



# Türkçe Haberlerde Yeni Olay Bulma ve İzleme: Bir Deney Derleminin Oluşturulması

Fazlı Can, Seyit Koçberber, Özgür Bağlıoğlu,  
Süleyman Kardaş, H. Çağdaş Öcalan, Erkan Uyar

Bilkent Bilgi Erişim Grubu  
Bilgisayar Mühendisliği  
Bilkent Üniversitesi



# Konu Başlıkları

- Giriş
  - Yeni Olay Belirleme ve İzleme Sistemi
  - Deney Derleminin Önemi
- İlgili Çalışmalar
  - TDT
- Deney Derleminin Hazırlanması
  - Bilginin elde edilmesi
  - Geliştirilen uygulama (ETracker)
  - Koleksiyon hakkında istatistiksel bilgiler
- Sonuç



# Giriş

- Olay
  - Belli bir **yer** ve **zaman** bilgisi içeren özel olgu.
- Konu
  - Yer, zaman bilgisi içermeyen genel olgu.
- Örnek
  - “British Telecom’ da bilgisayar virüsüne rastlandı, 3 Mart 1993” olay olarak tanımlanabilir.
  - “Bilgisayar virüsü salgını” konu olarak tanımlanabilir.



# Giriş

- Yeni olay belirleme: Haber akışı içerisinde daha önceden görülmemiş hikayelerin belirlenmesidir.
  - Uçağın düşmesi, deprem, seçimler, vb.
- Yeni olayın belirlenmesinde kullanılan nitelikler,
  - Olayın zamanı
  - Olaydaki kişiler
  - Olayın gerçekleştiği yer
  - ...



# Giriş

- Yeni olay belirleme ve izleme sistemlerinin kullanım alanları;
  - Sınıflandırma sistemleri
  - En güncel bilgilere ihtiyaç duyan iş alanları
    - İstihbaratçılar, Finansal analizciler, Borsacılar
  - Elektronik postaların olaylara ayrıştırılması ve takibi
- Bilgi süzme sistemlerinden farkı

<b>Bilgi süzme sistemi</b>	<b>Yeni olay belirleme ve izleme sistemi</b>
Kullanıcı tarafından belirlenen uzun süreli profil kullanır ve bilgi akışı içerisindeki ilgili belgeleri bu bilgiye dayanarak belirler.	Bilgi akışı içerisindeki olayları önceye dayalı hiçbir bilgi kullanmadan belirler ve takip eden ilgili belgeleri bulur.



# Giriş

- Deney derlemi YOBİ sistemlerinde kullanılan algoritmaların etkinliklerini ölçmeye ve literatürdeki diğer çalışmalarla karşılaştırmaya olanak sağlamaktadır.
- Standart deney derlemlerinin yapılan araştırmaların düzeyini yükseltici olumlu etkisi literatürde kanıtlanmıştır. [Voorhees, 2005]
- Türkçe'de ilk olması açısından bu konuda çalışma yapacak araştırmacılara altyapı sağlamaktadır.



# İlgili Çalışmalar

- **Konu Belirleme ve İzleme**  
(Topic Detection and Tracking)
  - 1997 – 2004 yılları arasında gerçekleştirilmiştir.
  - Çeşitli dillerde yazılı ve sözlü olarak yayımlanan haber hikayeleri üzerinde çalışılmıştır.
  - Araştırma alanları 5'e ayrılır;
    - Hikaye Bölümleme (Story Segmentation)
    - Konu İzleme (Topic Tracking)
    - Konu Belirleme (Topic Detection)
    - İlk Hikaye Belirleme (First Story Detection)
    - Bağlantı Belirleme (Link Detection)



# İlgili Çalışmalar

- Konu Belirleme ve İzleme altında geliştirilen deney derlemleri

	Hikaye Sayısı	İngilizce Konu Sayısı	Zaman Aralığı
TDT 1	<b>15,836</b>	<b>25</b>	Temmuz 1994 - Haziran 1995
TDT 2	<b>70,000</b>	<b>100</b>	Ocak 1998 - Haziran 1998
TDT 3	<b>35,000</b>	<b>115</b>	Ekim 1998 - Aralık 1998
TDT 4	<b>28,500</b>	<b>70</b>	Ekim 2000 - Ocak 2001
TDT 5	<b>278,109</b>	<b>63</b>	N/A
TDT-BilCol	<b>209,305*</b>	<b>80*</b>	Ocak 2005 - Aralık 2005

\* Türkçe



# Deney Derleminin Hazırlanması

- Haberler 5 kaynaktan toplanmıştır;
  - CNN Türk (<http://www.cnnturk.com>),
  - Haber 7 (<http://www.haber7.com>),
  - Milliyet Gazetesi (<http://www.milliyet.com.tr>)
  - TRT (<http://www.trt.net.tr>),
  - Zaman Gazetesi (<http://www.zaman.com.tr>).
- 2005 yılının zaman bilgisi (tarih, saat) içeren haberleri alınmıştır.



