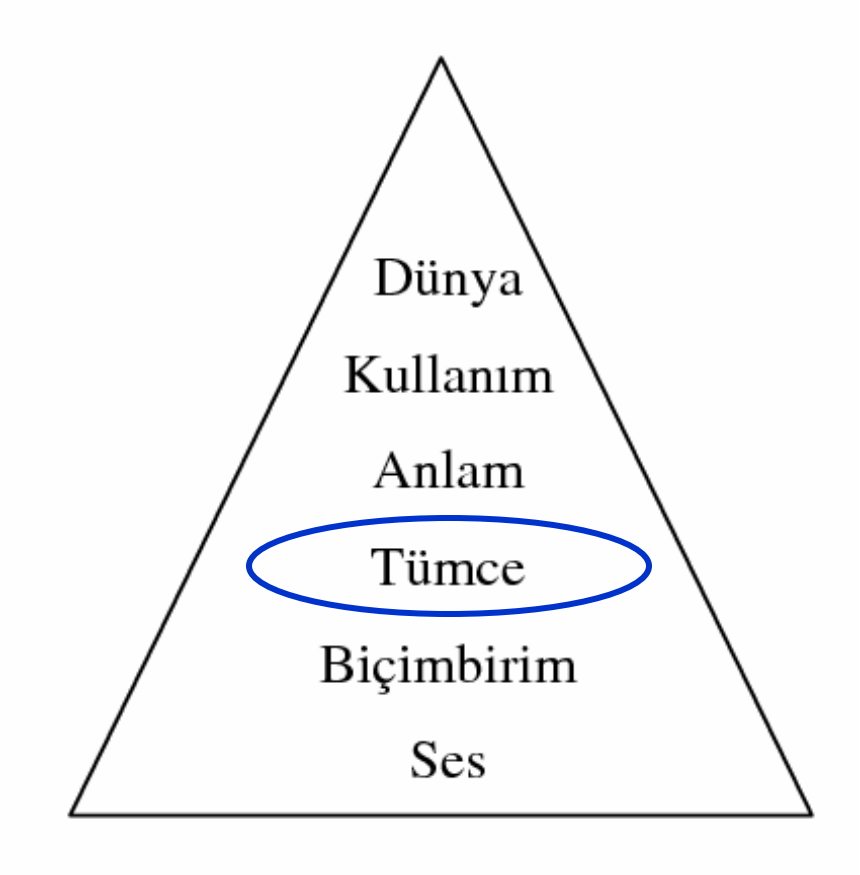
Türkçe'nin Bağlılık Ayrıştırması

Gülşen Cebiroğlu Eryiğit

Bağlılık Ayrıştırması

Doğal Dil İşleme ve Bölümleri



Bağlılık Ayrıştırması

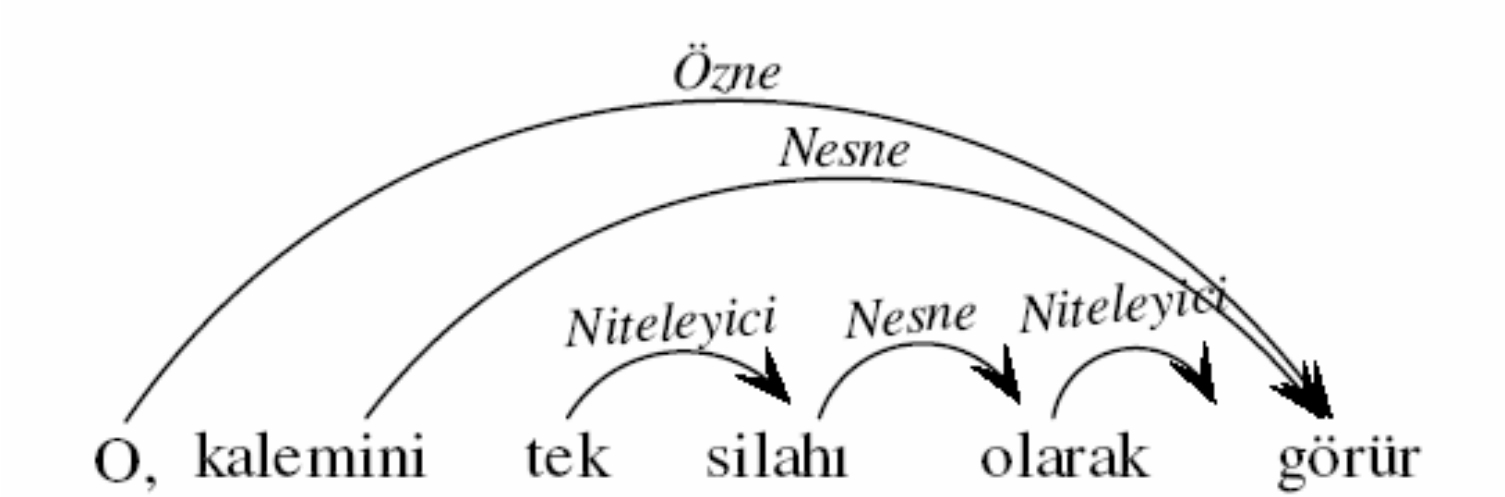
- Sözcükler arasındaki ikili bağlılık ilişkilerinin ayrıştırma başarımındaki önemli etkisinin anlaşılması,
- Bağlılık Ayrıştırması yönteminin, tümce içi sözcük dizilişleri serbest diller üzerindeki yetenekleri,
- Üst düzey uygulamalar için anlamlı bilgi üretmesi,

bu yöntemin son yıllarda sıkça kullanılır hale gelmesini sağlamıştır.

