Uygurca ile Türkçe Birleşik Sözcüklerin Karşılaştırılması

Murat Orhun
İstanbul Bilgi Üniversitesi, Bilgisayar Bilimleri Bölümü, İstanbul
muratmehmet@cs.bilgi.edu.tr

Özet: Bu bildiride, Uygurca birleşik sözcüklerin özelliği araştırılmış ve Türkiye Türkçesi ile karşılaştırma yapılmıştır. Doğal dil işlemlerinde, özellik ile bilgisayarlı çeviri çalışmalarında, birleşik sözcüklerin tespit edilmesi ve çözümlenmesi herhangi bir dil için çok zorluluklar oluşturmaktadır. Bunun temel nedeni ise, birleşik sözcüklerin çok karmaşık yapısından ve yazım kurallarında kaynaklanmaktadır. Eğer araştırılacak birleşik sözcükler, bir bitişken dil ile ilgili olduğunda, çalışmanın zorluluk derecesi daha artmaktadır. Çünkü bitişken dillerde bir kök sözcük birden fazla ek alabilir ve her yeni bir ek eklendiğinde, bir önceki sözcüğün anlamı, türü ve bazı sözcülerdeki harfler ses uyum sağlamsından dolayı değişir. Bir sözcüğün ek almasında sayı olarak her hangi bir kısıtlama yoktur. Bunun anlamı ise, bir sözcük ile sonsuz sözcükler oluşturulabilinir. Uygurca ve Türkçe ise bitişken dillerdir ve çok zengin eklere sahiptir. Dolaysı ile bitişken sözcülerin yapısı karmaşık ve çeşitli olarak oluşmaktadır. Bu bildiride, öncelik ile Uygurcanın dil yapısı kısaca tanıtılmıştır ve daha sonra birleşik sözcükler araştırılmış ve Türkçe ile karşılaştırılmıştır.

Anahtar Sözcükler: Uygurca Birleşik Sözcükler, Türkçe Birleşik sözcükler, Uygur Gramer, Bilgisayarlı Çeviri, Uygurca Doğal Dil İşleme.

1. Giriş

Bir dilde birleşik sözcük kullanılmasının temel nedeni ise, yeni bir kavramı karşılamak üzere oluşturulmasıdır. Bu çeşit sözcükler genel olarak birden fazla sözcüğün bir araya gelerek kaynaşmasından oluşur. Birleşik sözcükler bazen birleşik olarak yazılırken, bazen ayrı ayrı sözcükler şeklinde yazılır. Bundan dolayı, doğal dil çalışmasında, özellik ile bilgisayarlı çeviri çalışmalarında kaynak dildeki sözcüklerin doğru tespit edilmesi ve bu sözcüklerin hedef dile doğru aktarılması çok önemlidir. Dillerin doğal yapısından dolayı, birleşik sözcüklerin yapısı da farklı olmaktadır ve genel bir kural ve genel bir çözüm önermek oldukça zordur ve mümkün olmamaktadır. Birleşik sözcükler ile ilgili çalışmaları, bilgisayar ortamında yaparken, en çok karşılaşan zorluk ise birleşik sözcüklerin doğru tespit edilmesidir. Bu çeşit çalışmalar, doğrudan insanlar tarafından yapıldığında, çok kolay bir şekilde yapılırken, bir bilgisayar algoritması (programı) için muazzam zorluluk oluşturmaktadır. Çünkü, birleşik sözcükler oluşur ve o sözcükler ayrı olarak bir anlam verirken, birleşik olarak kullanılındığında da bir başka farklı anlam vermektedir. Söz konusu çalışılacak dil bitişken dil olduğunda, ilk önce yapılması gerek adım ise, biçim bilimsel çözümleyicinin oluşturulmasıdır. Çünkü bitişken dillerde bir kök sözcüğe, ya da bir sözcüğe bir ek eklendiğinde, ses uyumunun sağlanması nedeniyle bazı harfler değişir. Bitişken dillerdeki bu çeşit yapıların doğru çözümlenmesi için genel olarak, sözcük kökünün bulunması, tüm eklerin çözümlenmesi, dolaysıyla her farklı ek eklenirken sözcüğün nasıl etkilendiğinin araştırılması önerilir. Bu nenden ile Türkçe, Kırım Tatarcası,Türkmence, Fince ve Japonca gibi pek çok bitişken diller için biçim bilimsel çözümleyiciler geliştirilmiştir [1-5]. Örneğin, Türkçe'de bulunan "tarla faresi" sözcüğü, ayrı ayrı sözcük olarak Türkçe'nin biçim bilimsel çözümleyicisi [1] ile çözümlendiğine alttaki çözümlemeleri üretir.