# Deney Derleminin Hazırlanması

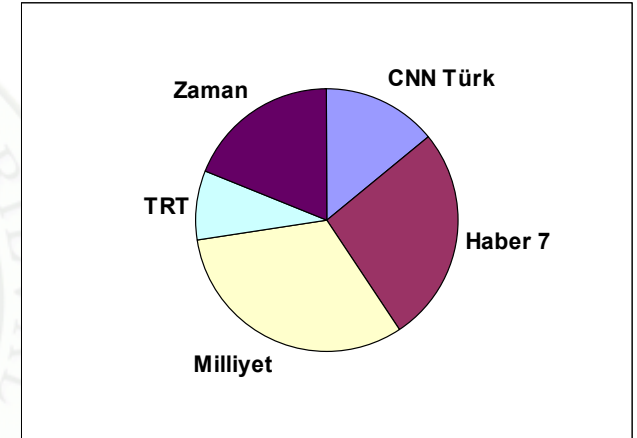
- Her kaynak farklı bakış açılarının temsilcisi olarak seçilmiştir
  - CNN Türk – Uluslararası, Amerikan tarzı
  - TRT – Hükümeti temsilen, Daha sınırlı
  - Milliyet Gazetesi – Modern bakış açısı
  - Zaman Gazetesi – Daha tutucu
  - Haber 7 – Çeşitlilik sağlamak amaçlı



# Deney Derleminin Hazırlanması

- Kaynaklar hakkında istatistikler

Haber Kaynağı	Haber Sayısı	Tüm Haberlere % Katkısı	Ortalama haber uzunluğu (kelime sayısı)
CNN Türk	23.644	11,3	271
Haber 7	51.908	24,8	238
Milliyet Gazetesi	72.233	34,5	218
TRT	18.990	9,1	121
Zaman Gazetesi	42.530	20,3	97
Toplam	209.305	100	200



- Haberler toplandıktan sonra, html dosyalarından etiketleri ayıklanmıştır.



# Deney Derleminin Hazırlanması

## ETracker

- ETracker, YOBI'de deney ve eğitim verilerini toplamak amacıyla geliştirilmiş, yarı-özdevimsel bir uygulamadır.
- Bilgi Erişim sistemini temel almaktadır. (f5-mf8)
- Kullanıcıların haber profilleri yapmalarına ve profilleri takip eden haberleri bulmalarına olanak sağlamaktadır.
- Ek olarak sistem, profiller hakkında bazı istatistiksel bilgileri de vermektedir.



# Deney Derleminin Hazırlanması

## ETracker

### • Olay profilinin içeriği;

- **Başlık (Topic Title):** Olayı çağrıştıracak ve kolayca akılda kalan on kelimedenden az bir cümle ya da kelimeler grubu,
- **Olay Tanımı (Event Summary):** Haber başlığını ayrıntılı hale getiren 1-2 cümle ile olayın tanımı,
- **Ne (What):** Olay sırasında ne olduğu,
- **Kim (Who):** Olayı gerçekleştiren veya olayda etkilenen kişiler,
- **Ne zaman (When):** Olayın gerçekleştiği zaman,
- **Nerede (Where):** Olayın gerçekleştiği yer,
- **Sayı (Topic Size):** Tahmini haber sayısı,
- **Tohum (Seed):** Konu ile ilgili ilk haber
- **Haber Türü (Event Type):** Haberin türü ya da türleridir.

