmadan hepsi bir arada kümelendirilmiştir. Bu yüzden k=50 olarak belirlenmiştir. word2vec ile elde edilen kök kategorilerinden bazıları Tablo 3'te verilmiştir.

### *C. Bilgi Teorisi Modeli ve Olasılıksal Model*

Sözcük köklerinin kategorilerinin bulunmasında kökler arasındaki benzerliği hesaplayabilmek için Baek [8] tarafından önerilen bir metrik kullanılmıştır. Bu metrik için bir veri kümesinde bulunan iki morfemin benzerliğinin hesaplanmasında, bu morfemlerin diğer morfemlerle olan karşılıklı bilgisine (mutual information - MI) ihtiyaç duyulur. Başka bir deyişle bu metrik, iki MI değeri arasındaki benzerliği ölçmektedir. Bu çalışmada ise belirtilen metrik iki sözcük kökü arasındaki benzerliğin ölçülmesi için uyarlanmıştır.

$s_1$ ve $s_2$ sözcük kökleri arasındaki benzerliği hesapladığımız metriğin formülü aşağıda verilmiştir:

$$Sim(s_1; s_2) = \frac{\sum_{m_i \in M_1 \cup M_2} \min(MI(s_1,m_i),MI(s_2,m_i))}{\sum_{m_i \in M_1 \cup M_2} \max(MI(s_1,m_i),MI(s_2,m_i))} \quad (1)$$

Burada $M_1$, $s_1$ kökünün aldığı veri kümesi içerisindeki ek kümesini göstermekteyken, $M_2$ de $s_2$ kökünün, aynı veri kümesi içerisinde aldığı ek kümesini göstermektedir.

Kümeleme işlemi için hiyerarşik yığınsal (agglomerative) kümeleme algoritması kullanılmıştır. Bu algoritmaya göre başlangıçta her sözcük kökü kendi kategorisi altında yer almaktayken, yukarıda verilen metriğe göre en çok benzeyen kategoriler her iterasyonda birleştirilmektedir. Her bir iterasyonda sadece iki kategori tek bir kategori altında birleştirilmiştir. Her birleşmeden sonra yeni kategorinin diğer kategorilere olan benzerliği tekrar hesaplanarak, bu işlem hesaplanan benzerlik oranı 0 olana kadar devam ettirilmiştir. Sonunda 152 kategori elde edilmiştir.

Oluşan kategoriler yeterince büyük olmadıkları için oluşan kök kümeleri simetrik Kullback-Liebler (KL) ıraksama (divergence) [9] yöntemi ile bir kez daha birleştirilmiştir. Jensen-Shannon ıraksama olarak adlandırılan simetrik KL ıraksama, KL ıraksamasının her iki yön için aritmetik ortalamasını verir. $S_1$ ve $S_2$ kök kategorileri için KL ıraksaması aşağıda verildiği gibi hesaplanmaktadır:

$$D_{KL}(S_1||S_2) = \sum_{m_i \in M} p_{S_1}(m_i) \frac{p_{S_1(m_i)}}{p_{S_2(m_i)}} \quad (2)$$

| Pencere | Küme sayısı | Saflık |
|---|---|---|
| 4 | 47 | 0.91 |
| 5 | 46 | 0.92 |

**Tablo 5.** CBOW kümeleme sonuçları

Burada $M = M_1 \cup M_2$ şeklinde ifade edilir. $M_1$ $s_1$ köküyle birlikte görülen ekler kümesi ve $M_2$ de $s_2$ köküyle birlikte görülen ekler kümesini ifade etmektedir. KL ıraksamanın da uygulanmasının ardından sonuç olarak 89 kök kategorisi elde edilmiştir.

| Pencere | Küme sayısı | Saflık |
|---|---|---|
| 4 | 50 | 0.93 |
| 5 | 50 | 0.92 |

**Tablo 6.** Skip-gram kümeleme sonuçları

İsimler ve fiillerin farklı kategorilerde kümelenmesi gerekmektedir. Fakat Türkçedeki sesli harf uyumlarından dolayı fonetik olarak farklı, görevsel olarak aynı ekleri alan sözcük kökleri farklı kategorilerde yer almıştır. Örneğin *lar* ve *ler* çoğul eklerini alan kökler farklı kategorilerde toplanmıştır. Bu problemi ortadan kaldırmak adına sesli harfler allofanlarıyla [17] değiştirilerek (örneğin *lar* ve *ler lAr* olarak birleştirilerek) yukarıdaki işlemler yine aynı sırada önce MI metriği ardından Jensen-Shannon ıraksama (divergence) yöntemi uygulandıktan sonra 33 adet kök kategorisi elde edilmiştir. Elde edilen kök kategorilerinden bazıları Tablo 4'te verilmiştir.