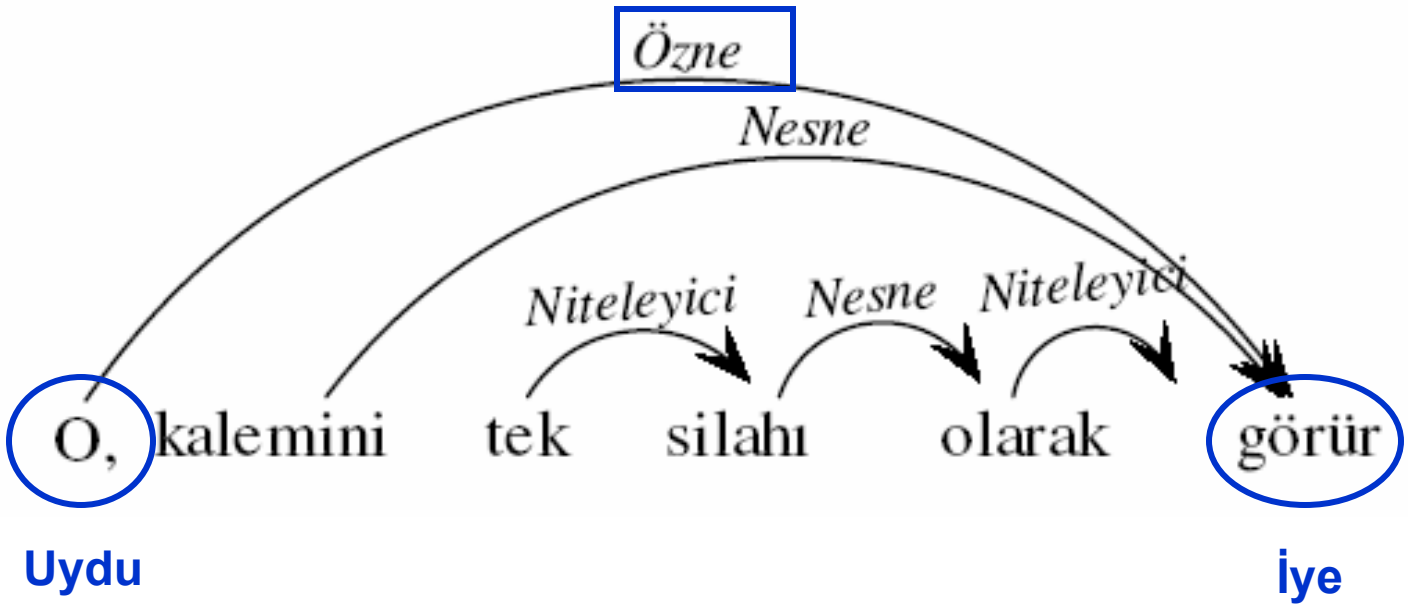
Bağlılık Ayrıştırması

- Güncel bağlılık grameri kuramının, Tesnière'in 1959'daki çalışmasına dayandığı söylenebilir.
- Tesnière'e göre ``Tümce, kendisini oluşturan öğeleri sözcükler olan düzenli bir topluluktur''
- ``Zihin, tümceyi oluşturan sözcükler ve komşuları arasında ilişkileri bulur ve bu ilişkilerin bütünü tümcenin iskeletini oluşturur. Her bir ilişki bir alt terimi bir üst terime bağlamaktadır.''
- Günümüzde DDA alanında kullanılan bağlılık gramerlerinde bu ilişki uydu (alt terim) - iye (üst terim) ilişkisi olarak tanımlanmaktadır. Bağlılık grameri tabanlı metin ayrıştırmasının amacı metin içerisinde geçen her tümce için tümceyi oluşturan sözcükler arasındaki uydu-iye ilişkilerini bulmaktır.

Bağlılık Ayrıştırması



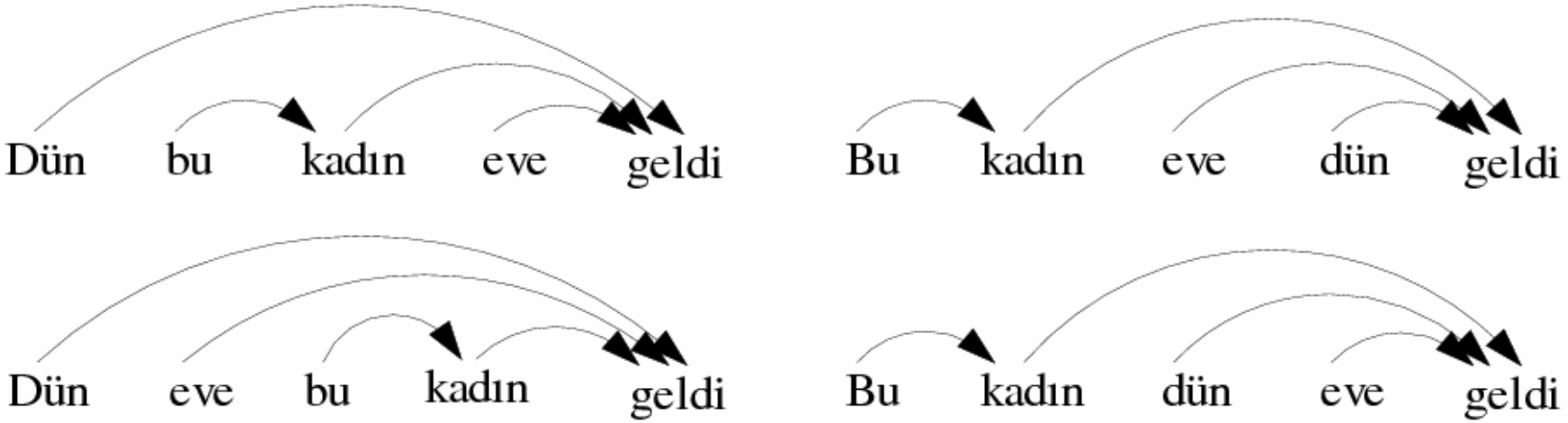
Bağlılık Ayrıştırması



Türkçe

Tümce içi öğe dizilişleri serbest

- Genelde ÖNY veya NÖY kalıpları



Türkçe

Çok zengin bitişken biçimbirimsel yapı

- Eklerin sona eklenmesiyle yüzlerce farklı yeni sözcük

gidiyorum
gidiyorsun
gideceğim
gideceksin
gidebilirim
gitmekteyim
.....

Türkçe

Çok zengin bitişken biçimbirimsel yapı

- Eklerin sona eklenmesiyle yüzlerce farklı yeni sözcük
- Sözcük çeşitliliğindeki zenginlik
- Sözcük etiketlerinin durum, kişi, sayı, cinsiyet gibi birçok bilgiyi taşımaları ve bu nedenle çok sayıda etiket oluşması