**ETracker**

**Profile View**

<b>Topic Title:</b>	Koreli bilim adamının kök hücre araştırması sahte
<b>Event Summary:</b>	Koreli Profesör Hwang Woo-suk'un kök hücre araştırması sahte çıktı ve üniversitedeki görevinden istifa etti. Seul Ulusal Üniversitesi araştırma yaptı ve sahtekarlık yapıldığı sonucuna varıldığını açıkladı.
<b>What:</b>	Koreli bilim adamı Hwang Woo-suk'un kök hücre araştırması sahte
<b>Who:</b>	Koreli Profesör Hwang Woo-suk
<b>When:</b>	12 Aralık 2005
<b>Where:</b>	Kore
<b>Topic Size:</b>	30
<b>Seed ID:</b>	202176
<b>News Type:</b>	Bilim ve keşif haberleri
<b>Annotator:</b>	Fazlı Can



# Deney Derleminin Hazırlanması

## ETracker

- Olay profilinin içeriği;

**ETracker**

**Profile View**

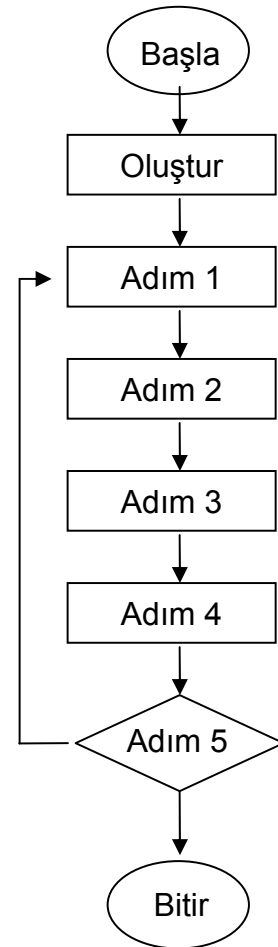
<b>Topic Title:</b>	Koreli bilim adamının kök hücre araştırması sahte
<b>Event Summary:</b>	Koreli Profesör Hwang Woo-suk'un kök hücre araştırması sahte çıktı ve üniversitedeki görevinden istifa etti. Seul Ulusal Üniversitesi araştırma yaptı ve sahtekarlık yapıldığı sonucuna varıldığını açıkladı.
<b>What:</b>	Koreli bilim adamı Hwang Woo-suk'un kök hücre araştırması sahte
<b>Who:</b>	Koreli Profesör Hwang Woo-suk
<b>When:</b>	12 Aralık 2005
<b>Where:</b>	Kore
<b>Topic Size:</b>	30
<b>Seed ID:</b>	<a href="#">202176</a>
<b>News Type:</b>	Bilim ve keşif haberleri
<b>Annotator:</b>	Fazlı Can



# Deney Derleminin Hazırlanması

## ETracker

- İzleyen haberlerin 5 adımda belirlenmesi;
  - Adım 1: Tohum ile arama
  - Adım 2: Profil ile arama
  - Adım 3: İlgili haberleri kullanarak arama
  - Adım 4: Yaratıcı sorgular ile arama
  - Adım 5: Kalite kontrol
- Adım 5'de, sistem yöneticileri profil kalitesini denetler; ya adımları tekrarlatır ya da profili onaylar.





# Deney Derleminin Hazırlanması

## ETracker

- Bu adımlarda, değerlendiriciler (annotator) haberleri “evet”, “hayır”, “belirsiz” ya da “bakılmadı” olarak etiketleyebilir.
- Değerlendiriciler;
  - Adım 1’de 200 haber,
  - Adım 2’de 300 haber
  - Adım 3’de 400 haber,
  - Adım 4’de ise her sorgu için 200 haber değerlendirir.



# Deney Derlemi Hakkında İstatistikler

- Şu ana kadar öğrencilerin ve akademisyenlerin katkılarıyla (~40 kişi) yaklaşık 80 profil tamamlanmıştır.

	Retireved	Tracking	Non-Tracking	Not-Sure	Not-Evaluated	Time-Spend
<b>Avg.</b>	546	89	378	1	77	130
<b>Min.</b>	221	2	14	0	0	20
<b>Max.</b>	1129	454	761	37	614	825

- Toplanan profillerin yanlış olmasını engellemek için ayrıca önem gösterilmiştir. Kişilerin yarattığı profil sayıları, haber türleri dengede tutulmaya çalışılmıştır.



