Building a National Corpus of Turkish
Building a National Corpus of Turkish

- Information on *Turkish National Corpus* Project
- Design of Turkish National Corpus
  - Construction of an Electronic Database
    - Selection and sampling procedures of texts
    - Computerizing data
- Annotation
  - Parts of Speech Tagging
    - Developing parts of speech software
- Developing the corpus interface
  - Interface features
  - Releases of the corpus
Information on Turkish National Corpus Project

Funded by the Scientific and Technological Research Council of Turkey for a period of three years (2008-2011)

Size: 50 million words

Time Period: 1990-2008

Annotation: Parts of speech tagged

Medium: 95% writing, 5% speech

User friendly web-based interface

Following the BNC model
Building a National Corpus of Turkish

**Turkish National Corpus will be**

- **a general corpus**: it will not be specifically restricted to any particular subject field, register or genre
- **a mixed corpus**: it will contain examples of both written and spoken language
- **a monolingual corpus**: it will consist of text samples which are produced by Turkish language speakers
- **a sample corpus**: it will be composed of text samples generally no longer than 45,000 words
- **a synchronic corpus**: it will include imaginative and informative texts from 1990
Building a National Corpus of Turkish

1. Work Package I
   CONSTRUCTING ELECTRONIC DATABASE

   Compiling Data
     - Sampling Procedure
     - Selection Procedure
     - Data Sources

   Computerizing Data
     - Obtaining Electronic Sources
     - Scanning
     - Typing
     - Computerizing Spoken Data

   Checking Data
     - Checking Formats
     - Checking Orthography
     - Checking Text Information
     - Metadata Headers

TIME
Building a National Corpus of Turkish

Work Package II

ANNOTATION: POS TAGGING

Constructing Electronic Lexicon
- Roots
- Suffixes

POSS Tagging Software
- Identifying Parts of Speech of Words
- Tagging Parts of Speech of Words

Training of POS Tagging Software
- Automatic Control
- Manuel Control
Building a National Corpus of Turkish
Building a National Corpus of Turkish

Work Package IV

TNC PRERELEASE CONTROL

- Beta Release I
- Local Testing
- Beta Release II
- National Testing
- Candidate Release I
- International Testing
- Candidate Release II
Building a National Corpus of Turkish

Time / Workflow

1. Work Package I: Constructing Electronic Database
   - Compiling Data
     - Sampling Procedure
     - Selection Procedure
   - Computerizing Data
     - Obtaining Electronic Sources
     - Scanning
     - Typing
     - Computerizing Spoken Data
   - Checking Data
     - Checking Formats
     - Checking Orthography
     - Checking Text Information (Metadata/Entities)

2. Work Package II: Annotation: POS Tagging
   - Constructing Electronic Lexicon
     - Roots
     - Suffixes
   - POS Tagging Software
     - Identifying Parts of Speech of Words
     - Tagging Parts of Speech of Words
   - Training of POS Tagging Software
     - Automatic Control
     - Manuel Control

3. Work Package III: Developing the Interface
   - Query Criteria
   - Visual Design
   - Cross-Platform Compatibility

4. TNC PreRelease
Corpus Representativeness

Balance:
The range of genres included in a corpus

Intended uses of a corpus
Covering a wide range of text categories which are representative of a language
Sampling text categories proportionally
No scientific measure for corpus balance
Adopting an existing corpus model when building your corpus

Sampling:
How text chunks for each genre are selected

View of a corpus:
Static or dynamic language model
Selection Features

Written data selection

Written texts included in the database:

**Domain:** Subject matter

- **Imaginative:** Fiction
- **Informative:** Social science, art, commerce-finance, belief-thought, world affairs, applied science, natural-pure science, leisure

**Time:** 1990-2008

**Medium of text:** Books, periodicals, miscellaneous (published-unpublished), written to be spoken
Building a National Corpus of Turkish

Construction of an Electronic Database II

Selection Features

Sampling

Sampling population for a general-purpose corpus

Sampling from produced and received language population

Sampling frame

Information on published material

Catalogues of books, books in print lists, best seller lists, particularly prize winners of competitions (Yunus Nadi, Orhan Kemal Roman, Haldun Taner short story prizes), library lending statistics, lists of current magazines and periodicals, periodical circulation figures
Selection Features

Sampling procedure

- For books: a target sample size of 45,000 words will be chosen.

- A convenient break point will be chosen for text samples.

- Samples will be taken randomly from the beginning, middle or end of longer texts.