Türkçe

Zengin türetim yapısı

Veda – laş – ma

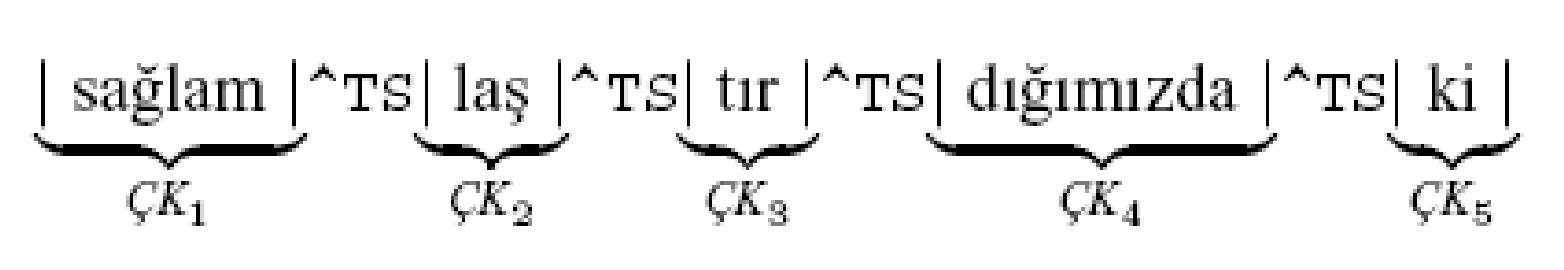
Oku – t – ul – an

Türkçe
Çekim Kümeleri

sağlamlaştırdığımızdaki

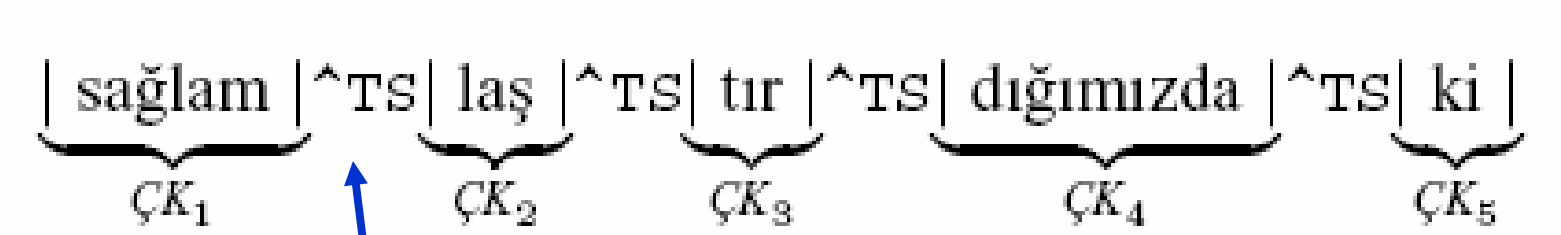
Türkçe - Çekim Kümeleri

sağlamlaştırdığımızdaki



Türkçe - Çekim Kümeleri

sağlamlaştırdığımızdaki



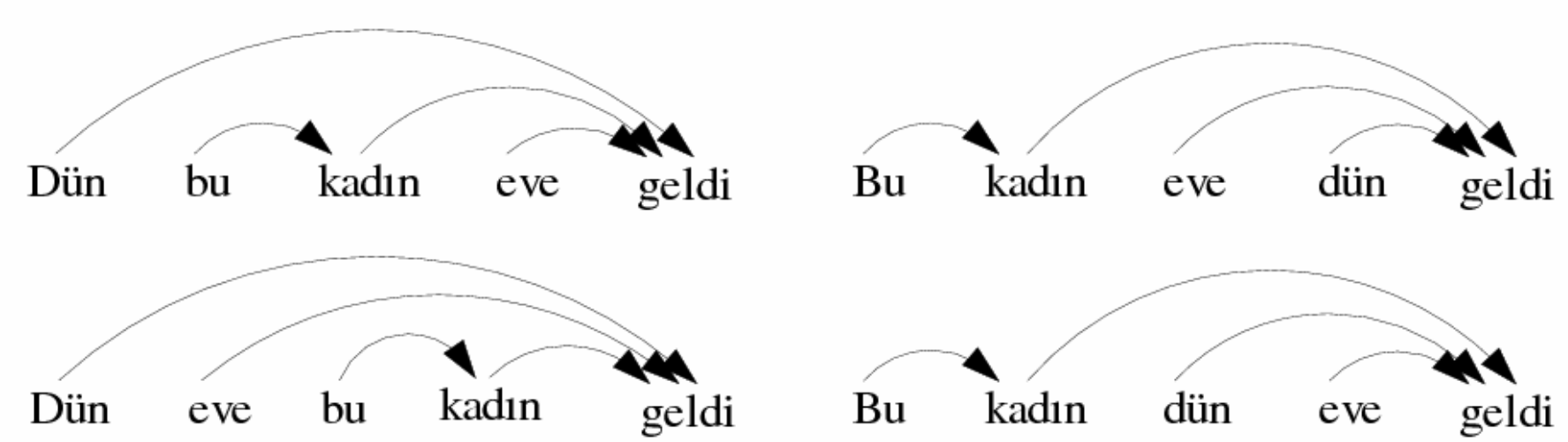
Sözcük başına ortalama 1.26 ÇK

Çekim Kümesi

Türetim Sınırı

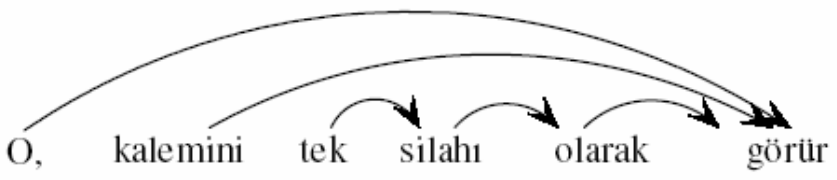
Türkçe

- Çoğunlukla sağa bağımlı türde bağlılıklar

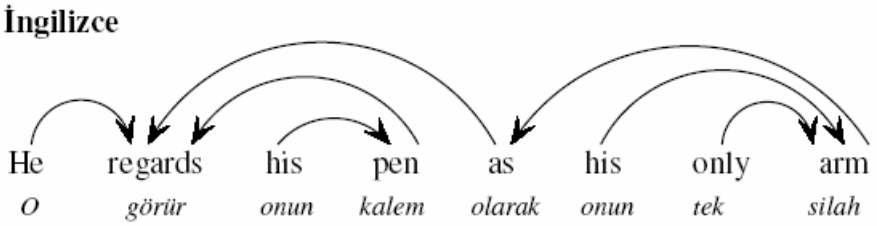


Türkçe

Türkçe

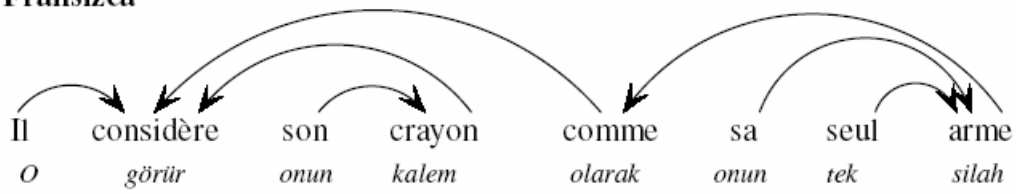


İngilizce

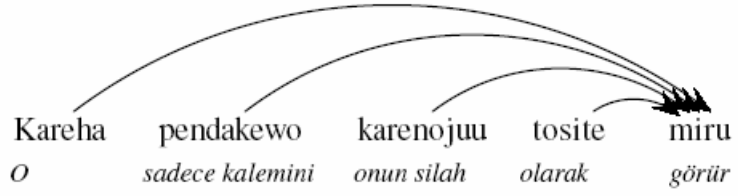


Fince

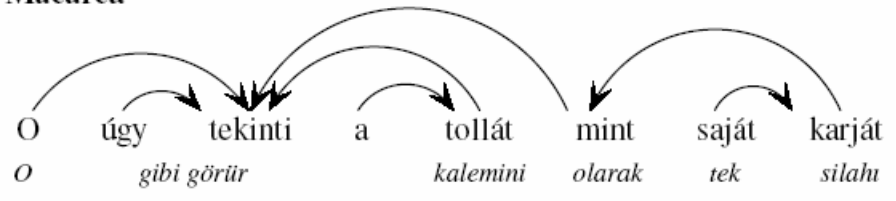
Fransızca



Japonca



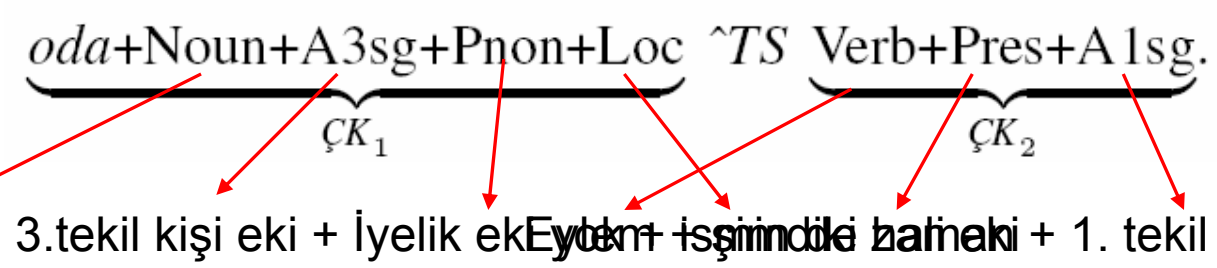
Macarca



Türkçe - Bağlılık Yapısı

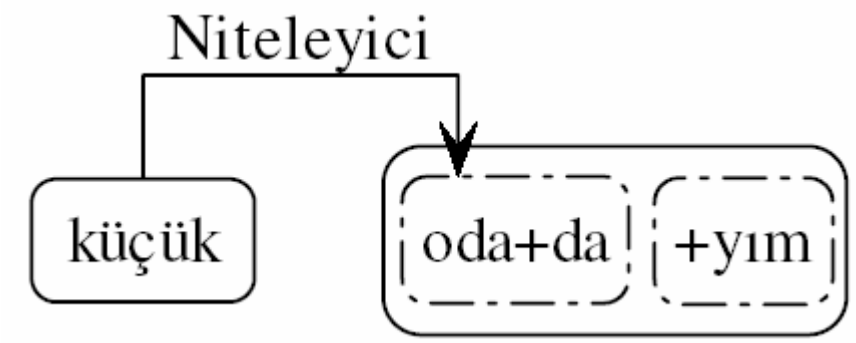
Bağlılıklar ÇK'ler arasında kurulur

küçük odadayım



Türkçe - Bağlılık Yapısı

Bağlılıklar ÇK'ler arasında kurulur



Türkçe Ağaç Yapılı Derlem

ODTÜ-Sabancı Türkçe Ağaç Yapılı Derlem

- 5635 tümce
- yetkin biçimbirimsel etiketler
- etiketli bağlılıklar
- bağlılıkların %95'i sağa bağımlı bağlılıklar
- XML biçiminde
- ÇK'ler arası bağlılıklar

Derlem

```
<?xml version="1.0" encoding="windows-1254" ?>
<Set sentences="1">
    <S No="1">
      <W IX="1" LEM="" MORPH="" IG="[(1,"bu+Det")]"
         REL="[2,1,(DETERMINER)]">Bu</W>
      <W IX="2" LEM="" MORPH=""
         IG="[(1,"okul+Noun+A3sg+Pnon+Loc")(2,"Adj+Rel")]"
         REL="[3,1,(MODIFIER)]">okuldaki</W>
      <W IX="3" LEM="" MORPH=""
         IG="[(1,"öğrenci+Noun+A3pl+Pnon+Gen")]"
         REL="[5,3,(POSSESSOR)]">öğrencilerin</W>
      <W IX="4" LEM="" MORPH="" IG="[(1,"en+Adv")]"
         REL="[5,2,(MODIFIER)]">en</W>
      <W IX="5" LEM="" MORPH=""
         IG="[(1,"akıl+Noun+A3sg+Pnon+Nom")(2,"Adj+With")
         (3,"Noun+Zero+A3sg+P3sg+Nom")]"
         REL="[9,2,(SUBJECT)]">akıllısı</W>
