Resolution of Dropped Pronouns in a Phrase Structure Grammar

Meltem Turhan
mturhan@ceng.metu.edu.tr

Onur T. Şehitoğlu
onur@lcsl.metu.edu.tr

Dept.of Computer Engineering, Middle East Technical University
Ankara/Turkey

Abstract

In this work intra-sentential resolution of Turkish dropped pronouns in a phrase structure grammar is studied. Turkish is a pro-drop and free constituent order language. The resolution scheme for dropped pronouns depends on the constituent order. We introduce resolution rules for different surface orders and propose an implementation for an HPSG based parser. Implementation is based on incremental processing of non-local referential index sets during parsing.

1 Introduction

In this work, we propose a method for intra-sentential resolution of dropped pronouns in Turkish. Turkish is a pro-drop language where pronouns in the subject position of a sentence (and subordinate clauses including relative clauses, sentential complements and adverbs) and possessor position of a possessive noun group can be omitted. Agreement features (PERSON and NUMBER) of dropped pronouns are marked by the agreement suffix in the head constituent. However properties like reflexivity are not available resulting multiple binding possibilities.

The possible antecedent of a dropped pronoun is first constrained by the linear order of the antecedent and pronoun within the sentence. The antecedent should precede the NP with the dropped pronoun unless this is the first NP in the sentence.

(1) a. Adam$_i$ (*pro*$_i$) karısına$_j$ (*pro*$_{i/j}$) abisini anlattı.
man (his) wife-*3SP-Dat* (his/her) brother-*3SP-Acc* tell-*Tense*
"The man$_i$ told his$_i$ wife$_j$ about his$_i$/her$_j$ brother."

b. Adam$_i$ (*pro*$_i$) abisini$_j$ (*pro*$_{i/j}$) karısına anlattı.
man (his) brother-*3SP-Acc* (his) wife-*3SP-Dat* tell-*Tense*
"The man$_i$ told his$_{i/j}$ wife about his$_i$ brother$_j$."

Most of the native speakers interpret the sentence (1a) as "the man" being the antecedent of both pronouns at first interpretation and "the brother" as being the antecedent of the second pronoun in the second interpretation. However, a few native speakers claim that the "brother" could be the antecedent of the first pronoun as well. However they also state that if this is the case they prefer not to omit the pronoun or change the order. Since most of the native speakers agree that this interpretation is invalid we prefer to neglect it. Although (1b) is just the scrambled version of the same sentence[1] dropped pronouns have different antecedents. These examples show that the antecedent should precede the dropped pronoun. We thus consider the resolution as an incremental process and assume that pronouns are bound to one of the known antecedent candidates during left-to-right interpretation of the sentence.

(2) [(*pro*$_i$) Duyduğu]$_{RC}$ ses adamı$_i$ şaşırttı.
he hear-*Rel-3sg* voice man-*Acc* surprise-*Caus-Tense*
"The voice that he$_i$ heard surprised the man$_i$."

(3) a. Çocuk$_i$ (*pro*$_{i/j}$) abisini sormuş Mehmet'e$_j$.
child (his) brother-*3SP-Acc* ask-*Tense* Mehmet-*Dat*
"The child$_i$ asked Mehmet$_j$ for his$_{i/j}$ brother."

b. Çocuk$_i$ (*pro*$_{i/*j}$) abisini Mehmet'e$_j$ sormuş.
child (his) brother-*3SP-Acc* Mehmet'-*Dat* ask-*Tense*
"The child$_i$ asked his$_{i/*j}$ brother to Mehmet$_j$."

One exception on the linear ordering between the antecedent and the pronoun exists when the dropped pronoun occurs within the first NP. In (2) the dropped pronoun in the first NP (dropped pronoun is the subject of the relative clause in this case) is bound to a succeeding noun. Another exception arises when the arguments are scrambled to postverbal position. In (3a) "Mehmet" is in the postverbal position which can also be the antecedent of the pronoun. Nevertheless in the same sentence where "Mehmet" is in the pre-verbal position (3b) it cannot be the antecedent.

