World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:5, 2007

A Computational Model for Resolving Pronominal Anaphora in Turkish Using Hobbs' Naïve Algorithm

Pınar Tüfekçi and Yılmaz Kılıçaslan

***Abstract*—**In this paper we present a computational model for pronominal anaphora resolution in Turkish. The model is based on Hobbs' Naïve Algorithm [4, 5, 6], which exploits only the surface syntax of sentences in a given text.

***Keywords*—**Anaphora Resolution, Pronoun Resolution, Syntax-based Algorithms, Naïve Algorithm.

I. Introduction

ANAPHORIC dependence is a relation between two linguistic expressions such that the interpretation of one, called anaphora, is dependent on the interpretation of the other, called antecedent. The problem of anaphora resolution is to find the antecedent(s) for every anaphora [7]. A model or algorithm for carrying out such a resolution process will be an essential component of any speech or text understanding system intended to handle realistic discourse or text fragments satisfactorily [2].

To speak more specifically, anaphora resolution, which most commonly appears as pronoun resolution, is the problem of resolving references to other items in the discourse. These items are usually noun phrases representing objects in the real world called referents but can also be verb phrases, whole sentences or paragraphs.

Anaphora resolution is classically recognized as a very difficult problem in Natural Language Processing [2, 12, 13]. Work on anaphora resolution in the open literature tends to fall into three domains: *artificial intelligence* (as a specialty of computer science, including computational linguistics and natural language processing), *classical linguistics* (as distinguished from computational linguistics), and *cognitive psychology*. Psychologists tend to be interested in this topic because of their interest in how the brain processes language. Linguists are interested in anaphora resolution simply because this is a classical problem in the field [2]. For our purposes we are primarily interested in the AI/computational linguistics approach. We will only be concerned with *computational* approaches to pronominal anaphora resolution algorithm that have been implemented on a computer in Prolog.

Manuscript received March 31, 2005.

Pınar Tüfekçi is with the Electronics and Communication Engineering Department, Çorlu Faculty of Engineering, Trakya University, Tekirdağ, Turkey (phone: + 90-282-652 94 75; fax: + 90-282-652 93 72; e-mail: pinart@corlu.edu.tr).

Yılmaz Kılıçaslan is with the Computer Engineering Department, Faculty of Engineering and Architecture, Trakya University, Edirne, Turkey (e-mail: yilmazk@trakya.edu.tr).

The aim of this paper is to implement a system that is based on Hobbs' Naïve Algorithm for pronominal anaphora resolution in Turkish. The system processes low level information by using syntactic knowledge to collect possible antecedents of pronouns. Then the future work will be determining the most plausible candidate by means of higher level information by using semantic and pragmatic pieces of knowledge. The relevant literature on pronoun interpretation ([5], [8], [15]) showed that a success rate of 80% is feasible when employing syntactic information alone for English. Again, as part of our future work we intend to compare Turkish and English with respect to their rate of success.

To the best of our knowledge, [18]'s BABY-SIT is the sole computational work that is intended to deal with anaphora resolution in Turkish, along with many other aspects of the language [20]. [18] uses situation-theoretic tools and notions. [20] is an another computational work that is based on Centering Theory to deal with pronominal anaphora resolution in Turkish and it particularly exploits the findings arrived by applying this theory to Turkish.

II. The Syntactic Approach

A. Types of Anaphora

There are primarily three types of anaphora:

- Pronominal: This is the most common type where a referent is referred to by a pronoun.
- Definite noun phrase: The antecedent is referred to by a phrase of the form "<the><noun phrase>".
- Quantifier/Ordinal: The anaphor is a quantifier such as 'one' or an ordinal such as 'first'[14].

Pronominal anaphora are the most commonly encountered in general usage. This category includes three subclasses: Personal, demonstrative and reflexive [2]. Pronominal anaphora in English and Turkish are shown in Table I [21].