      <W IX="6" LEM="" MORPH=""
         IG="[(1,"şura+Noun+A3sg+Pnon+Loc")]"
         REL="[7,1,(LOCATIVE.ADJUNCT)]">şurada</W>
      <W IX="7" LEM="" MORPH=""
         IG="[(1,"dur+Verb+Pos")(2,"Adj+PresPart")]"
         REL="[9,1,(MODIFIER)]">duran</W>
      <W IX="8" LEM="" MORPH="" IG="[(1,"küçük+Adj")]"
         REL="[9,1,(MODIFIER)]">küçük</W>
      <W IX="9" LEM="" MORPH=""
         IG="[(1,"kız+Noun+A3sg+Pnon+Nom")
         (2,"Verb+Zero+Pres+Cop+A3sg")]"
         REL="[10,1,(SENTENCE)]">kızdır</W>
      <W IX="10" LEM="" MORPH="" IG="[(1,".+Punc")]"
         REL="[,( )]">.</W>
    </S>
</Set>
```

Derlem

```
<?xml version="1.0" encoding="windows-1254" ?>
<Set sentences="1">
    <S No="1">
      <W IX="1" LEM="" MORPH="" IG="[(1,"bu+Det")]"
        REL="[2,1,(DETERMINER)]">Bu</W>
      <W IX="2" LEM="" MORPH=""
        IG="[(1,"okul+Noun+A3sg+Pnon+Loc")(2,"Adj+Rel")]"
        REL="[3,1,(MODIFIER)]">okuldaki</W>
      <W IX="3" LEM="" MORPH=""
        IG="[(1,"öğrenci+Noun+A3pl+Pnon+Gen")]"
        REL="[5,3,(POSSESSOR)]">öğrencilerin</W>
      <W IX="4" LEM="" MORPH="" IG="[(1,"en+Adv")]"
        REL="[5,2,(MODIFIER)]">en</W>
      <W IX="5" LEM="" MORPH=""
        IG="[(1,"akıl+Noun+A3sg+Pnon+Nom")(2,"Adj+With")
        (3,"Noun+Zero+A3sg+P3sg+Nom")]"
        REL="[9,2,(SUBJECT)]">akıllısı</W>
      <W IX="6" LEM="" MORPH=""
        IG="[(1,"şura+Noun+A3sg+Pnon+Loc")]"
        REL="[7,1,(LOCATIVE.ADJUNCT)]">şurada</W>
      <W IX="7" LEM="" MORPH=""
        IG="[(1,"dur+Verb+Pos")(2,"Adj+PresPart")]"
        REL="[9,1,(MODIFIER)]">duran</W>
      <W IX="8" LEM="" MORPH="" IG="[(1,"küçük+Adj")]"
        REL="[9,1,(MODIFIER)]">küçük</W>
      <W IX="9" LEM="" MORPH=""
        IG="[(1,"kız+Noun+A3sg+Pnon+Nom")
        (2,"Verb+Zero+Pres+Cop+A3sg")]"
        REL="[10,1,(SENTENCE)]">kızdır</W>
      <W IX="10" LEM="" MORPH="" IG="[(1,".+Punc")]"
        REL="[,( )]">.</W>
    </S>
</Set>
```