- Individual texts that newspapers and magazines contain will be separated and classified on the basis of the selection and classification features.
Selection Features

- Books from bestsellers, prize winners of competitions lists, books in print lists:
  
  Publication date: 2000-2007

- Periodicals
  
  National newspapers:
  Radikal, Milliyet, Türkiye, Cumhuriyet, Zaman
  Local newspapers:
  Bugün Mersin, Bolu Detay
  Magazines:
  Birikim, Sızıntı, Aksiyon, Aktüel, Telapati

Unpublished written material

Student essays (Primary, secondary, and higher education)
Title: Excerpt from Benim adım kırmızı. Sample containing about 38.764 words

Spoken or written: Written

Number of words: 38.764

Derived text type: Fiction

Text type: Fiction: Prose

Publication date: 1998

Age of author: 46

Sex of author: Male

Type of author: Sole

Age of audience: Adult
Text domain: Imaginative: Fiction

Medium of text: Book

Text sample: Middle sample

Name of author: Orhan Pamuk

Name of text: Benim Adım Kırmızı

Target audience sex: Mixed

Publisher: İletişim

Place of publication: İstanbul

Key words: History, crime and adventure
<table>
<thead>
<tr>
<th><strong>Title:</strong></th>
<th>Art news from Radikal. Sample containing about 25.184 words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spoken or written:</strong></td>
<td>Written</td>
</tr>
<tr>
<td><strong>Number of words:</strong></td>
<td>25.184</td>
</tr>
<tr>
<td><strong>Derived text type:</strong></td>
<td>Newspaper National</td>
</tr>
<tr>
<td><strong>Text type:</strong></td>
<td>Written News</td>
</tr>
<tr>
<td><strong>Publication date:</strong></td>
<td>2005</td>
</tr>
<tr>
<td><strong>Age of author:</strong></td>
<td>Unknown</td>
</tr>
<tr>
<td><strong>Sex of author:</strong></td>
<td>Mixed</td>
</tr>
<tr>
<td><strong>Type of author:</strong></td>
<td>Multiple</td>
</tr>
<tr>
<td><strong>Age of audience:</strong></td>
<td>Adult</td>
</tr>
</tbody>
</table>
Construction of an Electronic Database IX

Text domain: Informative: Arts

Medium of text: Periodical

Text sample: Composite

Target audience sex: Mixed

Publisher: Radikal

Place of publication: Istanbul
Spoken data selection procedure

Spoken data included in the database:

- Demographic part of spoken database
- Transcriptions of spontaneous natural conversations:
  - Equal numbers of men and women recruits will use a personal stereo to record all their conversations over two to seven days, and logged details of each conversation in a notebook.
  - Permission will be obtained from the participants of the conversations upon each recording.
  - Information about the participants, such as age, sex, accent, occupation will be recorded when available.
Spoken data included in the database:

**Context-governed part of spoken database**

- **Educational-informative events**: Lectures, news broadcasts, classroom discussions
- **Institutional-public events**: Political speeches, parliamentary proceedings, council meetings
- **Leisure events**: Sport commentaries, club meetings, after-dinner speeches
Computerizing and checking data - I

**Re-use of existing electronic texts:**
Newspapers, magazines, journals, books, other materials

**Scanning:**
Scanning and hand editing of books

**Keyboarding:**
Typing of hand-written items, leaflets, recorded speech
Construction of an Electronic Database XIII

Computerizing and checking data - II

WORKFLOW

1. Scanning & Typing
   - Typing
   - Scanning
   - Annotation

2. OCR
   - Correction & SpellCheck

3. SVN 1
   - Final Check
   - SVN 2

PHASE 1

SVN 1 Layout
- Scanning & Typing
  - On Editing
  - FINAL VERSION
- Digital Data
  - On Editing
  - FINAL VERSION

PHASE 2

PHASE 3

SVN Accounts
- SVN 1
  - Undergraduate Students
  - Graduate Students
- SVN 2
  - Graduate Students

Job Management
- Undergraduate Students
- Graduate Students
MGK: Önce laiklik
AKP hükümeti ile Genelkurmay arasında Milli Görüş genelgesiyle başlayıp, TBMM resesiyonunun boykotu ile süren gerilimin gölgesinde toplanan Milli Güvenlik Kurulu (MGK), 'ince ve diplomatik' laiklik uyanısı ile son buldu.

'Postmodern darbe' olarak nitelenen 28 Şubat 1997 tarihli toplantısının ardından en uzun toplantısını yapan MGK, bildirisinde laikliği, Irak ve Kıbrıs gibi konuların önüne koydu.