TABLE I. PRONOMINAL ANAPHORA

Pronominal Anaphora in English			Pronominal Anaphora in Turkish		
Personal	Demonstrative	Reflexive	Personal	Demonstrative	Reflexive
he	this	himself	o	bu	kendi
she	that	herself	onu	bunu	kendisi
it	these	itself	onun	bunun	kendim
his	those	themselves	onlar	bunlar	kendin
her	others		onları	bunları	kendimiz
him			onların	bunların	kendiniz
its				şu	kendileri
they				şunu	
them				şunun	
their				şunlar	
				şunları	
				şunların	

For the purpose of this study, we will narrow down the scope of anaphoric phenomena and focus on a sub-problem of anaphora resolution, namely, the resolution of 3rd person singular pronominal anaphora to noun-phrase antecedents. Most algorithms in the literature resolve the pronouns 'he', 'she', 'it', 'her', 'him', 'his', 'her', 'its' in English whenever they have an antecedent which is a noun phrase. The algorithm we offer in this study will resolve the pronouns 'o', 'onu', 'onun', and 'kendi' in Turkish whenever they have an antecedent which is a noun phrase.

B. The Naïve Algorithm

In his 1977 paper, Hobbs presents two algorithms of pronominal anaphora resolution: - a syntax-based algorithm, known as the Naïve Algorithm, and a semantic algorithm. We will concentrate on the Naïve Algorithm for finding antecedents of pronouns here.

The Naïve Algorithm consists of a single resolution procedure based on traversing full parse trees starting from the pronoun in a search for an appropriate antecedent. The algorithm assumes that the data is presented in the format of parse trees produced by a particular grammar- namely, the one where an NP node dominates an N-bar node, to which arguments of the head noun attach. The algorithm traverses the tree, from the pronoun up, stopping on certain S, NP and VP nodes, searching left-to-right breadth-first in the subtrees dominated by these nodes.

It will be necessary to assume that an NP node has an Nbar node below it, as proposed by Chomsky [1], to which a prepositional phrase containing an argument of the head noun may be attached. Truly adjunctive prepositional phrases are attached to the NP node in English. This assumption, or something equivalent to it, is necessary to distinguish between sentences (1) and (2) in English [6]. It is worth noting that where English has a prepositional phrase we use an NP which has a locative case in Turkish.

(1) Mr. Smith$_i$ saw a driver$_k$ in his$_{i,k}$ truck.
(2) Mr. Smith$_i$ saw a driver of his$_i$ truck.

In sentence (1) 'his' may refer to Mr. Smith or the driver, but in sentence (2) it may not refer to the driver. The structures for the relevant noun phareses in sentences (1) and (2) are shown in Fig. 1.

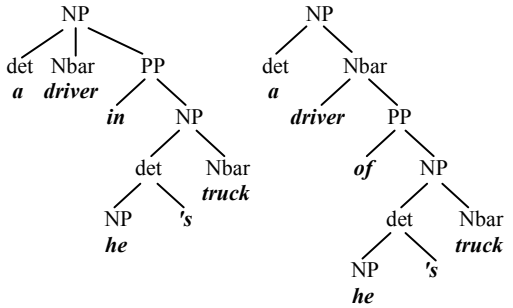

Fig. 1. The structures for NPs of sentences (1) and (2).

We translate sentence (1) from English to Turkish in four different forms as indicated in sentences (3), (4), (5) and (6).

(3) Mr. Smith bir şoför-ü$_i$ o-nun$_{i,k}$
Mr. Smith one driver-ACC s/he-GEN-3.SG
kamyon-u-n-da gör-dü.
truck-POSS-3.SG-LOC see-PAST.
'Mr.Smith saw a driver in his truck.'

(4) Mr. Smith$_i$ bir şoför-ü$_k$ Ø$_{i,k}$ kamyon-u-n-da
Mr.Smith one driver-ACC truck-POSS-3.SG-LOC
gör-dü.
see-PAST.
'Mr.Smith saw a driver in (his) truck.'

(5) Mr. Smith$_i$ o-nun$_k$ kamyon-u-n-da
Mr. Smith s/he-GEN-3.SG truck-POSS-3.SG-LOC
bir şoför gör-dü.
one driver see-PAST.
'Mr.Smith saw a driver in his truck.'

(6) Mr. Smith$_i$ Ø$_i$ kamyon-u-n-da bir şoför
Mr.Smith truck-POSS-3.SG-LOC one driver
gör-dü.
see-PAST.
'Mr.Smith saw a driver in (his) truck.'

In sentences (3), (4), (5) and (6) there are some ambiguous states. Let us look at them one by one:

In sentence (3) "onun" may be co-referential with "şoför" or another person in the previous sentences as the parse tree of (3) shows in Fig. 2. The syntactic tree structures of Turkish which are used in this study are based on [11, 9].