Derlem

```
<?xml version="1.0" encoding="windows-1254" ?>
<Set sentences="1">
    <S No="1">
      <W IX="1" LEM="" MORPH="" IG="[(1,"bu+Det")]"
        REL="[2,1,(DETERMINER)]">Bu</W>
      <W IX="2" LEM="" MORPH=""
        IG="[(1,"okul+Noun+A3sg+Pnon+Loc")(2,"Adj+Rel")]"
        REL="[3,1,(MODIFIER)]">okuldaki</W>
      <W IX="3" LEM="" MORPH=""
        IG="[(1,"öğrenci+Noun+A3pl+Pnon+Gen")]"
        REL="[5,3,(POSSESSOR)]">öğrencilerin</W>
      <W IX="4" LEM="" MORPH="" IG="[(1,"en+Adv")]"
        REL="[5,2,(MODIFIER)]">en</W>
      <W IX="5" LEM="" MORPH=""
        IG="[(1,"akıl+Noun+A3sg+Pnon+Nom")(2,"Adj+With")
        (3,"Noun+Zero+A3sg+P3sg+Nom")]"
        REL="[9,2,(SUBJECT)]">akıllısı</W>
      <W IX="6" LEM="" MORPH=""
        IG="[(1,"şura+Noun+A3sg+Pnon+Loc")]"
        REL="[7,1,(LOCATIVE.ADJUNCT)]">şurada</W>
      <W IX="7" LEM="" MORPH=""
        IG="[(1,"dur+Verb+Pos")(2,"Adj+PresPart")]"
        REL="[9,1,(MODIFIER)]">duran</W>
      <W IX="8" LEM="" MORPH="" IG="[(1,"küçük+Adj")]"
        REL="[9,1,(MODIFIER)]">küçük</W>
      <W IX="9" LEM="" MORPH=""
        IG="[(1,"kız+Noun+A3sg+Pnon+Nom")
        (2,"Verb+Zero+Pres+Cop+A3sg")]"
        REL="[10,1,(SENTENCE)]">kızdır</W>
      <W IX="10" LEM="" MORPH="" IG="[(1,".+Punc")]"
        REL="[,( )]">.</W>
    </S>
</Set>
```

Derlem

```
<?xml version="1.0" encoding="windows-1254" ?>
<Set sentences="1">
    <S No="1">
      <W IX="1" LEM="" MORPH="" IG="[(1,"bu+Det")]"
        REL="[2,1,(DETERMINER)]">Bu</W>
      <W IX="2" LEM="" MORPH=""
        IG="[(1,"okul+Noun+A3sg+Pnon+Loc")(2,"Adj+Rel")]"
        REL="[3,1,(MODIFIER)]">okuldaki</W>
      <W IX="3" LEM="" MORPH=""
        IG="[(1,"öğrenci+Noun+A3pl+Pnon+Gen")]"
        REL="[5,3,(POSSESSOR)]">öğrencilerin</W>
      <W IX="4" LEM="" MORPH="" IG="[(1,"en+Adv")]"
        REL="[5,2,(MODIFIER)]">en</W>
      <W IX="5" LEM="" MORPH=""
        IG="[(1,"akıl+Noun+A3sg+Pnon+Nom")(2,"Adj+With")
        (3,"Noun+Zero+A3sg+P3sg+Nom")]"
        REL="[9,2,(SUBJECT)]">akıllısı</W>
      <W IX="6" LEM="" MORPH=""
        IG="[(1,"şura+Noun+A3sg+Pnon+Loc")]"
        REL="[7,1,(LOCATIVE.ADJUNCT)]">şurada</W>
      <W IX="7" LEM="" MORPH=""
        IG="[(1,"dur+Verb+Pos")(2,"Adj+PresPart")]"
        REL="[9,1,(MODIFIER)]">duran</W>
      <W IX="8" LEM="" MORPH="" IG="[(1,"küçük+Adj")]"
        REL="[9,1,(MODIFIER)]">küçük</W>
      <W IX="9" LEM="" MORPH=""
        IG="[(1,"kız+Noun+A3sg+Pnon+Nom")
        (2,"Verb+Zero+Pres+Cop+A3sg")]"
        REL="[10,1,(SENTENCE)]">kızdır</W>
      <W IX="10" LEM="" MORPH="" IG="[(1,".+Punc")]"
        REL="[,( )]">.</W>
    </S>
</Set>
```

Sınıflandırıcı Tabanlı Ayrıştırıcı

- Bağlılık grafiğini oluşturmak için kullanılan **gerekirci bir ayrıştırma algoritması**, (Kudo ve Matsumoto, 2002; Yamada ve Matsumoto, 2003; Nivre,2003)
- Ayrıştırıcının bir sonraki hareketini belirlemek üzere kullanılan **geçmişe dayalı ayrıştırma modeli** (Black ve diğ., 1992; Magerman, 1995; Collins, 1999)
- Geçmişte olan olayları ayrıştırıcının hareketleri ile ilişkilendirmek üzere kullanılan **ayırdedici sınıflandırıcı** (Veenstra ve Daelemans, 2000; Kudo ve Matsumoto,2002; Nivre ve diğ., 2004) (KDM: karar destek makineleri)

Geçmişe dayalı özellik modeli

Hedef birimler ve bunlarla ilişkili birimler için özellik vektöründe kullanılabilecek özellikler:

- Görünüm bilgisi (tümü veya gövdesi)
- Sözcük sınıfı (ana sınıf veya alt sınıf)
- Biçimbirimsel özellikler
- Bağlılık türü (Eğer bağlanmışsa)

Tasarım Modelleri

- Birim Seçim Modelleri
  - Sözcük Tabanlı Model
  - ÇK Tabanlı Model
  - ÇK Tabanlı Belirlenimci Model
- Biçimbirimsel Özelliklerin Kullanımı ile ilgili Modeller
  - ÇK tabanlı (INF birleşik) model
  - ÇK tabanlı (INF parçalı) model