7.5 saatlik toplantı

İrticayla mücadele, Irak, Kıbrıs, Petrol Boru Hatları gibi konuların ele alındığı toplantı 7.5 saat sürdü. Yemek arası vermeyen MGK üyesi sandviç atıtırarak toplantıya devam ettiler.

'Laiiklik titiz korunsun'
Toplantının ardından yayımlanan kısa bildiride şöyle denildi:
1- MGK 30 Nisan 2003 tarihinde aylık yapılan toplantısını yapmıştı.
2- Toplantıda ülkemizin güvenlik ve asayişini etkileyen iç ve dış gelişmeler gözden geçirilmiş. Bu kapsamda;
   a) Devletin temel niteliklerinden olan laiklik ilkesinin önemi ve titizlikle korunması vurgulanmıştı.
   b) Irak’a ilişkin gelişmeler ayrıntılı değerlendirilmiş gelişmelerin yakından izlenmeye devam edilmesi ve gerekli temaslardırını sürdürlmesi kararlaştırılmıştı.
   c) Kıbrıs’da ilgili tüm gelişmeler kapsamlı olarak gözden geçirilmişti.
   d) Ayrıca, petrol boru hatlarına ilişkin son gelişmeler hakkında kurula bilgi sunulmuş ve alınması gereken önlemler görülümüştü.

Toplantıda tartışmaların önemli bir bölümü, son günlerde devleter zirvesinde büyük gerilim yaratan Milli Görüş genelgesi, türban tartışmaları, kadrolaşma gibi konular tartışılarka geçti.
Annotation - I

Parts of Speech Tagging

Parts of Speech Tagger

Stage I
- Morphological Analyzer
  - Root List
  - Suffix List
  - Rule Set
- Manually Tagged Corpus
- Probabilistic Disambiguator

Stage II
- Statistical Testing
- Automatically Tagged Text
- Manual Editing
- Tagged Text

Corpus
Rule-based morphological analyser:
Construct an electronic lexicon of Turkish, run it on the corpus data (at least 10 million words), and obtain the most likely Turkish grammatical categories - such as nouns, verbs, adjectives, adverbs, conjunctions, etc.

Probabilistic analyser:
It will run on a corpus data containing manually tagged parts of speech annotation, and will turn the most probable word classes of Turkish.
Comparison:

Compare the tagged word classes of rule-based analyser to word classes of probabilistic analyser. To achieve precision in POS tagging software, human post-editing will be taken place on the outputs of software.
It is a dream come true for me she said I love you.
Developing the Corpus Interface I

Interface features

Query
Detailed specification of the *categories of text*, and/or *sections of texts* to constrain a search

Concordance output
Navigating through concordance output and manipulating the concordance

Text Information - Statistical information
Facility to access *metatextual information* and *frequency data*
## Restricted Range of Written Texts

**Query string:**

- Text input field

**Query mode:**

- Simple query (case insensitive)

**Number of hits per page:**

- 50

---

### Publication Date:

- 1960-1974
- 1975-1984
- 1985-1993

### Medium of Text:

- Book
- Periodical
- Miscellaneous: published
- Miscellaneous: unpublished
- To-be-spoken

### Text Sample:

- Whole text
- Beginning sample
- Middle sample
- End sample
- Composite

### Domain:

- Imaginative prose
- Informative: Natural and pure sciences
- Informative: Applied science
- Informative: Social science

### Derived text type:

- Academic prose
- Fiction and verse
- Non-academic prose and biography
- Newspapers
## Developing the Corpus Interface IV

### Visual Design

<table>
<thead>
<tr>
<th>Estimated Circulation Size:</th>
<th>Perceived Level of Difficulty:</th>
<th>Domicile of Author:</th>
</tr>
</thead>
<tbody>
<tr>
<td>□ Low</td>
<td>□ Low</td>
<td>□ UK and Ireland</td>
</tr>
<tr>
<td>□ Medium</td>
<td>□ Medium</td>
<td>□ Commonwealth</td>
</tr>
<tr>
<td>□ High</td>
<td>□ High</td>
<td>□ Continental Europe</td>
</tr>
<tr>
<td></td>
<td></td>
<td>□ USA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>□ Elsewhere</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age of Author:</th>
<th>Sex of Author:</th>
<th>Type of Author:</th>
</tr>
</thead>
<tbody>
<tr>
<td>□ 0-14</td>
<td>□ Male</td>
<td>□ Corporate</td>
</tr>
<tr>
<td>□ 15-24</td>
<td>□ Female</td>
<td>□ Multiple</td>
</tr>
<tr>
<td>□ 25-34</td>
<td>□ Mixed</td>
<td>□ Sole</td>
</tr>
<tr>
<td>□ 35-44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>□ 45-59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>□ 60+</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Target Audience Age:</th>
<th>Target Audience Sex:</th>
</tr>
</thead>
<tbody>
<tr>
<td>□ Child</td>
<td>□ Male</td>
</tr>
<tr>
<td>□ Teenager</td>
<td>□ Female</td>
</tr>
<tr>
<td>□ Adult</td>
<td>□ Mixed</td>
</tr>
<tr>
<td>□ Any</td>
<td></td>
</tr>
</tbody>
</table>
### Visual Design