# Deney Derlemi Hakkında İstatistikler

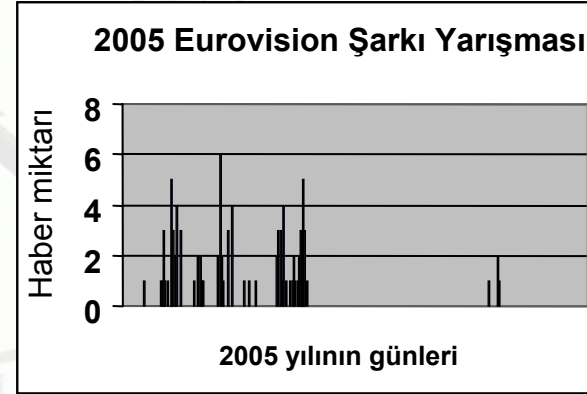
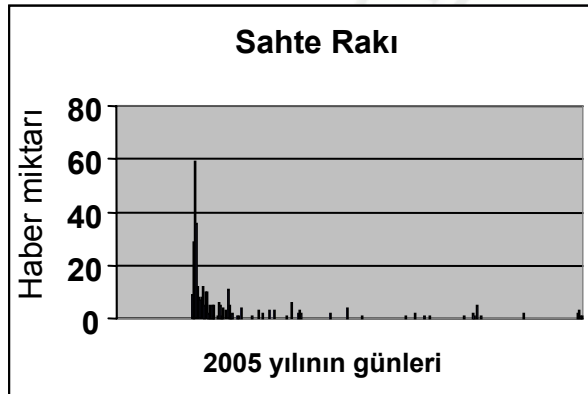
- Örnek profiller ve profillerin ömür istatistikleri

Haber Profillerinin Bazıları	İzleyen Haber Sayısı	Haber Ömrü (Gün)	İlk n Günde İzleyen Haber Sayısı			
			n=100	n=50	n=25	n=10
Londra metrosunda patlama	454	175	440	419	376	236
Sahte rakı	323	182	316	291	255	197
400 koyun intihar etti	10	8	10	10	10	10
Mortgage Türkiye'de	375	356	60	41	25	13
Onur Air'in Avrupa'da yasaklanması	159	203	154	154	148	105
İlk yüz nakli	14	17	14	14	14	10
Attilâ İlhan vefat etti	40	69	40	37	36	32
Su anda Toplam Profil Sayısı : 80	Ortalamalar :	72	64	54	47	36



# Deney Derlemi Hakkında İstatistikler

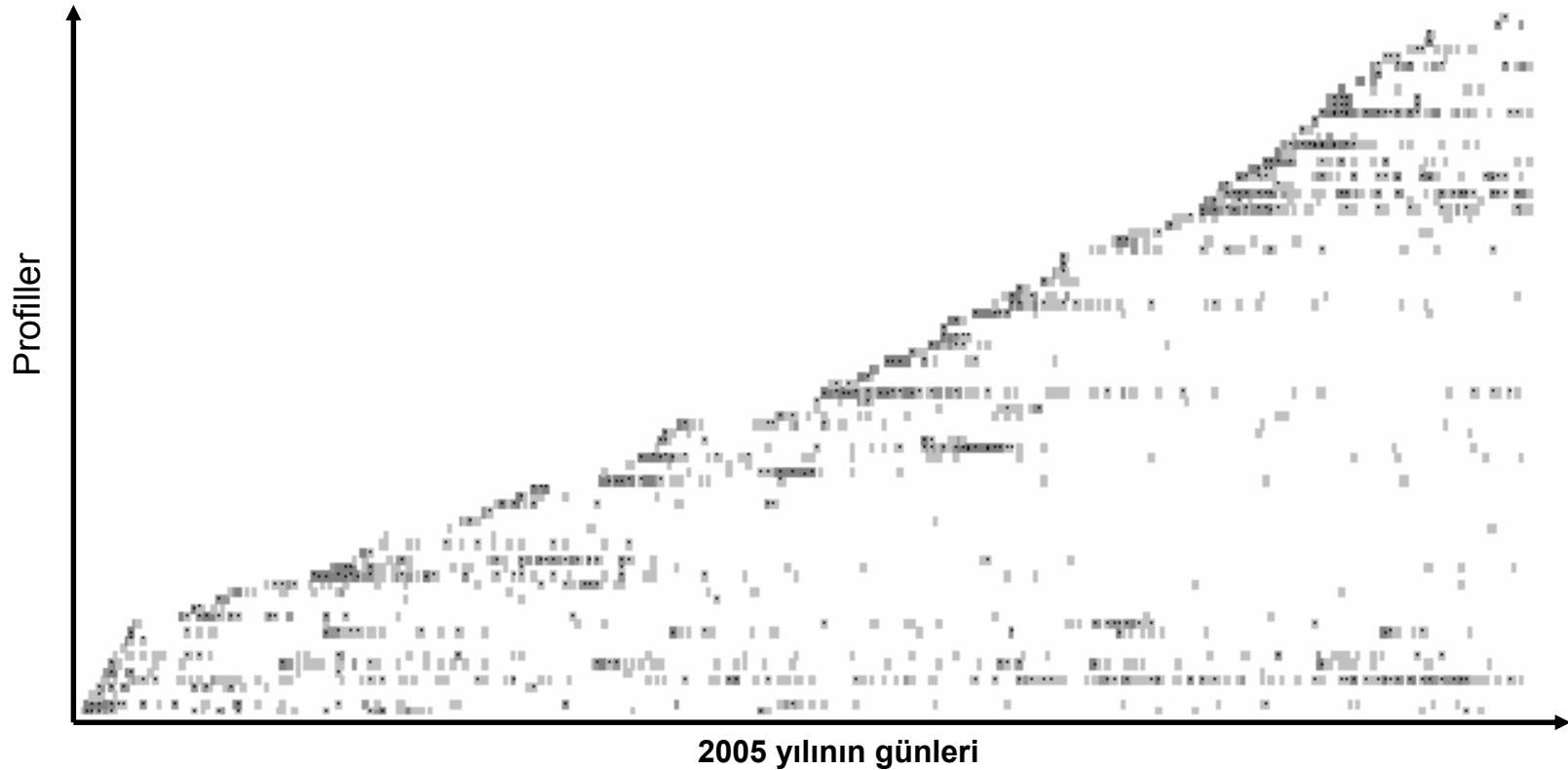
- ETracker kullanılarak tamamlanan iki profil için haberlerin yıl içindeki dağılımı