Birim Seçim Modelleri

Sözcük Tabanlı Model

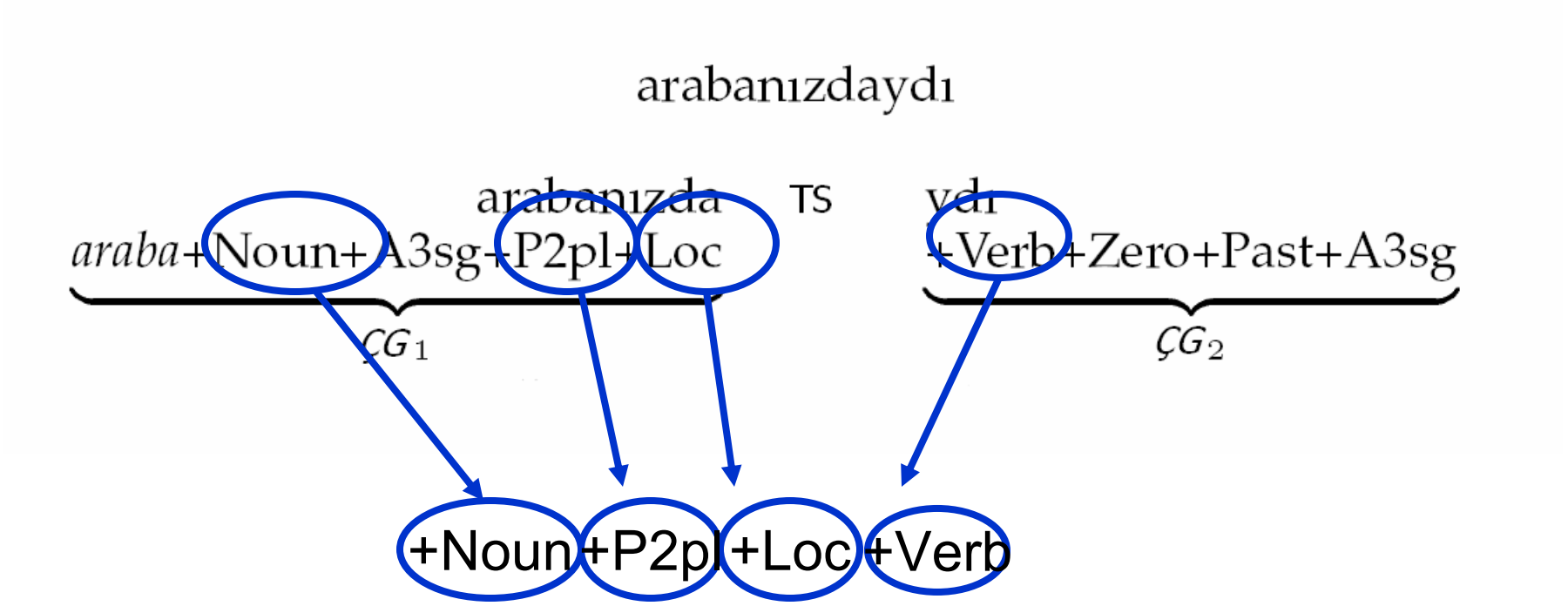
arabanızdaydı

arabanızda TS ydı

$\underbrace{araba\text{+Noun+A3sg+P2pl+Loc}}_{ÇG_1}$ $\underbrace{\text{+Verb+Zero+Past+A3sg}}_{ÇG_2}$

Birim Seçim Modelleri

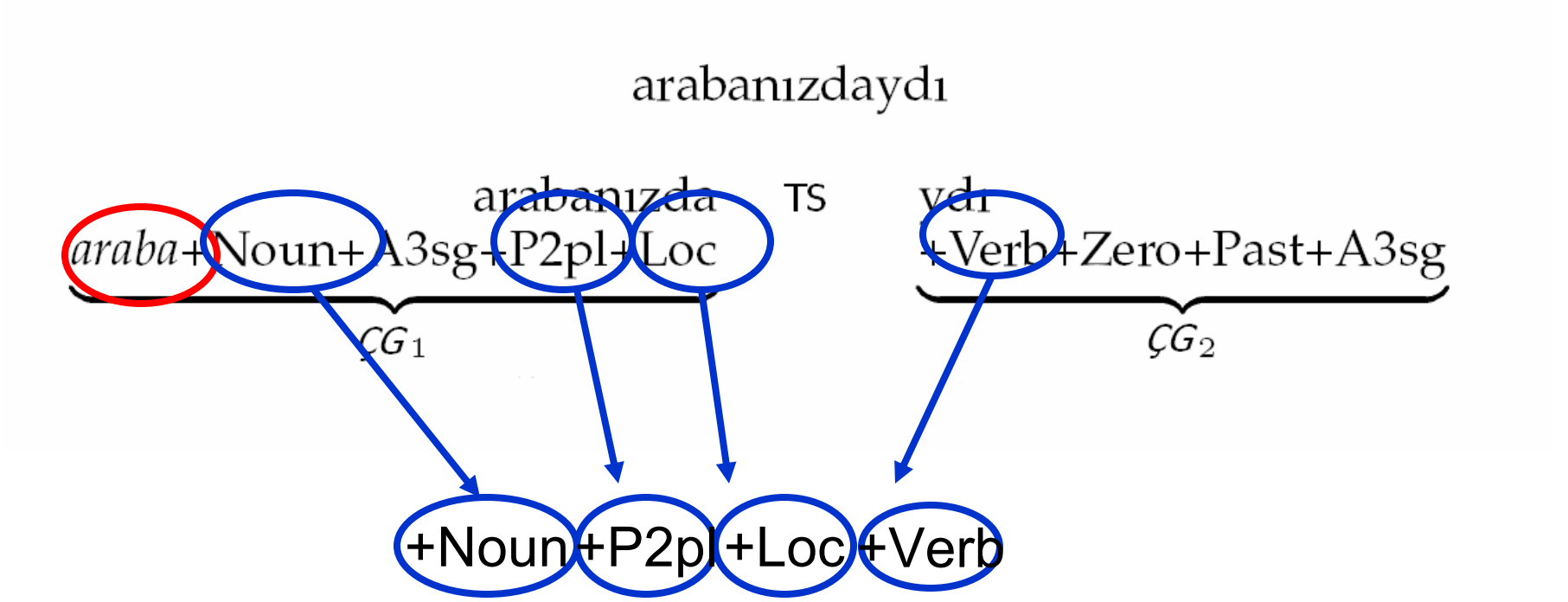
Sözcük Tabanlı Model



(Eryiğit and Oflazer (2006)'e benzer şekilde)