<table>
<thead>
<tr>
<th>Genre:</th>
</tr>
</thead>
<tbody>
<tr>
<td>W:ac:humanities_arts</td>
</tr>
<tr>
<td>W:ac:medicine</td>
</tr>
<tr>
<td>W:ac:nat_science</td>
</tr>
<tr>
<td>W:ac:polit_law_edu</td>
</tr>
<tr>
<td>W:ac:soc_science</td>
</tr>
<tr>
<td>W:ac:tech_engin</td>
</tr>
<tr>
<td>W:admin</td>
</tr>
<tr>
<td>W:advert</td>
</tr>
<tr>
<td>W:biography</td>
</tr>
<tr>
<td>W:commerce</td>
</tr>
<tr>
<td>W:email</td>
</tr>
<tr>
<td>W:essay:school</td>
</tr>
<tr>
<td>W:essay:univ</td>
</tr>
<tr>
<td>W:fict:drama</td>
</tr>
<tr>
<td>W:fict:poetry</td>
</tr>
<tr>
<td>W:fict:prose</td>
</tr>
<tr>
<td>W:hansard</td>
</tr>
<tr>
<td>W:institut_doc</td>
</tr>
<tr>
<td>W:instructional</td>
</tr>
<tr>
<td>W:letters:personal</td>
</tr>
<tr>
<td>W:letters:prof</td>
</tr>
<tr>
<td>W:misc</td>
</tr>
<tr>
<td>W:news_script</td>
</tr>
<tr>
<td>W:newsp:brdsht_nat:arts</td>
</tr>
<tr>
<td>W:newsp:brdsht_nat:commerce</td>
</tr>
<tr>
<td>W:newsp:brdsht_nat:editorial</td>
</tr>
<tr>
<td>W:newsp:brdsht_nat:misc</td>
</tr>
<tr>
<td>W:newsp:brdsht_nat:report</td>
</tr>
<tr>
<td>W:newsp:brdsht_nat:science</td>
</tr>
<tr>
<td>W:newsp:brdsht_nat:social</td>
</tr>
<tr>
<td>W:newsp:brdsht_nat:sports</td>
</tr>
<tr>
<td>W:newsp:other:arts</td>
</tr>
<tr>
<td>W:newsp:other:commerce</td>
</tr>
<tr>
<td>W:newsp:other:report</td>
</tr>
<tr>
<td>W:newsp:other:science</td>
</tr>
<tr>
<td>W:newsp:other:social</td>
</tr>
<tr>
<td>W:newsp:other:sports</td>
</tr>
<tr>
<td>W:newsp:tabloid</td>
</tr>
<tr>
<td>W:non_ac:humanities_arts</td>
</tr>
<tr>
<td>W:non_ac:medicine</td>
</tr>
<tr>
<td>W:non_ac:nat_science</td>
</tr>
<tr>
<td>W:non_ac:polit_law_edu</td>
</tr>
<tr>
<td>W:non_ac:soc_science</td>
</tr>
<tr>
<td>W:non_ac:tech_engin</td>
</tr>
<tr>
<td>W:pop_lore</td>
</tr>
<tr>
<td>W:religion</td>
</tr>
</tbody>
</table>
Cross-platform compatibility

Character set:
Texts in the corpus data will be saved in UTF-8 character code.

Platform-free:
Corpus interface will be constructed in terms of w3c standards and it will be accessed through any web-browser on any platform.

Result set:
Results of the corpus will be used on any platform and they will be compatible with other softwares.
Prerelease of the Corpus

Releases

Beta Release I - Local Testing:
Bringing out the corpus on the local network for testing

Beta Release II - National Testing:
Upon local testing, revealing the corpus to national users and academicians

Release Candidate I - International Testing:
Upon national testing, making the corpus available to international users

Release Candidate II - Turkish National Corpus:
The final candidate release before the release of the corpus
And the volunteers

Working...
And the volunteers

Working...
And the volunteers

Working...
And the volunteers

Working...
Thank you