Turkish Coreference Annotation Manual*

Barış Gün Sürmeli Kübra Cıngıllı Ferit Tunçer Peter Schüller
Computer Engineering Department, Faculty of Engineering
Marmara University, Istanbul, Turkey

1.12.2016 (Version 2)

We say these parts of a text are **coreferent** because they **point to the same thing** in the world. Computers can learn how to find coreference automatically if we provide **training data**. Such training data contains text plus **manually created annotations** that show which parts of the text are coreferent. This is a guide for creating such coreference annotations for Turkish. As an example, a computer cannot easily know that "Atatürk", "Mustafa Kemal", and "Türkiye Cumhuriyeti'nin kurucusu" refers to the same person. Annotation principles are numbered in the margin as (Px).

Coreference annotation consists of two parts: (i) marking **mentions**: mentions are parts of the text that can be coreferent; afterwards (ii) storing all phrases that point to the same object or being into a **coreference chain**. In practice (i) requires to mark all phrases, and (ii) requires to select a set of mentions marked in (i) and storing them as a chain by giving it a name.

1 Which parts of the text are mentions that can be coreferent?

Noun phrases, pronouns, and nominalized adjectives are candidates for coreference. (P1)

- A noun phrase is a sequence of words that specifies a concrete entity, for example names and descriptions: "Atatürk", "arabam", "benim çok şirin ama nerdeyse her zaman bozuk olan ... arabam".
- Pronouns (zamirler) are "ben", "sen", "o", "biz", ..., "onun", ..., "onca", "odur", ...
- Nominalized adjectives (sıfat) are "yenisi", "eskisi", ...

2 Which coreferences should be annotated?

All cases where two phrases point to the same concrete entity. Also names that are used repeatedly ("Ankara", "Ankara'da", "Ankara") are annotated. (P2)

2.1 Nouns phrases with pronouns

The most basic coreference is between nouns and pronouns.

"Ahmet okula gitti. O, orayı çok sevdi."

Here the noun phrase "Ahmet" and the pronoun "O" should be marked as coreferent. Also "okul" and "orayı" should be marked as coreferent.

"Bizim araba eski, sizinkisi bizimkisinden yeni."

The noun phrase "Bizim araba" and the pronoun "bizimkisinden" are coreferent.

Important: we do not split "bizimkisinden", we mark the **complete word**.

2.2 Noun phrases with noun phrases

2.2.1 Proper and common nouns

If one noun phrase is a proper noun (name) and the other is a common noun and they clearly refer to same object or being, they are coreferent. (P3)

"Ahmet Hoca evlendi. Umarım çok mutlu olur, profesörümüz her şeyin en iyisini hak ediyor."

"Ahmet Hoca" and "profesörümüz" refers to same person.

2.2.2 Different proper nouns

If two different proper nouns (names) refer to same object or being, they are coreferent. (P4)

"Prof. Dr. Kemal Doe teşrif etti. Kemal Bey bize çalışmalarının gidişatından bahsetti."

*This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) Grant 114E430.

2.3 Nominalized adjectives and verbs

2.3.1 Nominalized Adjectives (adlaşmış sıfat)

"Yeni araba ile eski araba karşı karşıya geldi. Yenisi, eskisinden çok daha hızlıydı."

Here noun phrase "Yeni araba" is coreferent with nominialized adjective "yenisi" and noun phrase "eski araba" is coreferent with "eskisi".

2.3.2 Nominalized Verbal Adjectives (adlaşmış sıfat-fiil)

"Karşıdan gelen adam emin adımlarla yürüyordu. Mustafa: 'Geleni tanıyorum.' Dedi."

Here noun phrase "Karşıdan gelen adam" is coreferent with "Geleni".

2.4 Examples and Special Cases

"Mustafa kendine gel. Sen buralara gelmek için çok çalıştın. Sana bu salmışlık hali yakışmıyor." — "Ben de farkındayım. Düzeleceğim."

All underlined words refer to the same person.

Important: "farkındayım" and "Düzeleceğim" contain references to Mustafa (-yım and -im), BUT these words refer to what Mustafa is doing, they are NOT coreferent! (P5)

"Kristof Kolomb'un, öyle insanlık adına keşiflere çıkmış bir seyyah değil, yeni zenginlikler peşinde koşan ve tayfasına kan kusturan zalim bir kaptan olduğunu fark ediyorsunuz."

Here, the name and a long proper noun are coreferent.

Important: we take the largest possible coreferent mention (marking only "kaptan" would be wrong). (P6)

"... iktidar arayışıyla ilgili yazı yazma krizine girmişken dün akşamüstü icat ettim ... Başlığı bulunca rahatladım, krizden kurtuldum ..."

Here the first coreferent is a long noun phrase that is part of the subject.

"T. C. başbakanı Ankara'ya gitti, onun katılımıyla Cumhuriyetin 95'inci yıldönümü kutlamaları başladı."

A mention can contain another mention. In this example there are 2 chains and 4 mentions. (P7)

3 Which cases should not be annotated?

"Big bang bir söylencedir."

This is predication and not coreference: the purpose of the sentence is to say that "bir söylencedir" is the property of "Big bang". (P8)

"Boş, kiralık apartman dairesi, bir ev değildir; o, kiralanması beklenen bir konuttur."

"Boş, kiralık apartman dairesi" is not a specific apartment. There is not coreference with "o". (P9)

"Derin adamdı. Pek konuşmazdı."

The subjects of these sentences are coreferent, but they are not tokens, so we cannot annotate them. (P10)

"Arının peteği, arıdan arıya, kırlangıcın yuvası, kırlangıcından kırlangıcına değişmez, onlar yalnızca birer barınaktır ..."

Collections of objects are not considered coreferent. Therefore "onlar" is not annotated at all. (P11)

Annotations are done in CoNLL coreference format. A few examples are as follows.

The best way to edit such files is a text editor with fixed-width font (Word or libreoffice will **not** be helpful).

```
0    1   Ahmet          (1
0    2   Hoca           1)
0    3   evlendi         -
0    4   .               -
1    1   Umarım          -
1    2   çok             -
1    3   mutlu           -
1    4   olur            -
1    5   ,               -
1    6   profesörümüz   (2)
1    7   her_şeyin       -
1    8   en              -
1    9   iyisini         -
1   10   hak             -
1   11   ediyor          -
1   12   .               -
1=2 Ahmet Hoca
```

```
0   1   Karşıdan     (1
0   2   gelen         -
0   3   adam         1)
0   4   emin           -
0   5   adımlarla      -
0   6   yürüyordu      -
0   7   .              -
1   1   Mustafa        -
1   2   :              -
1   3   ‘              -
1   4   Geleni      (2)
1   5   tanıyorum      -
1   6   .              -
1   7   ’              -
1   8   Dedi           -
1   9   .              -
1=2 gelen adam
```

```
0   1   T.C.          (1(2)
0   2   başbakanı       1)
0   3   Ankara’ya       -
0   4   gitti           -
0   5   ,               -
0   6   onun          (3)
0   7   katılımıyla     -
0   8   Cumhuriyetin  (4)
0   9   95’inci         -
0  10   yıldönümü       -
0  11   kutlamaları     -
0  12   başladı         -
0  13   .               -
1=3 T.C. Başbakanı
2=4 Türk Cumhuriyeti
```