tarla:
tarla+Noun+A3sg+Pnon+Nom

faresi: fare+Noun+A3sg+Pnon +Nom^DB+ Adj+JustLike

faresi: fare+Noun+A3sg+P3sg +Nom

"Tarla" sözcüğü ile ilgili etiketlerin anlamlan ise: sözcüğünün kökü "tarla" "Noun (isim)", "A3sg (tekil)", "Pnon (şahıs eki yok)" ve "Nom" (durum eki yok). "Faresi" sözcüğü iki farklı sözcük olarak çözümlenmiştir. Birinci çözüm ise, sözcüğün kökü ise "fare" ve isimdir. Ancak "si" ekinin eklenmesi ile, "fare gibi" anlam oluşmaktadır ve bu durm "JustLike" etiketi ile belirtilmiştir. Sonucunda "faresi" sözcüğü, "isim" tipinden "sıfat" tipine değişmiş oluyor ve bu durum "Adj" etiketi ile belirtilmiştir. Çözümlemede bulunana "^DB" ise türetim etiketidir. Yani "faresi" sözcüğü aslında bir "isimdir" ve "si" ekinin eklenmesi ile "sıfata" dönüşmüştür.

"Faresi" sözcüğü ile ilgili ikinci çözümde, "DB" etiketi oluşturulmamıştır. Yanı "isim" olarak çözümlenmiştir. Ancak "P3sg" etiketi oluşmuştur. Bu etiketin anlamı ise, "fare" nesnesinin 3- bir şahısa ait olduğunu göstermektedir.

Sonuç olarak. "tarla faresi" sözcüğü ile ilgili biçim bilimsel bilgiler birleştirildiğinde, "tarla fare gibi" ve "tarlanın faresi" gibi bir anlam oluşuyor. Ancak "tarla faresi" ise tarlada yaşayan bir çeşit faredir. Bu çeşit fare için yeni bir isim de oluşturula bilirdi. Tıpkı gönümüzde sürekli gelişen bilgisayar teknolojileri ve ürünleri için yeni isimi üretildiği gibi.

Bu örnekte kullanılan "tarla faresi" gerçi ayrı sözcük şeklinde yazılsa bile, onlar birleşik sözcüklerdir. Bu neden ile bu sözcükleri, tek sözcükler gibi analiz yapmak yanlış sonuçlara neden olacaktır.

Bu bildiride Uygurca için geliştirilen biçim bilimsel çözümleyici kullanılarak [6-7], Uygurca sözcükler üzerinde çözümleme yapılmıştır ve birleşik sözcüklerin tespit edilmesi ve ayni yapıdaki Türkçe birleşik sözcükler üzerinde karşılaştırma yapılmıştır.

Bu bildirinin 2- bölümünde Uygurca sözcük yapıları ve ekleri, 3- bölümünde Uygurca birleşik sözcükler incelenmiş ve Türkçe birleşik sözcükler ile karşılaştırma yapılmıştır. Son bölümde ise Uygurca birleşik sözcüklerin bilgisayar ortamında Türkçe'ye çevirilerken olası önerilerde bulunmuştur.

2. Uygurca Sözcükler

Uygurca Ural-Altay dilleri ailesinin Türk dilleri grubuna ait bitişken dildir. Yoğun olarak Doğu Türkistan yerlileri Uygurlar tarafından kullanılır. Uygurlar geçmişte pek çok farklı alfabe kullansa bile, günümüze Arap alfabesi temelinde geliştirilen Uygur- Arap alfabesi kullanıyorlar. Ancak bu bildiride, örnekler gösterilirken, Latin alfabe temelinde geliştirilen Uygur-Latin alfabesi (Uygur komputer alfabesi) kullanılır [8-9].