# Deney Derlemi Hakkında İstatistikler

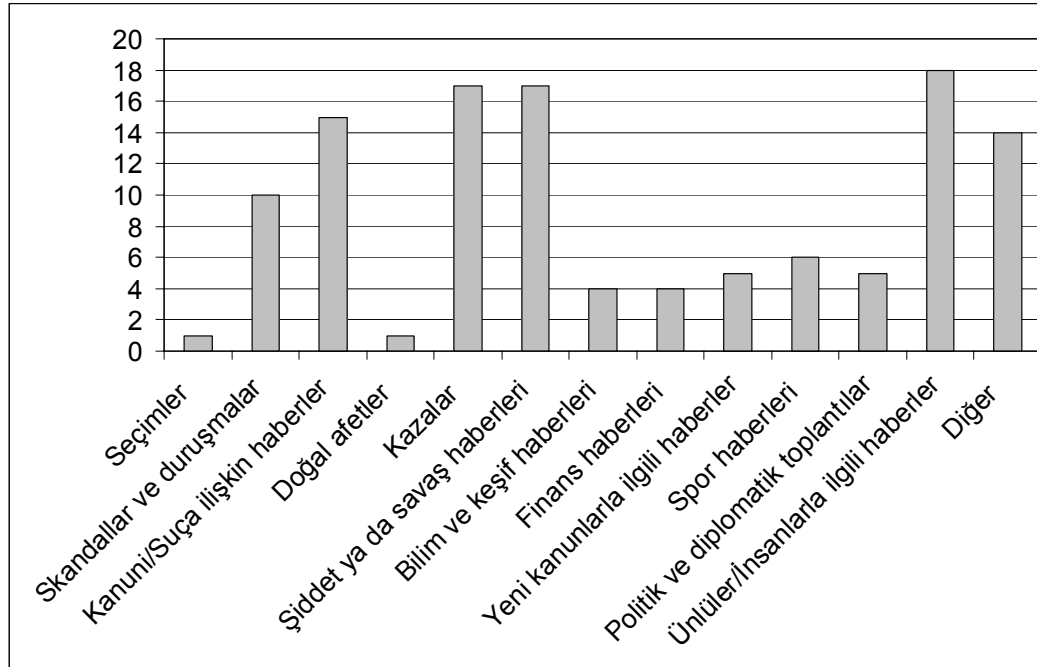
- Bütün profiller için haberlerin yıl içindeki dağılımı