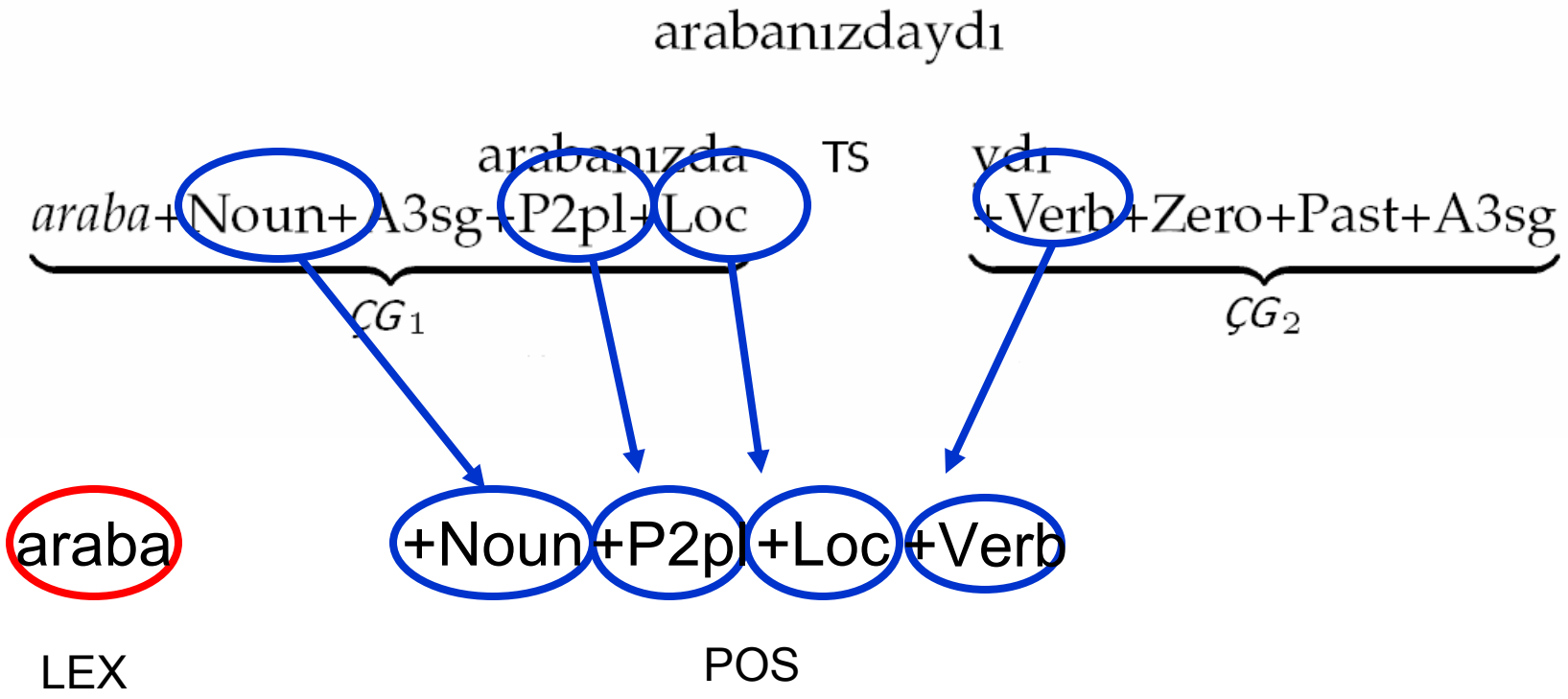
Birim Seçim Modelleri

Sözcük Tabanlı Model



Birim Seçim Modelleri

Sözcük Tabanlı Model



Birim Seçim Modelleri

ÇK Tabanlı Model

- Ayrıştırma Birimi : ÇK'ler
- Sözcük İçi bağlılıklar, gerçek bağlılıklar gibi KDM tarafından belirlenirler.

Birim Seçim Modelleri

ÇK Tabanlı Model

- Ayrıştırma Birimi : ÇK'ler
- Sözcük İçi bağlılıklar, gerçek bağlılıklar gibi KDM tarafından belirlenirler.

ÇK Tabanlı Belirlenimci Model

- Sözcük İçi bağlılıklar, KDM'ye başvurulmadan belirlenimci bir şekilde işlenirler.

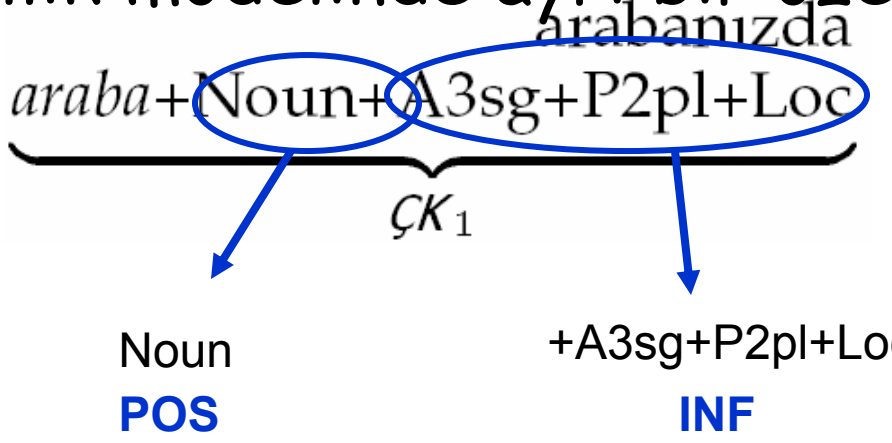
Birim Seçim Modelleri

Model	Görünüm Bilgisi Eklenmemiş		Görünüm Bilgisi Eklenmiş	
	ÇKB	$ÇKB_E$	*ÇKB*	$ÇKB_E$
Sözcük tabanlı	$67{,}2_{\pm 0{,}3}$	$57{,}9_{\pm 0{,}3}$	$70{,}7_{\pm 0{,}3}$	$62{,}0_{\pm 0{,}3}$
ÇK tabanlı	$68{,}3_{\pm 0{,}2}$	$58{,}2_{\pm 0{,}2}$	$73{,}8_{\pm 0{,}2}$	$64{,}9_{\pm 0{,}3}$
ÇK tabanlı belirlenimci	$70{,}6_{\pm 0{,}3}$	$60{,}9_{\pm 0{,}3}$	$73{,}8_{\pm 0{,}2}$	$64{,}9_{\pm 0{,}3}$

Biçimbirimsel Özelliklerin Kullanımı

ÇK tabanlı (INF birleşik) model

- Biçimbirimsel özellikler üzerinde indirgeme yok
- Tüm biçimbirimsel özellikler kullanımda
  - Ana sözcük sınıfına ek olarak ve
  - Özellik modelinde ayrı bir özellik olarak



Biçimbirimsel Özelliklerin Kullanımı

ÇK tabanlı (INF birleşik)

+A3sg+P2pl+Loc

ÇK tabanlı (INF parçalı): her bir parçacık ayrı bir özellik olarak kullanılır.

