TAG Analysis of Turkish Long Distance Dependencies

Elif Eyigöz
University of Rochester
Computer Studies Bldg.
Rochester, NY 14627, USA
eyigoz@cs.rochester.edu

Abstract

All permutations of a two level embedding sentence in Turkish is analyzed, in order to develop an LTAG grammar that can account for Turkish long distance dependencies. The fact that Turkish allows only long distance topicalization and extraposition is shown to be connected to a condition -the coherence condition- that draws the boundary between the acceptable and inacceptable permutations of the five word sentence under investigation. The LTAG grammar for this fragment of Turkish has two levels: the first level assumes lexicalized and linguistically appropriate elementary trees, where as the second level assumes elementary trees that are derived from the elementary trees of the first level, and are not lexicalized.

1 Introduction

The formal power of lexicalized TAG (LTAG) (Joshi et al., 1975; Schabes et al., 1988; Schabes, 1990) is adequate to assign appropriate structural descriptions to Turkish long distance scrambling. This provides an uncomplicated ground for the investigation of the mechanisms behind long distance scrambling in Turkish. In this paper, all permutations of a five word two level embedding structure are analyzed and an LTAG grammar is developed for this fragment of Turkish. Sentences involving scrambling from more than two levels of embedding are difficult to interpret, therefore the optimum compromise between the complexity of the structure and the validity of the analysis is determined by restricting the number of the words in the structure under investigation, which as a result limits the number of permutations to a manageable quantity.

The use of the adjunction operation to explain several linguistic phenomena such as raising, extraction, and long distance dependencies has been demonstrated in (Kroch and Joshi, 1985; Kroch and Joshi, 1987; Kroch and Baltin, 1989; Frank, 2000; Frank, 1992). However, it has been shown that German long distance scrambling can not be adequately described within the framework of lexicalized TAGs, as elements from subordinate clauses can scramble to any position in the matrix clause in German (Becker et al., 1991; Becker et al., 1992; Rambow, 1994). As a consequence, multi-component TAG (MC-TAG) (Weir, 1988; Becker et al., 1991; Rambow, 1994) grammars have been proposed for German and Korean scrambling (Rambow and Lee, 1994). Since Turkish, unlike German, allows only long distance topicalization and long distance extraposition, the formal power of LTAG is adequate to explain Turkish long distance dependencies.

The detailed analysis of the two level embedding sentence in section 2 brings forth a condition -the coherence condition- that draws the boundary between the acceptable and inacceptable permutations of the five word sentence. The LTAG grammar for this fragment of Turkish developed in section 3 and 4 serves multiple purposes. First, it was complied into a linear indexed grammar as explained in (Schabes and Shieber, 1992), and parsed with a parser written in Prolog (Shieber et al., 1995). Second, it shows that the set of derivations can be meaningfully partitioned according to the coherence condition. Finally, it reveals a connection between the coherence condition and the semantic function of long distance scrambling in Turkish.[1]

2 Turkish Long Distance Scrambling

Turkish is an head-final SOV language. Yet, there is no restriction on the order of arguments and adjuncts of simple sentences, as long as they are not referentially dependent and the sentence does not contain non-specific NPs or WH-phrases (Kural, 1992). Scrambling in Turkish causes different semantic interpretations. Scrambling to the sentence initial position marks the constituent as the *topic*, the immediately preverbal position marks it as the *focus*, and the post-verbal position as the *background* information (Ergüvanlı, 1984). Scrambling of case marked

[1] Following the literature on the free word order phenomena in Turkish, the term *scrambling*, in this paper, refers to any word order variation from the unmarked word order.

arguments and adjuncts out of subordinate clauses to sentence initial and sentence final positions, i.e. long distance topicalization and extraposition are also grammatical. Long distance scrambling to positions other than these two, i.e. scrambling without a *semantic function* is unnatural, and is considered ungrammatical.

This section gives an analysis of the two level embedding structure in (1) to determine the grammatical, acceptable and inacceptable permutations this five word sentence. This structure has two subordinate clauses, the subject positions of which are empty.[2] The sentence has two noun phrases: the most embedded verb has an NP complement *NP3*, and the matrix sentence has an NP subject *NP1*. The matrix verb *V1* has an infinitival complement (INF) with an ablative case (ABL). Likewise, *V2* has a verbal noun (VN) complement with accusative case (ACC). The most embedded verb *V3* has an NP complement with accusative case (ACC).[3]

(1) **Unmarked Order**

Mary çocukları susturmayı denemekten
Mary children-ACC silence-VN-ACC try-INF-ABL
yoruldu.
tired-PAST

'Mary is tired of trying to silence the children.'