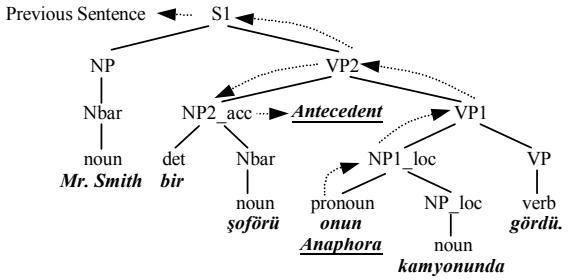

Fig. 2. The illustration of the parse tree of sentence (3) and the algorithm working on it.

The subject of the possessive NP can be null in Turkish [19]. In sentence (4) there is a null pronoun just before the object "kamyonunda" and it may be co-referential with "Mr. Smith" or "şoför". This null pronoun behaves either like the genitive-3 singular pronoun, "onun", or like the reflexive pronoun, "kendi", when the NP has a possessive-3 singular noun. If the null pronoun behaves like a GEN.3.SG pronoun, it is interpreted as co-referential with "şoför". If the null pronoun behaves like a reflexive pronoun, it is interpreted as co-referential with "Mr. Smith" as the parse tree of sentence (4) shows in Fig. 3.

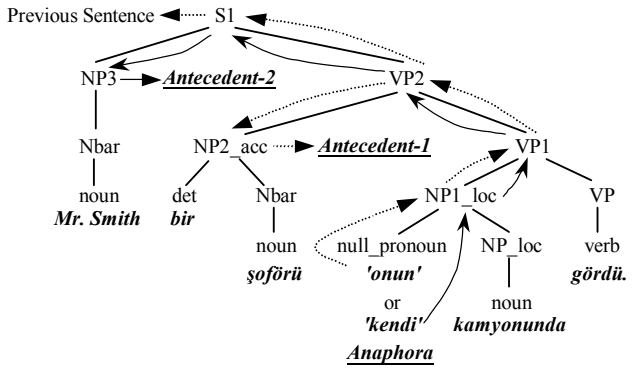

Fig. 3. The illustration of the parse tree of sentence (4) and the algorithm working on it.

In sentence (5) "onun" may be co-referential with another phrase in the previous sentences, as the parse tree of sentence (5) shows in Fig. 4.

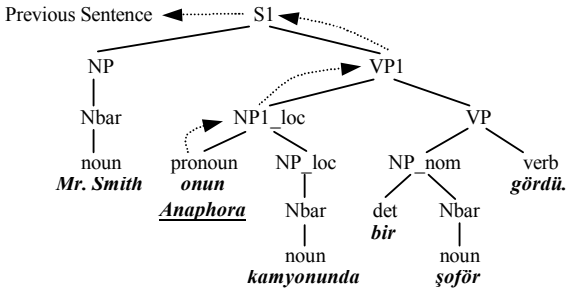

Fig. 4. The illustration of the parse tree of sentence (5) and the algorithm working on it.

In sentence (6) there is also a null pronoun just before the phrase "kamyonunda". The null pronoun behaves like the reflexive pronoun "kendi" and, hence, it becomes co-referential with "Mr. Smith". The parse tree of sentence (6) is shown in Fig. 5.

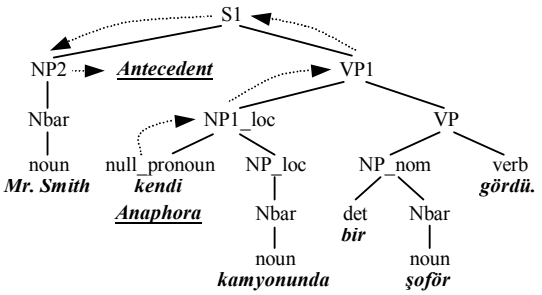

Fig. 5. The illustration of the parse tree of sentence (6) and the algorithm working on it.