[1]Turkish is a free word order language where constituents scramble freely

Pronouns are bound by nominal objects on the lower levels of nesting as long as surface rules hold. Consider the sentences :

(4) a. [Kızın$_i$ konuştuğu]$_{RC}$ adam$_j$ (*pro*$_{j/i}$) annesinden söz etti.
girl-*Gen* talk-*Rel-3sg* man (his/her) mother-*3SP-Abl* mention-*Tense*
"The man$_j$ that the girl$_i$ talked with has talked about his$_j$/her$_i$ mother."

b. [[[(*pro*$_j$) Yavrusunu$_i$]$_{NP}$ besleyen]$_{RC}$ kuş$_j$]$_{NP}$ (*pro*$_{i/j}$) kanadını kırdı.
(its) infant feed-*Rel-3sg* bird (its) wing-*3SP-Acc* break-*Tense*
"The bird$_j$ that feeds its$_i$ infant$_i$ has broken its$_{i/j}$ wing."

It is clear from the previous examples that binding is possible even in two level nesting. The following is the example for relativized nouns:

(5) Bahçedeki$_i$ adam$_j$ (*pro*$_{*i/j}$) çiçeklerini kokladı.
garden-*Loc-Relv* man (his/its) flower-*3Pl-Acc* smell-*Tense*
"The man$_j$ in the garden$_i$ has smelled his$_{*i/j}$ flowers."

In (5) the possible candidates for the owner of the "flowers" are "man" and "garden". "garden" has a relativization suffix "`-ki`" that prevent it to become the antecedent of the dropped pronoun. Therefore this forms an additional exception concerning relativized nouns.

Binding is even possible for three level nested sentences like:

(6) [[[Kızın$_i$ çıktığı]$_{RC}$ adamın$_j$]$_{NP}$ pantolonunu]$_{NP}$ (*pro*$_{i/j}$) annesi ütüledi.
girl-*Gen* be-*Rel-3sg* with man-*Gen* trousers-*3SP-Acc* (his/her) mother-*3SP* iron--*Tense*
"The girls$_i$ mother has ironed the trousers of the man whom she$_i$ is with./
The mother of the man$_i$ whom the girl is with has ironed his$_i$ trousers"

However constraining the resolution process only by surface order rules is overgenerative in the sense that it tries to produce all possible interpretations most of which are semantically invalidated by native judgments. One way to cope with this problem is to extend the semantic information of the reference indices unified. In the normal case *ref* sort consists of agreement structure (person and number) and this is enough for binding two nominal objects. However it is possible to have *ref* sort also contain selectional restrictions on the object. These may include gender[2], animacy, human vs. animal, physical and classificational properties of object etc. All

[2]Turkish pronouns do not mark gender

this information can be put into a type inheritance hierarchy and resolved by unification during parsing.

(7a) and (7b) are the scrambled versions of the same sentence. When resolution is done without semantic constraints structurally similar interpretations are obtained which will be semantically ill-formed. For example in (7b) "dress" is an artifact and it cannot have a "wife" and the second interpretation is disambiguated by semantic information.

(7) a. Ahmet$_i$ (*pro*$_i$) karısına$_j$ (*pro*$_{i/j}$) giysisini verdi.
Ahmet (his) wife-*3SP-Dat* (his/her) dress-*3SP-Acc* give-*Tense*
"Ahmet$_i$ gave his$_i$ wife$_j$ his$_i$/her$_j$ dress."

b. Ahmet$_i$ (*pro*$_i$) giysisini$_j$ (*pro*$_{i/*j}$) karısına$_j$ verdi.
Ahmet (his) dress-*3SP-Acc* (his) wife-*3SP-Dat* give-*Tense*
"Ahmet$_i$ gave his$_i$ dress to his$_i$ wife."

In summary we have the following rules:

- Nominal objects preceding the dropped pronoun are candidate antecedents.
- If the dropped pronoun belongs to the first NP of the sentence, following nouns are candidate antecedents.
- Nominal objects in the post-verbal position are candidate antecedents of any dropped pronoun before the verb.
- Relativized nouns cannot be antecedents of the dropped pronouns.

2 Implementation

We have tried to formalize this linguistic phenomena in an HPSG parser for Turkish [5]. In HPSG, all nonlocal dependencies of a constituent are carried outside of the phrase by NON-LOCAL feature. We have added a new feature called PRO under INHERITED and TO-BIND features consisting of a set [3] of noun references (cf. 8). Dropped pronouns insert their reference into "INHERITED|PRO" set where an overt noun inserts it into "TO-BIND|PRO" set. These

[3] We have used the "set" term as unordered lists of distinct objects rather than the set descriptions mentioned in [3] where several feature structure elements in the set might describe the same object.

are carried along the phrases while producing the possible interpretations and preserving the linear order of results at each step.

