Turkish Coreference Annotation Manual*

Barış Gün Sürmeli Peter Schüller
Computer Engineering Department, Faculty of Engineering
Marmara University, Istanbul, Turkey

August 27, 2015 (Version 1)

A computer cannot easily know that "Atatürk", "Mustafa Kemal", and "Türkiye Cumhuriyeti'nin kurucusu" refers to the same person. 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.

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.

- A noun phrase is a connected sequence of words that specifies an entity.
  Examples are names and descriptions, they can be short or long: "Atatürk", "arabam", "benim çok şirin ama nerdeyse her zaman bozuk olan ... arabam".
- Pronouns (zamirler) and derived forms are "ben", "sen", "o", "biz", ..., "onun", ..., "onca", "odur", ...
- Nominalized adjectives (sıfat) are "yenisi", "eskisi", ...

2 Which coreferences should be annotated?

All cases where two mentions point to a clearly identifiable entity.

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** even though marking only "bizimkisi" can seem more logical.

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 refer to same object or being, they should be marked as coreferent.

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

"Ahmet Hoca" and "profesörümüz" refers to same person, the first is a proper noun and the second is a common noun.

2.2.2 Different proper nouns

If two different proper nouns (names) refer to same object or being, they should be marked as coreferent. This often happens when a person is called with full name or surname or with a title.

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

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

Here "Prof. Dr. Kemal Doe" and "Kemal Bey" should be marked as coreferent.

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". Both coreferences should be marked.

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" should be marked as 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 and should be stored into the same co-reference chain.

Important: even though "farkındayım" and "Düzeleceğim" contain references to Mustafa (-yım and -im), these words refer to what Mustafa is doing and not to the person and so they are not coreferent!

"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).

"... 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.

3 Which cases should not be annotated?

The following cases are not coreference and should not be annotated.

"Big bang bir söylencedir."

Here "Big bang" refers to an event and the sentence describes a property. This is not coreference, so "bir söylencedir" is not marked as coreferent with "Big bang".

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

Here, "Boş, kiralık apartman dairesi" and "o" is not coreferent because it is not clear which apartment flat is the object that "o" is pointing to. (It describes a set of possible flats.)

"Derin adamdı. Pek konuşmazdı."

Here there are two subjects that are coreferent. However they are not visible as words so we cannot annotate them.

"Bilim hayatın her alanında rehber alınmalıdır. Bilim ufkumuzu açmıştır, açacaktır. O geleceğe ışık tutmaktadır ..."

Here "bilim" is first used in a general objective manner, and then in a more specific subjective manner. Only the second "bilim" is coreferent with "O" and should be annotated.

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

Here "onlar" refers to a collection of objects in the first part of the sentence. This collection of objects cannot be marked as one phrase, so we do not annotate it. Therefore "onlar" is not annotated at all.