Uygur alfabesi, 24 sessiz ve 8 sesli olamak üzere toplam 32 harften oluşur. Uygurcada bulunan tüm sözcüklere çeşitli çekim ve yapım ekleri eklenebilinir. Ekler eklenirken, genel olarak ses uyumu uyuşma şartı aranır. Ancak Uygurca çok eski bir dil olduğundan, farklı dillerden kabul edilen sözcükler için ses uyuma kuralı geçersizdir [10-12]. Örneğin Arapça, Farsça, Rusça, İngilizce ve son zamanlar Çince'den kabul edilen sözcüklerin çoğalmasıyla, bu çeşit sözcüklere ekleme yaparken ses uyum kurallarını uygulanması mümkün olamamaktadır. Bu neden ile bu çeşit sözcükler için özel kurallar tanımlanması gerekmektedir. Uygurca tek sözcüklere eklerin eklenmesi, Türkçeye ve Türk dilleri grubunda bulunan başka dillere benzemektedir [13-14]. Örneğin:

otan (odun)
otan+ ci (oduncu)

ash (ash, yemek)
ash+xane (yemekhane)

sawaq (eğitim)
sawaq+dash (öğrenci, talebe)

ana (anne)
ana+m (annem)

sen (sen)
sen+mu? (sen mi?)

ish (iş)
ish+le (çalış)
ish+le+watimen (çalışıyorum)

Yukarıda belirtilen örneklerde, Uygurca basit sözcükler ve onlara eklenen çekim ve yapım ekleri ve Türkçe karşılıkları belirtilmiştir. Bu çeşit sözcükler analiz edilirken, biçim bilimsel çözümleyici ile doğrudan çözümlenebilinmektedir. Bu neden ile bu çeşit sözcükleri bir başka dile çeviri yaparken, biçim bilimsel bilgilere dayanarak doğrudan (burada biçimbilimsel ve sözcüğün anlamsal belirsizlik giderildiği kabul edilmiştir) çeviri yapılabilinir.

3. Uygurca ve Türkçe Birleşik Sözcükler

Uygurca'da yeni deyimleri ve yeni olayları adlandırılmak üzere, tıpkı başka dillerde olduğu gibi birleşik sözcükler kullanılır. Bazı birleşik sözcükleri kullanmak için 2 sözcük kullanılırken, bazı durumlarda ise 3 ya da 4 sözcüklerden oluşmaktadır.

3.1 İki Sözcükten Oluşan Birleşik Sözcükler

Uygurcada çoğu zaman 2 sözcükten oluşan birleşik sözcükler bitişik yazılır. Örneğin:

köz (göz)
eynek (ayna)
közeynek (gözlük) = köz + eynek

ash (yemek)
qazan (tencere)
ashqazan (mide) = ash + qazan

ete (yarın)
yaz (yaz mevsim)
eteyaz (bahar) = ete + yaz

Yukarıda gösterilen örneklerde, Uygurcada gösterilen birleşik sözcükler iki farlı sözcüklerden oluşmuştur bitişik yazılmıştır. Türkçe kaşılığı da tek sözcük ile çevrilmiştir. Bu çeşit birleşik sözcükler Türkçe'de gösterilebilinir. Örneğin:

bilgisayar = bilgi + sayar
yıldönümü = yıl + dönümü
sivrisinek = sivri + sinek
köpekbalığı = köpek + balığı
pazartesi = pazar + ertisi

Yukarıda gösterilen sözcükler konuşulurken iki ayrı sözcük gibi söylenir, ancak yazılırken bitişik yazılır.

3.2 İkiden Fazla Sözcükten Oluşan Birleşik Sözcükler

Uygurcada bazı birleşik sözcükler ise ikiden fazla basit sözcüklerin katılması ile oluşur ve bu sözcükler ayrı ayrı yazılır. Bu çeşit sözcükler genel olarak en az iki sözük gibi ayrı yazılır. Örneğin:

ghaz (gaz)
qanat (kanat)
ot (ot)
ghazqanat ot = ghaz + qanat ot
(bir çeşit bitkidir)

köz (göz)
eyenklik (aynalı) = eynek (ayna)
+li

yilan (yilan)
közeyneklik yilan (korba) = köz + eyneklik yılan

edir (owa)
ghaz (gaz)
qanat (kanat)
ot (ot)
edir ghazqanat ot = edir ghaz + qanat ot
(owada bir çeşit bitkidir)