# Deney Derlemi Hakkında İstatistikler

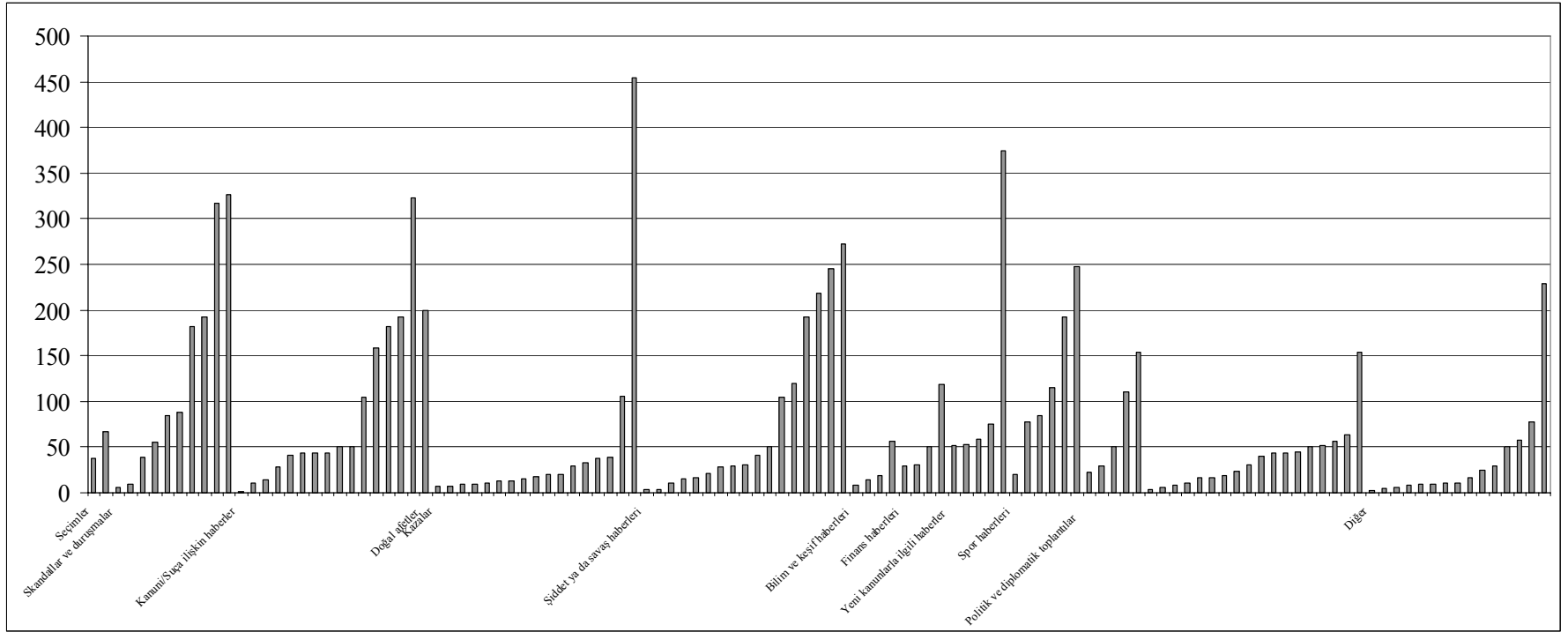
- Profillerin haber türlerine göre dağılımları





# Deney Derlemi Hakkında İstatistikler

## ● Profillerin haber türlerine göre dağılımları





# Sonuç

- Geliştirmekte olduğumuz deney derlemi büyüklüğü ve kapsamıyla Türkçe için yapılacak olan YOBİ çalışmaları için ilk standart olması amacıyla hazırlanmıştır.
- TDT 2004'den [TDT, 2004] esinlenerek geliştirilen derlem oluşturma yaklaşımı YOBİ için verimli ve etkin bir yöntemdir.
- Elde edilen sonuçlar Bilkent Haber Portalının tasarımında kullanılmaktadır.



# Referanslar

- Allan, J., Lavrenko, V., Jin, H. (2000). First story detection in TDT is hard. In Proceedings of the 9th Conference of Information and Knowledge Management (ACM CIKM'00) (ss. 374-381). Washington, DC: ACM.
- Allan, J., Papka, R., Lavrenko, V. (1998). On-line new event detection and tracking. In Proceedings of the 21st International Conference on Research and Development in Information Retrieval (ACM SIGIR '98) (ss. 37-45). Melbourne: ACM.
- Can, F, Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H. C., Vursavas, O. M. (2006a). First large scale information retrieval experiments on Turkish texts. (Poster makale) Proceedings of the 29th International Conference on Research and Development in Information Retrieval (ACM SIGIR '06) içinde (ss. 627-628). Seattle: ACM.
- TDT (2004). TDT 2004: Annotation manual: Version 1.2. August 4, 2004. Jan 9 2007 <http://projects ldc.upenn.edu/TDT5/Annotation/TDT2004V1.2.pdf>
- Voorhees, E. (2005). TREC: Improving information access through evaluation. Bulletin of the American Society for Information Science and Technology, 32.



# Sorular?

Sabrınız için teşekkürler...