## III. DENEYLER VE SONUÇLAR

Yapılan deneyleri değerlendirmek için Morpho Challenge 2009'dan elde ettiğimiz 1 milyon cümlelik veri kümesinin ilk 5 bin cümlesi ele alınmıştır. Bu 5000 cümleden 5630 adet biricik sözcük kökü elde edilmiştir. Bu köklerden de hiç ek almamış ve sadece bir ek almış kökler çıkartılmıştır. Bunu sebebi bu köklerin kendisinden sonra gelen eklerle sınıflandırılabilmesi için yeterli bilgi içermemesidir. Özellikle bağlaç gibi zaten ek alması beklenmeyen sözcüklerin bu yöntemle sınıflandırılması mümkün değildir. Sonuç olarak elimizde 3049 adet biricik kök kalmıştır.

Elde edilen köklerin değerlendirilmesinde saflık (purity) değerlendirme ölçütü baz alınmıştır. Elde edilen kategorilerin saflık değerleri hesaplanırken karakteristik özellikleri daha belirgin olan isim, eylem ve sıfat türleri ele alınmıştır. Doğru sözcük kökü kümelerinin (gold clusters) oluşturulmasında sözcük köklerinin türlerini de verdiği için Zemberek [2] kullanılmıştır. Kök türlerinden sadece isim, eylem ve sıfat kümeleri ele alınmıştır. Saflık değerinin hesaplanmasında aşağıda verilen formül uygulanmıştır:

$$\sum purity(S; C) = \frac{1}{N} \sum_k \max_j |s_k \cap c_j| \quad (3)$$

| |
|---|
| ısıt, taşın, yararlan, çoğal, azal, yansı, payla, ula |
| temiz, ince, taze, pek, nitelik, sakin, geniş, güzel, derin |
| anket, geçmiş, cennet, çekirdek, klinik, mektep, disk, siyaset |

**Tablo 7.** Doğru kök kategorilerinden bazıları

*S* uyguladığımız kümeleme algoritması sonucunda elde edilen kök kümelerini ifade ederken, *C* ise Zemberek tarafından elde edilen doğru sözcük kökü kümelerini ifade etmektedir. Böylece her sonuç kümesi doğru kümelerden hangisiyle en fazla ortak köke sahipse o doğru kümeye eşleştirilmektedir. Doğru kök kategorilerinden bazıları Tablo 7'de verilmiştir.

Deneylerde CBOW ve skip-gram modelleri ayrı ayrı seçilerek farklı pencere boyutları ile sözcük köklerinin kategorize edilmesi sağlanmıştır. Sözcüklerin başlangıç ve bitişlerini belirten sınır karakteri de eklenerek de gerçekleştirilen deneylerde maksimum öbek sayısı 50 olarak belirlenmiş ve daha önceden seçilen sözcük kökleriyle son kümeler oluşturulmuştur.

İlk olarak CBOW modeli ile deneyler yapılmıştır. Bu deneyler sonunda saflık değeri en fazla 0.91 olarak bulunmuştur (bkz. Tablo 5). Pencere sayısının 5 olarak ayarlandığı deney en yüksek saflık değerini vermiştir. Aynı deneyler skip-gram modeli seçilerek tekrarlandığında saflık değerleri CBOW modeline göre daha yüksek çıkmıştır (bkz. Tablo 6). CBOW modelinden farklı olarak, skip-gram modelinde pencere sayısını azalması daha olumlu sonuç vermiştir.

Olasılıksal modelde ise oluşan son 33 kategorinin saflık değeri 0.88 olarak bulunmuştur.