Yukarıda gösterilen örneklerde birleşik sözcükleri oluşturan sözcüklerden bazı sözcük-

ler ayrı yazılırken, bazı sözcükler ise bitişik yazılmıştır. Yanı birleşik sözcükleri oluşturan sözcükler kendileri bile birleşik sözcüklerden oluşmuşutur. Örneğin, "edir ghazqanat ot" bir birleşik sözcüktür. Bu çeşit birleşik sözcüklerin bazıları Türkçe'ye tek sözcük ile çevrilirken, örenğin (közeyneklik yilan= korba), bazı sözcükler ise bir den fazla sözcükler ile çevrilmek zorundadır.

3.3 Birleşik Kişi isimleri

Uygurcada kişi isimleri oluşturulurken, tek isim kullanılırken, bazen ikiden fazla kişi isimleri oluşan birleşik isimlar kullanıır. Uygurcada iki farklı sözcüklerden oluşan kişi adları genelde bitişik yazılır. Bu adları oluşturan sözcükler bitişik yazılırken, Uygurcada bulunan sesli ve sessiz harflerin uyumu sağlanmaz. Yanı sözcükle olduğu gibi birleştirilerek yazılır. Örneğin:

Tursunmemet = Tursun + Mehmet
Memetsidiq = Mehmet + Sidiq
Qadiraxun = Qadir + Axun
Toxtiqari = Toxti + Qari
NurMemet = Nur + Memet
AbidiQadir= Abdi+ Qadir

Yukarıda gösterilen kişi adları Türkçe'ye çevrilirken Uygurca'da kullanılan "Memet" sözcüğü genelde "Mehmet" olarak çevrilir. Bunun temel nedeni ise, Türkçe'de genelde "memet" isimi bulunmamaktadır.

Uygurca'da bazı tek sözcükten oluşan kişi isimları özel anlamlar içerir. Örneğin:

Polat (çelik)
Yolwas (qaplan)
Xehriman (kahriman)
Sübihi (sabah)
Leyligül (layla)

Yukarıda gösterilen kişi adları Türkçe'de kullanılan kişi adlarına benzer. Türkçe'de genel olarak özel isimların ilk harfi büyük yazılır. Ancak Uygurca'da Arap alfabesi kullanıldığından, büyük-küçük karakter farkı gösterilemez. Bu nenden ile, Uygurcada tek sözcüklerden oluşan kişi adlarını ayrıt etmek için özel çalışmalar yapılması gerekmektedir. Sözcüklerin anlamına göre ayrıt edilmesine, kendi başına ayrı bir araştırma konusudur ve bu konuda, Türkçe için çok değerli çalışmalar yapılmıştır [15-17].

4. Sonuç ve Öneriler

Bu bildiride Uygurca birleşik sözcükler ve yapısı incelenmiş ve Türkçe birleşik sözcükler ile karşılaştırılmıştır. Doğal dil çalışmalarında özel isimlerin tespit edilmesinde yaygın kullanılan yöntemlerden biri ise derleme dayalı çalışan istatistiksel yöntemlerdir. Ancak Türk dilleri ailesinde sadece Türkçe için geliştirilmiş derlem bulunurken [18], Uygurca için derlem çalışmaları devem etmektedir [19-20]. Dolayısıyla, Uygurca birleşik sözcüklerin tespit edilmesinde, istatistiksel yöntemlerin kullanılması başarısız olacaktır. Derlem çalışmaları ise uzun zaman alan ve karmaşık ve pahalı bir çalışma olduğundan, Uygurca birleşik sözcüklerin tespit edilmesi için, ayrı bir birleşik sözcükler listesi oluşturulması ve özel etiketler ile etiketlenmesi önerilmektedir. Böylece bu sözcükler biçim bilimsel çözümleyici tarafından, özel olarak çözümlenir. Ayrıca, birleşik sözcüklerden söz ederken, özel deyimler ve ata-sözleri gibi yapılar için farklı bir çalışma yapılması gerekmektedir. Çünkü bir dilde bulunan bir atasöz, doğrudan ikinci bir dile çevrildiğinde çok farklı anlamlar vermektedir.