According to [19] and [3], the subject of a possessive NP must be null when it is coreferential with the matrix subject, as in sentence (7a); if the possessive is informationally focused, the reflexive pronoun kendi 'own/self' is used, as in sentence (7b). An overt genitive pronoun forces disjoint reference irrespective of whether the antecedent precedes or follows the pronoun, as shown in sentences (7c) and (7d):

(7) a. Ahmet$_i$ [Ø$_i$ anne-si-n-i] sev-er.
Ahmet mother-POSS-3.SG-ACC love-AOR.
'Ahmet loves (his) mother.'

b. Ahmet$_i$ [kendi$_i$ anne-si-n-i] sev-er.
Ahmet self/own mother-POSS-3.SG-ACC loveAOR.
'Ahmet loves own mother.'

c. Ahmet$_i$ [o-nun$_k$ anne-si-n-i]
Ahmet he-GEN-3.SG mother-POSS-3.SG-ACC
sev-er.
love-AOR.
'Ahmet loves his mother.'

d. [O-nun$_k$ anne-si-n-i] sev-er
He-GEN-3.SG mother-POSS-3.SG-ACC love-AOR
Ahmet$_i$.
Ahmet.
'Ahmet loves his mother.'

In our opinion, if there is no accusative NP node preceding a possessive NP which has a null pronoun, the null pronoun is used just like the reflexive pronoun "kendi" as in sentence (6). This reflexive pronoun co-refers with the subject of the sentence as in sentences (6) and (7a). If there is an accusative NP preceding a possessive NP which has a null pronoun, the null pronoun is used like 'kendi' or 'onun' as in sentence (4). For this reason, the null prononun may co-refer with the subject of the sentence, when 'kendi' is used. On the other hand, the null pronoun may co-refer with an accusative NP preceding a possessive NP which has a null pronoun, when 'onun' is used.

C. Reformulation of the Naïve Algorithm for Turkish

We have reformulated Hobbs Naïve Algorithm so that it can be applied to Turkish. We have incorporated some new rules to the algorithm, as indicated below:

1. Begin at the NP node which immediately dominates a pronoun ('o', 'onu', 'onun' or 'kendi') or a null pronoun. If NP node immediately dominates a pronoun, continue to step 3.
2. Convert the null pronoun immediately dominated by the NP node to the pronoun 'onun' and the pronoun 'kendi' and apply the rest of the algorithm for each of these conversions separately. Firstly, apply the algorithm for 'kendi' and continue Step 4.
3. Secondly, apply the algorithm for 'onun' and continue to step 4.
4. Go up the tree to the first NP or VP node encountered. Call this node *X* and call the path used to reach it *p*.

5. If the pronoun is 'kendi', continue to step 8.
6. If *X* is an NP node, traverse all branches below node *X* to the left of path *p* in a left-to-right, breadth-first fashion. Propose as the antecedent any accusative NP node which is immediately dominated by *X* or propose as the antecedent any accusative NP node that is encountered which has an NP, VP or S node between it and *X.*
7. If *X* is an VP node, traverse all the other branches below node *X* except path *p.* Propose as the antecedent any accusative NP node which is immediately dominated by *X* or propose as the antecedent any accusative or genitive NP node that is encountered which has an NP, VP or S node between it and *X.*
8. From node *X* go up the tree to the first NP, VP or S node encountered. Call this new node *X,* and the path traversed to reach it *p.* If *X* is an NP node or a VP node, continue to step 5. If *X* is an S node, continue to step 9.
9. If the pronoun is "kendi", the antecedent is a nominative or genitive case-marked NP preceding it. If the pronoun is not "kendi", continue to step 10.
10. If node *X* is the highest S node in the sentence, traverse the surface parse trees of previous sentences in the text in order of recency, the most recent first; each tree is traversed in a left-to-right, breadth-first manner, and when an NP node is encountered, it is proposed as the antecedent. If *X* is not the highest S node in the sentence, continue to step 11.
11. From node *X*, go up the tree to the first NP, VP or S node encountered. Call this new node *X*, and call the path traversed to reach it *p.*
12. If *X* is an NP node and if the path *p* to *X* did not pass through the Nbar node that *X* immediately dominates, propose *X* as the antecedent.
13. If *X* is an NP node and if the path *p* passed through the N-bar node that *X* immediately dominates, traverse all branches below node *X* to the *left* of path *p* in a left-to-right, breadth-first manner. Propose any NP node encountered as the antecedent.
14. If *X* is a VP or S node, traverse all branches of node *X* to the *right* of path *p* in left-to-right, breadth-first manner, but do not go below any NP or VP or S node encountered. Propose any NP node encountered as the antecedent.
15. Go to step 10.