(8)PRO structure

$$\text{PRO} \quad \left\{ \left[\text{INDEX} \quad \textit{ref} \right], \ldots \right\}$$

a. $$\left[\text{TO-BIND} \mid \text{PRO} \quad \left\{ .., \boxed{1}, .. \right\} \right] \ll \left[\text{INHERITED} \mid \text{PRO} \quad \left\{ .., \boxed{1}, .. \right\} \right] \ll \begin{bmatrix} \text{LEX} & + \\ \text{HEAD} & \text{verb} \end{bmatrix}$$

b. $$\perp < \left[\text{INHERITED} \mid \text{PRO} \quad \left\{ .., \boxed{1}, .. \right\} \right] \ll \left[\text{TO-BIND} \mid \text{PRO} \quad \left\{ .., \boxed{1}, .. \right\} \right]$$

c. $$\left[\text{INHERITED} \mid \text{PRO} \quad \left\{ .., \boxed{1}, .. \right\} \right] \ll \begin{bmatrix} \text{LEX} & + \\ \text{HEAD} & \text{verb} \end{bmatrix} \ll \left[\text{TO-BIND} \mid \text{PRO} \quad \left\{ .., \boxed{1}, .. \right\} \right]$$

Rules for the possible interpretations are given in (8a-c). (8a) describes that antecedent should precede the pronoun. $\ll$ has been used to indicate that the left operand precedes the second in the surface. PRO list. Similarly $<$ indicates the immediate precedence. Rule (8b) handles the exception of first noun phrase. It might be bound to a new object coming from the outside context or bound to the immediately succeeding noun object. And the last one (8c) is for the postverbal antecedents (cf. 7b,c).

Implementation is based on a HPSG style grammar of Turkish in ALE [1]. The existing parser implements a subset of Turkish including free constituent order, relative clauses, pronoun drops, adjuncts, and inflectional morphology. Since Turkish constituents can scramble freely and especially adverbs may occur anywhere within the sentence, arguments in the subcategorization structure are reduced one at a time during parsing. So that in the implementation, resolution process has to be applied incrementally as arguments are reduced. However to carry the information of the arguments in the postverbal position which is useful in (8c) the PRO structure is slightly modified.

(9) $$\text{PRO} \quad \left\{ \begin{bmatrix} \text{POSTV} & \textit{bool} \\ \text{INDEX} & \textit{index} \end{bmatrix}, \ldots \right\}$$

Where POSTV is a flag indicating whether the pronoun information is coming from a postverbal position or not. The elements in the INHERITED|PRO set are the indices of the dropped pronouns and the elements in the TO-BIND|PRO set are the candidate nominal indices within the phrase. In this configuration empty lexical entry in the place of a dropped pronoun will

have its index in the INHERITED|PRO and a lexical entry for a noun will contain its index in the TO-BIND|PRO. All the other lexical entries will have empty sets in these two features. The only exception to this is the relativized nouns which should not be bound to any pronoun, hence resulting in an empty TO-BIND|PRO feature.

In the subcategorization process these features are combined and bound, giving all possible binding combinations with the principles listed below:

1. Indices in the TO-BIND|PRO set of the left constituent are bound with the indices in the INHERITED|PRO set of the right constituent.

2. If TO-BIND|PRO feature of the right constituent has indices with POSTV $+$ then these indices are bound to the indices in the INHERITED|PRO feature with POSTV $-$ of the left constituent which handles the exception for the postverbal arguments.

3. If left constituent is the last item in the SUBCAT structure and INHERITED|PRO is non-empty, these indices are bound to the TO-BIND|PRO structure of the head phrase which handles the dropped pronouns in the sentence initial position.