5. Kaynaklar

[1] Oflazer, K., "Two-level Description of Turkish Morphology", Literary and Linguistic Computing, vol. 9, 137-148, (1995).

[2] Altıntaş, K., ve Çiçekli, İ., "A Morphological Analyser for Crimean Tatar", Proceedings of the 10th Turkish Symposium on Artificial Intelligence and Neural Networks, TAINN North Cyprus, pp. 180-189, (2001).

[3] Tantuğ A.C., Adalı E., Oflazer K., "Computer Analysis of The Turkmen Language Morphology", Proceedings of the 5th International Conference on Natural Language Processing, FinTAL, Turku, Finland, (2006).

[4] Koskenniemi, K., "An Application of the Two-Level Model to Finnish," University of Helsinki Department of General Linguistics , (1985).

[5] Alam, Y. S., "A Two-Level Morphological Analysis of Japanese", Texas Linguistics Forum, 22:229-252, (1983).

[6] Orhun, M., Tantuğ, A.C., Adalı, E., "Rule Based Analysis of the Uyghur Nouns", Proceedings of the International Conference on Asian Language Processing (IALP). Chiang Mai, Thailand, (2008).

[7] Orhun, M., Tantuğ, A.C. ve Adalı,E., "Rule Based Tagging of the Uyghur Verbs", Fourth International Conference on Intelligent Computing and Information Systems. Faculty of Computer &Information Science, Ain Shams University Cairo, Egypt, (2009).

[8] Duval J.R., Janbaz, W, A., "An Introduction to Latin-Script Uyghur", Middle East & Central Asia Politics, Economics, and Society Conference. Sept 7-9, University of Utah, Salt Lake City, USA, (2006).

[9] UKIJ, Uyghur Komputer Ilmi Jemiyeti. http://www.ukij .org/.

[10] Tömür, H., "Modern Uygur Grammar (Morphology)", Yildiz Teknik Üniversitesi, Fen-Ed Fak. T.D.E Bölümü, Istanbul, (İngilizce baskı), (1997).

[11] Osmanof, M., "Hazirqi Zaman Uyghur Edebiy Tilining İmla ve Teleppuz Lughiti", Shin Jiang Xeliq Neshiryatı, (Uygurca baskı), (1997).

[12] Ridivan Öztürk, "Yeni Uygur Türkçesi Grameri", Türk dil kurumu yayınları: 593, (1993).

[13] Hengirmen, M., "Türkçe Dilbilgisi", Engin yayınevi, Ankara, (2002).

[14] Tantug, A.C., "A Hybrid Model For Machine Translation Between Agglutinative and Related Languages", PhD Thesis, Istanbul Technical University, (2007).

[15] Orhan, Z. ve Altan, Z., "Anlam Belirsizliği İçeren Türkçe Sözcüklerin Hesaplamalı Dilbilim Uygulamalarıyla Belirginleştirmesi", XX. Ulusal Dilbilim Kurultayi, İstanbul., (2006)

[16] Orhan, Z. ve Altan, Z., Determining Efective Features for Word Sense Disambiguation in Turkihs, Journal of Electrical & Electronics Engineering, Istanbul University, Vol3,No:2, (2005).

[17] Oflazer, K., Çetinoğlu, Ö. ve Say, B., "Integrating Morphology with Multi-word Expression Processing in Turkish", Second ACL Workshop on Multiword Expressions: Integrating Processing, July pp:64-71, (2004).

[18] Oflazer, K., Say, B., Hakkani-Tür, D.Z., Tür, G., "Building a Turkish Treebank". In Abeille, A., ed.: Building and Exploiting Syntactically-annotated Corpora. Kluwer Academic Publisher, (2003).

[19] Yusup Aibaidulla and Kim-Teng Lua, "The development fo Tagged Uyghur Corpus", Proceedings of PACLIC17, 1-3 Oct, Sentosa, Singapore, P228-234, (2003).

[20] Abaidula, Y., Rezwangul, Sali, A., "The Research and Development of Computer Aided Contemporary Uighur Language Tagging System", Journal of Chinese Language and Computing , 15(4), pp: 203210, Vol.15, No:4, (2005).