As [6] points out, a breadth-first search of a tree is one in which every node of depth n is visited before any node of depth n+1.

Figures 2, 3, 4 and 5 illustrate the algorithm working on the sentences (3), (4), (5) and (6). Figures 6 and 7 illustrate the algorithm working on the sentence (8b) which is the translation of the sentence (8a) from English to Turkish for determining the antecedents of each anaphora.

(8) a. The castle in Camelot remained the residence of the king until 536 when he moved it to London[6].

b. Camelot-ta-ki kale, kral-ın o-nu
Camelot-LOC-REL castle, king-GEN it-ACC

Londra-ya taşı-dı-ğı 536-ya kadar,
Londra-DAT move-PAST-ACC 536-DAT until,

o-nun rezidans-ı kal-dı.
s/he-GEN-3.SG residence-ACC remain-PAST.

Beginning from node NP1 which is immediately dominating the pronoun 'onu', step 3 rises to node NP2. Step 4 does not apply, because the pronoun is not 'kendi'. It's passed from step 3 to step 5. Step 5 searches the left portion of NP2's tree but finds no eligible NP node. Step 6 does not apply. Step 7 rises to node NP3. It's passed from step 7 to step 4. Step 5 searches the left portion of NP3's tree but finds no eligible NP node. Step 6 does not apply. Step 7 rises to node VP1 and it's passed from step 7 to step 4. Step 5 does not apply, it's passed to step 6. Step 6 searches the all branches below node VP1 except path p and proposes NP4 as antecedent. NP4 correctly determines 'rezidansı' as the antecedent of 'onu', as shown in Fig. 6.

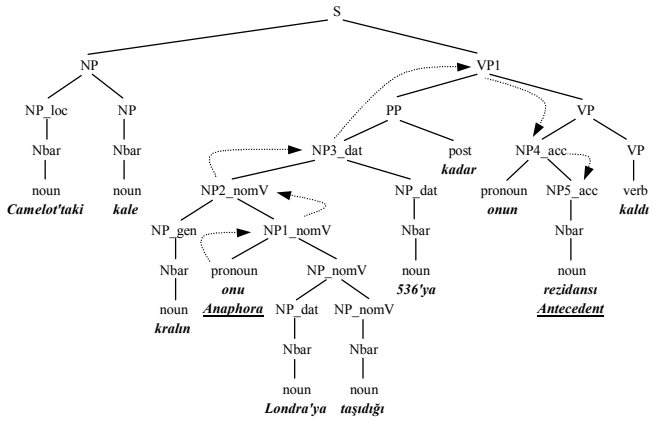

Fig. 6. The illustration of the parse tree of sentence (8b), the algorithm working on it and the determination of the antecedent of anaphora 'onu'.

If we search for the antecedent of 'onun', beginning from node NP1 immediately dominating the pronoun 'onun', step 3 rises to node VP1. Step 4 does not apply, because the pronoun is not 'kendi'. Step 5 does not apply and it's passed from step 3 to step 6. Step 6 searches the all branches below node VP1 except path p. Firstly it's proposed NP2 as antecedent in step 6. Thus, '536-ya' is recommended as the antecedent of 'onun'.

The algorithm can be improved somewhat by applying simple selectional constraints, such as; Dates and places and large fixed objects can't move [6].

After NP2 is rejected, it's proposed NP3 as antecedent in step 6. And finally ligting upon NP3 'kralın' as the antecedent of 'onun' in step 6 as shown in Fig. 7.

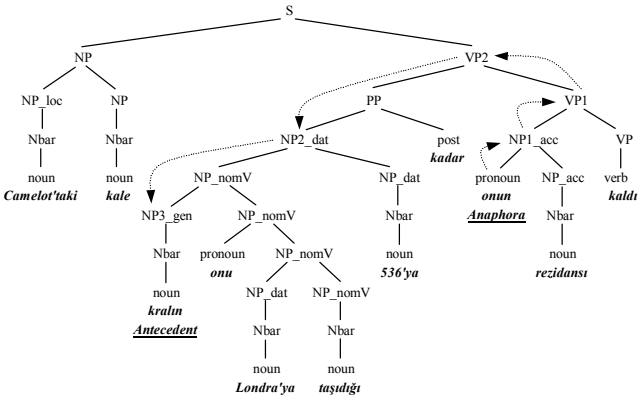

Fig. 7. The illustration of the parse tree of sentence (8b), the algorithm working on it and the determination of the antecedent of anaphora 'onun'.