4. TO-BIND|PRO feature of the upper phrase is simply the union of the two TO-BIND|PRO features.

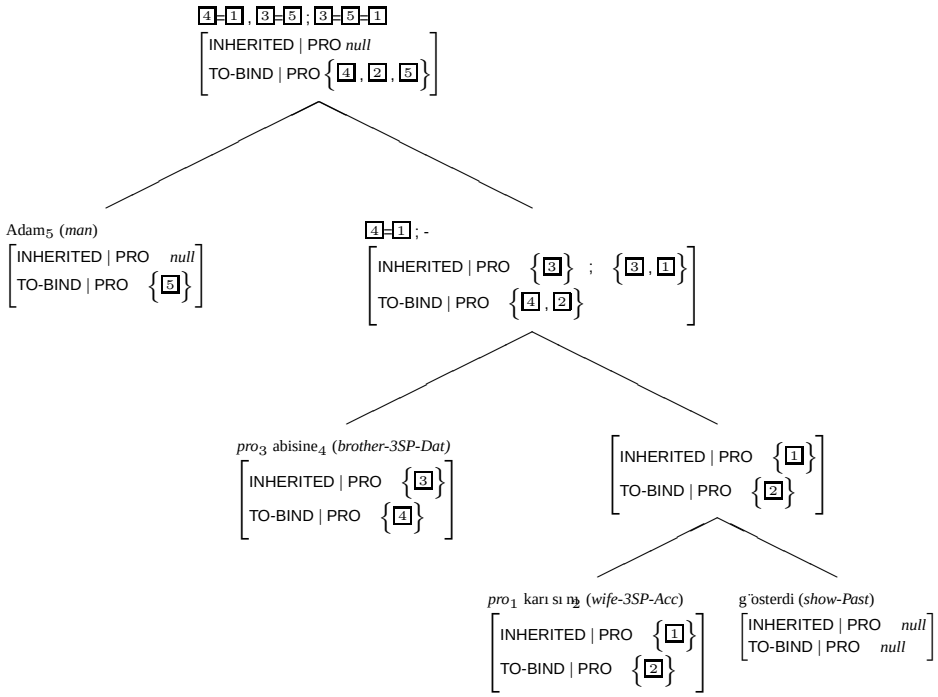
5. INHERITED|PRO feature of the upper phrase is the union of the INHERITED|PRO structures with bound indices deleted.

In the implementation, all these principles are coded as ALE definite clauses which are very similar to Prolog clauses. For generating all resolution ambiguities information in two sets should be unified in a special manner. An index in the TO-BIND|PRO set can bind multiple indices in the INHERITED|PRO set. An index in the INHERITED|PRO set on the other hand can be bound by only one index. Hence all combinations of many-to-one mappings from TO-BIND indices to INHERITED indices should be generated as ambiguities. Following example indicates how two index sets are bound. First set is the TO-BIND set from nominal objects and second set is the INHERITED set from dropped pronouns.

(10) $\texttt{bindset}(\{\boxed{1},\boxed{2}\},\{\boxed{a},\boxed{b}\}) = \{\boxed{1}=\boxed{a}=\boxed{b}\} \vee \{\boxed{1}=\boxed{a},\boxed{2}=\boxed{b}\} \vee \{\boxed{2}=\boxed{a},\boxed{1}=\boxed{b}\} \vee \{\boxed{2}=\boxed{a}=\boxed{b}\} \vee \{\boxed{1}=\boxed{a}\} \vee \{\boxed{1}=\boxed{b}\} \vee \{\boxed{2}=\boxed{a}\} \vee \{\boxed{2}=\boxed{b}\} \vee \{\}$

The system is setup to generate all the resolution ambiguities during parsing using the above clause for combining two sets. In the following example the resolution process of the sentence "Adam$_i$ *(pro$_i$)* abisine$_j$ *(pro$_{i/j}$)* karısını gösterdi. (*The man$_i$ showed his$_{i/j}$ wife to his$_i$ brother$_j$.*)" is illustrated. For simplicity parses with the unbound pronouns are not shown.

(11)



3 Conclusion

This work proposes a method for resolution of pronouns in a pro-drop language, Turkish. We did not concentrate on other pro-drop languages like Spanish and Chinese. Rules could be analyzed in a broader perspective to cover more pro-drop languages and extract some universal statements.

Native speakers basically use two main clues for resolving dropped pronouns: surface order and semantic information. Free word order feature of Turkish helps speakers to state the correct resolution form by changing the word order.

Semantic clues are as important as the surface order since native speakers eliminate most of the ambiguities semantically by their knowledge about the world and the objects. So it is essentially important to take semantics into account for producing only valid ambiguities.

References

[1] Bob Carpenter and Gerald Penn. *The Attribute Logic Engine User's Guide, Version 2.0*. Carnegie Mellon University, Pittsburgh, August 1994.

[2] Geoffrey L. Lewis. *Turkish Grammar*. Oxford University Press., Oxford, UK, 1967.

[3] Carl Polard and Ivan A. Sag. *Information Based Syntax and Semantics*. CSLI, 1987.

[4] Carl Pollard and Ivan A. Sag. *Head-driven Phrase Structure Grammar*. CSLI Chicago, 1994.

[5] Onur Şehitoğlu. A sign-based phrase structure grammar for Turkish. Master's thesis, Middle East Technical University, January 1996.

[6] Rasim Şimşek. *Örneklerle Türkçe Sözdizimi*. Kuzey Matbaacılık, 1987.