## IV. SONUÇ VE GELECEK ÇALIŞMALAR

Bu çalışmada kök kategorileri sözdizimsel olarak kümelendirilmiştir. Bunun için morfolojik olarak bölünmüş sözcükler haricinde herhangi bir etiketlenmiş veri kullanılmadan denetimsiz bir öğrenme gerçekleştirilmiştir. Kümeleme işlemi için dağılımsal bir model olan word2vec modeli ile bilgi teorisi tabanlı ve olasılıksal bir model kullanılmıştır. word2vec sonuçları göstermektedir ki, sözcük köklerinin aldıkları eklerden ötürü dağılımsal olarak modellenmesi uygundur. Olasılıksal modelin ise word2vec modeline göre başarısı daha düşük çıksa da daha az sayıda ve daha kapsamlı kök kategorileri elde edilebilmiştir.

Sözcük köklerinin aldığı eklerden ziyade aldığı eklerin kategorileri sözcük kategorilerinin sınıflandırılmasında daha anlamlı olacaktır. Örneğin *taş-ın* sözcüğüne eklenen ekin emir kipi mi yoksa iyelik eki mi olduğu bilgisi burada ayırt edici bir bilgi olarak kullanılmalıdır. Bunu ise gelecek çalışmalarda ele almayı hedefliyoruz.

## KAYNAKÇA


[1] Mikolov, T., Le, Q. V., and Sutskever, I., "Exploiting similarities among languages for machine translation", *Comouting Research Repository*, 2013.

[2] Akın, A.A. & Akın, M.D., "Zemberek, an open source nlp framework for Turkic languages", *Structure*, vol. 10, p. 1-5, 2007.

[3] Kurimo, M., Lagus, K. S.V.V.T.: Morpho challenge 2009. http://research.ics.aalto.fi/events/morphochallenge2009/datasets.shtml#download/ (October 2015)

[4] Oflazer, K., "Two-level description of turkish morphology", *Literary and linguistic computing*, vol. 9, p. 137-148, 1994

[5] word2vec - Tool for computing continuous distributed representations of words. https://code.google.com/p/word2vec/, 2013.

[6] Mikolov, T., Chen, K., Corrado, G. and Dean, J. "Efficient estimation of word representations in vector space", *Comouting Research Repository*, 2013.

[7] Fortescue, M., Harder, P. and Kristoffersen, L. Hengeveld, K., "Parts of Speech." *Layered Structure and Reference in a Functional Perspective*. Ed. Amsterdam: John Benjamins, p. 29-55, 1992.

[8] Baek, D.H., Lee, H., chang Rim, H., "Conceptual clustering of korean concordances using similarity between morphemes"

[9] Kullback, S., Leibler, R.A., "On information and sufficiency", *The Annals of Mathematical Statistics*, vol. 22, p. 79-86, 1951

[10] Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C. and Mercer, R. L. "Class-based n-gram models of natural language", *Computational Linguistics*, vol. 18, p. 467-479, 1992.

[11] Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C. and Mercer, R. L., "Class-based n-gram models of natural language", *Computational Linguistics*, vol. 18, p. 467-479, 1990.

[12] Martin, S., Liermann, J. o. and Ney, H., "Algorithms for bigram and trigram word clustering", *Speech Communication*, vol. 24, p. 19-37, 1998.

[13] Whittaker, E. W. D. and Woodland, P. C., "Efficient class-based language modelling for very large vocabularies", *Acoustics, Speech and Signal Processing, (ICASSP)*, p. 545-548, 2001.

[14] Kneser, R. and Ney, H., "Improved clustering techniques for class-based statistical language modelling", European Conference on Speech Communication and Technology, p. 973-976, 1993.

[15] Uszkoreit, J. and Brants, T., "Distributed word clustering for large scale class-based language modeling in machine translation", Proc. of ACL, p. 755–762, 2008.

[16] Hogenhout, W. R. and Matsumoto, Y., "Training stochastical grammars on semantical categories", Stefan Wermter, Ellen Riloff, and Gabriele Scheler, editors, Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing. Springer. 1996.

[17] Oflazer, K., Gocmen, E., Bozsahin, C., "An Outline of Turkish Morphology", Technical Report, Bilkent University, 1994.

[18] Niesler, T., Whittaker, E., and Woodland, P., "Comparison of part-of-speech and automatically derived category-based language models for speech recognition", Acoustics, Speech and Signal Processing , vol. 1, p. 177-180, 1998.