III. Conclusion

We have implemented a version of the Hobbs' Naive Algorithm for Turkish by reformulating and incorporating some new rules to the algorithm. For issues relating to Turkish, we have rested upon the thematic hierarchy proposed by [10, 20]. The algorithm so far has been tested successfully on 10 toy sentences.

The idea we propose is to implement a system for pronoun resolution that locates likely antecedents according to the syntactic information. Then better models resulting from our future work will be able to select the most suitable one according to whether the corresponding logical form of the sentence would be consistent with the axioms in semantic and pragmatic.

References

[1] Chomsky, N., "Remarks on nominalization." In: R.Jacobs and P.Rosenbaum(eds.), Readings in transformational grammar, 184-221. Waltham, Mass.:Blaisdell, 1970.
[2] Denber M., "Automatic Resolution of Anaphora in English", Technical Report, Eastman Kodak Co. Imaging Science Division, June 30,1998; http://www.wlv.ac.uk/~le1825/anaphora_resolution_papers/denber.ps
[3] Erguvanlı-Taylan E., "Pronominal versus Zero Representation of Anaphora in Turkish", Studies in Turkish Linguistics, 1986.
[4] Hobbs J.R., "Pronoun Resolution,".Research Report 76-1, Department of Computer Sciences, City College, City University of New York, August 1976.
[5] Hobbs J.R., "38 Examples of Elusive Antecedents from Published Texts," Research Report 77-2, Department of Computer Sciences, City College, City University of New York, August 1977.
[6] Hobbs J.R., "Resolving Pronoun References," Lingua, Vol. 44, pp. 311-338. Also in Readings in Natural Language Processing, B. Grosz, K. Sparck-Jones, and B. Webber, editors, pp. 339-352, Morgan Kaufmann Publishers, Los Altos, California, 1978.
[7] Huang Y., "Anaphora: A Cross-linguistic Approach," New York: Oxford University Press, 2000.
[8] Kennedy, C. and Boguraev, B., "Anaphora for everyone: Pronominal Anaphora Resolution without a Parser.", COLING 96 pages 113-118 (89), 1996.
[9] Kennelly, S.D., "Nonspecific External Arguments in Turkish", Dilbilim Araştırmaları 7, İstanbul, p.58-75, 1997.
[10] Kılıçaslan, Y., "Information packaging in Turkish." Unpublished MSc. Thesis, University of Edinburg, Edinburg, 1994.
[11] Kılıçaslan, Y., "A Situation-Theoretic Approach to Case Marking Semantics in Turkish", Lingua , 2005.
[12] Mitkov, R., "Anaphora Resolution:The State of te Art, COLING'98/ACL'98 tutorial on anaphora resolution, University of Wolverhampton,1999.
[13] Mitkov, R., "Anaphora Resolution", Pearson Education, ISBN 0 582 32505 6, 2002.
[14] Sayed, I,Q.,"Issues in Anaphora Resolution", http://www.ceng.metu.edu.tr/courses/ceng463/project/BurakAysegul/project_report.pdf
[15] Shalom, L. and Herbert, L., " An algorithm for Pronominal Anaphora Resolution.", Computational Linguistics 20(4): 535-561, 1993.
[16] Tetrault, J., "Analysis of Syntax-based Pronoun Resolution Methods". In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pages 602-605, 1999.
[17] Tetrault, J., "A Corpus-based Evaluation of Centering and Pronoun Resolution." Computational Linguistics,2000.
[18] Tın E. and Akman V., "Situated Processing of Pronominal Anaphora", Bilkent University, Ankara, 1998.
[19] Turan, Ü.D., "Null vs. Overt Subjects in Turkish Discourse : A Centering Analysis", Ph.D. Dissertation, 1996
[20] Yıldırım, S. and Kılıçaslan, Y. and Aykaç, R.E., "A Computational Model for Anaphora Resolution in Turkish via Centering Theory: an Initial Approach." , 124-128. ICCI 2004. ISBN 975-98458-1-4, 2004.
[21] Yüksel, Ö., "Contextually Appropriate Anaphor/Pronoun Generation for Turkish", MSc. Thesis of The Middle East Technical University,1997.