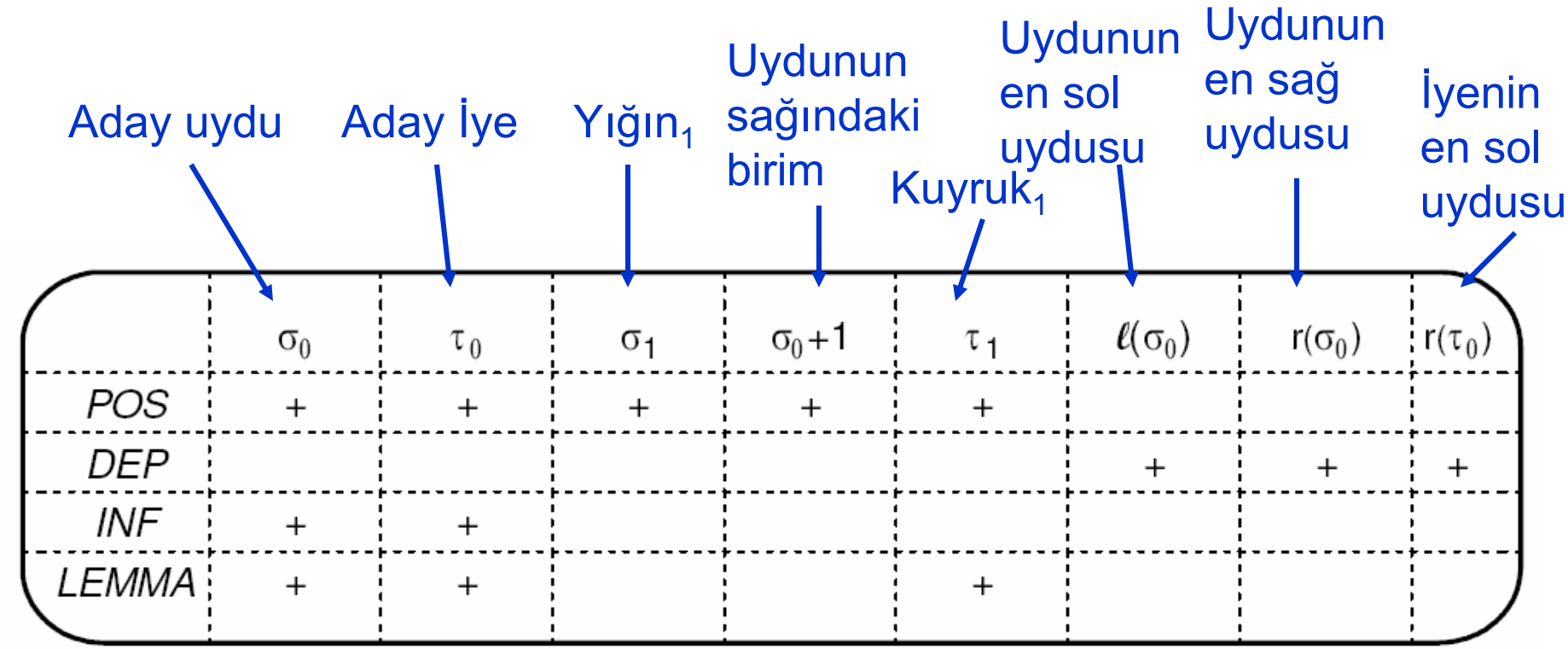
+A3sg +P2pl +Loc

Sınıflandırıcı Tabanlı Ayrıştırıcı
Deney Sonuçları – Tüm Derlem

Model	Görünüm Bilgisi Eklenmemiş		Görünüm Bilgisi Eklenmiş	
	ÇKB	$ÇKB_E$	*ÇKB*	$ÇKB_E$
Sözcük tabanlı	67,2±0,3	57,9±0,3	70,7±0,3	62,0±0,3
ÇK tabanlı	68,3±0,2	58,2±0,2	73,8±0,2	64,9±0,3
ÇK tabanlı belirlenimci	70,6±0,3	60,9±0,3	73,8±0,2	64,9±0,3
ÇK tabanlı (INF birleşik)	71,6±0,2	62,0±0,3	74,4±0,2	65,6±0,3
ÇK tabanlı (INF parçalı)	71,9±0,2	62,6±0,3	74,8±0,2	66,0±0,3
Eniyileştirilmiş			76,0±0,2	67,0±0,3

Özellik Kalıbı

En yüksek başarımların elde edildiği özellik kalıbı:



	σ_0	τ_0	σ_1	σ_0+1	τ_1	$\ell(\sigma_0)$	$r(\sigma_0)$	$r(\tau_0)$
POS	+	+	+	+	+			
DEP						+	+	+
INF	+	+						
LEMMA	+	+			+			

Ayrıştırıcı Başarımları

Ayrıştırıcı	*ÇKB*	
	KsmSb Derlem	*Tüm Derlem*
Yandakine bağlan (ilk *ÇK*)	63,9	56,0
Yandakine bağlan (son *ÇK*)	62,2	54,1
Kural tabanlı	73,4	70,5
Olasılık tabanlı	74,9±0,3	-
Sınıflandırıcı tabanlı	78,3±0,3	76,0±0,2

CoNLL-X Ortak Çalışması

- CoNLL-X (Conference on Natural Language Learning) Shared Task on Multi-lingual Dependency Parsing, Haziran 2006, New York
- 17 araştırma grubu
- 14 farklı dil

  Arapça, Çince, Çekçe, Danca, Macarca, Felemenkçe, Almanca, Japonca, Portekizce, Slovakça, İspanyolca, İsveççe, Türkçe, Bulgarca
- CoNLL-X veri biçimi, derlem dönüşümleri
- Başarım ölçütü $ÇKB_E$
- Türkçe için en yüksek başarım

CoNLL-X Ortak Çalışması

- Türkçe derlem, ortak çalışmanın en zor derlemi olarak gösterilmiştir. (Buchholz ve Marsi, 2006)
- Sekiz farklı türden metin, 25 farklı bağlılık türü
- Sınama verisinde yeni sözcük görülme oranı en yüksek dil
- Başarımlar %37.8 - %65.7 arasında

CoNLL-X Ortak Çalışması

Türkçe Bölümü

Katılımcılar	*ÇKB*	$ÇKB_E$
Sınıflandırıcı Tabanlı Ayrıştırıcı	**75,8**	**65,7**
Johansson ve Nugues	73,6	63,4
McDonald ve diğ.	74,7	63,2
Corston-Oliver ve Aue	73,1	61,7
Cheng ve diğ.	74,5	61,2
Chang ve diğ.	73,2	60,5
Yüret	71,5	60,3
Riedel ve diğ.	74,1	58,6
Carreras ve diğ.	70,1	58,1
Wu ve diğ.	69,3	55,1
Shimizu	68,8	54,2
Bick	65,5	53,9
Canisius ve diğ.	64,2	51,1
Schiehlen ve Spranger	61,6	49,8
Dreyer ve diğ.	60,5	46,1
Liu ve diğ.	56,9	41,7
Attardi	65,3	37,8

Sonuçlar

- Türkçe'nin bağlılık ayrıştırması konusunda literatürdeki en yüksek sonuçlar elde edilmiştir.
- Ayrıştırmada ana birim olarak sözcükler yerine ÇK'lerin kullanılmasının başarımı arttırdığı,
- Biçimbirimsel özelliklerin kullanılmasının Türkçe'nin ayrıştırmasında vazgeçilemez bir yere sahip olduğu gösterilmiştir.
- Görünüm bilgisi özelliklerini kullanmanın, Türkçe'nin bağlılık ayrıştırması başarımında önemli artışa neden olduğu gösterilmiştir.

Araçlar

- Türkçe Derlem, *Kemal Oflazer, Bilge Say, Nart Atalay*
- Biçimbirimsel Çözümleyici, *Kemal Oflazer*
- Sözcük Etiketleyici, *Deniz Yüret*
- Maltparser sınıflandırıcı tabanlı ayrıştırıcı platformu, *Joakim Nivre ve ekibi*
- LibSVM, *C.W. Hsu, C.C. Chang, C.J. Lin*

Referanslar

- Nivre, J., Hall, J., Nilsson, J., Chanev, A., **Eryiğit, G.**, Kübler, S., Marinov, S., and Marsi, E., 2007. ***MaltParser: A Language-Independent System for Data-Driven Dependency Parsing***, Natural Language Engineering Journal 13(1), 1-41 Cambridge Press.

- **Eryiğit, G.**, and Oflazer, K., 2006. ***Statistical dependency parsing of Turkish***. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, April.
- Nivre, J., Hall, J., Nilsson, J., **Eryiğit, G.** and Marinov, S., 2006. ***Labeled Pseudo-Projective Dependency Parsing with Support Vector Machines***. Proceedings of the Tenth Conference on Computational Natural Language Learning, New York, USA, June.
- **Eryiğit, G.**, Adalı, E. and Oflazer, K., 2006. ***Türkçe Cümlelerin Kural Tabanlı Bağlılık Analizi***. In Proceedings of the 15th Turkish Symposium on Artificial Intelligence and Neural Networks, Muğla, Turkey, June.
- **Eryiğit, G.**, Nivre, J. and Oflazer, K., 2006. ***The incremental use of morphological information and lexicalization in data-driven dependency parsing***, Proceedings of the 21st International Conference on the Computer Processing of Oriental Languages, Sentosa, Singapore, December.