NP1 NP3 V3 V2 V1

(2) *Çocukları* Mary susturmayı denemekten
children-ACC Mary silence-VN-ACC try-INF-ABL
yoruldu.
tired-PAST

[NP3] NP1 V3 V2 V1

(3) Mary susturmayı denemekten yoruldu
Mary silence-VN-ACC try-INF-ABL tired-PAST
çocukları.
children-ACC

NP1 V3 V2 V1 [NP3]

(4) ? *Çocukları susturmayı* Mary denemekten
children-ACC silence-VN-ACC Mary try-INF-ABL
yoruldu.
tired-PAST

[NP3 V3] NP1 V2 V1

(5) ? Mary denemekten yoruldu *çocukları*
Mary try-INF-ABL tired-PAST *children-ACC*
susturmayı.
silence-VN-ACC

NP1 V2 V1 [NP3 V3]

The most embedded argument *NP3* is long distance topicalized in (2), and is long distance extraposed in (3). *[NP3 V3]*, which is the complement of *V2*, is long distance topicalized in (4) and is long distance extraposed in (5).

(6) shows an ungrammatical sentence in which *NP3* extraposes and *V3* topicalizes. If *NP3*, *V2* and *V3* are separated into three as in (6), then the sentence not only becomes ungrammatical but also becomes inacceptable. Such a sentence is not more informative than a 'word salad' with respect to pragmatic inference. The coherence condition in (7) is proposed to rule out such inacceptable sentences.

(6) * *uğraşmaktan* Mary *bırakmaya* bıktı
try-INF-ABL Mary quit-VN-DAT tire-PAST-3SG
Sigarayı.
Cigarette-ACC

[V3] NP1 [V2] V1 [NP3]

'Mary is tired of trying to quit smoking.'

(7) *The Coherence Condition*
In acceptable sentences, **[[**NP3 V3**]** V2**]** is separated as **[**NP3 V3**]** - V2 or NP3 - **[**V3 V2**]**.

It is not the case that all sentences that do not violate the coherence condition are grammatical. The sentence in (8a) exemplify long distance topicalization of *NP3* when *[V3 V2]* is extraposed. Similarly in (8b), *[NP3 V3]* is topicalized and *V2* is extraposed. In both cases, an element of a subordinate clause is topicalized when its verb is extraposed, which results in an ungrammatical sentence.

(8) a. *? *Çocukları* Mary yoruldu
children-ACC Mary tired-PAST
susturmaya uğraşmaktan.
silence-VN-DAT try-INF-ABL

[NP3] NP1 V1 [V3 V2]

b. *? *Çocukları susturmaya* Mary
children-ACC silence-VN-ACC Mary
yoruldu *uğraşmaktan*.
tired-PAST try-INF-ABL

[NP3 V3] NP1 V1 [V2]

Since Turkish is a head-final language, embedding a sentence inside another one creates a center embedding structure. Moreover, long distance scrambling creates center embedding with crossing dependencies. Psycholinguistics studies indicate that such sentences in-

[2] Since the discussion on long distance scrambling does not hinge upon the existence of the silent PRO, it is left out in the analysis for the sake of the clarity of the presentation.

[3] The analysis proposed in this paper is independent of the choice of the verbs, the case markers on their complements, and the type of subordination. The analysis is intended to explain the least pragmatically restricted cases, the sentences that in fact can undergo long distance scrambling described in this work.

1	Long Distance Left Scrambling of NP3	• NP3 • [V3 V2]
	Long Distance Left Scrambling of [NP3 V3]	• [NP3 V3] • V2
2	Long Distance Right Scrambling of NP3	[V3 V2] NP3 • •
		[V3 V2] • NP3 •
		• [V3 V2] NP3 •
	Long Distance Right Scrambling of [NP3 V3]	V2 • [NP3 V3]•
3	Local Extraposition of [NP3 V3]	• V2 [NP3 V3] •
		V2 [NP3 V3] • •

Table 1: Permutations without a Semantic Function

crease processing load, which results in low acceptability judgments associated with these sentences. As indicated with the judgment '?' for (4) and (5), long distance topicalization and extraposition of *[V3 V2]* is more marked than long distance topicalization and extraposition of *NP3*.

Both the tendency to group the verbs as *[V3 V2 V1]*, and the coherence condition are reminiscent of the 'clause union' account of German and Dutch verb constructions (Evers, 1975). According to the 'clause union' hypothesis, verbs undergo a process by which they form a single complex verb. Similarly, the coherence condition seems to collapse the two level embedding structure into a one level embedding structure by either combining the *[V3 V2]* into one complex verb, or freezing *[NP3 V3]* as one complex object.

2.1 Semantic Function of Scrambling

Among the 120 permutations of the sentence in (1), only 42 word orders do not violate the coherence condition. However, 16 more sentences have to be ruled out because scrambling without a semantic function, i.e scrambling to positions other than the sentence initial *topic* and sentence final *background* positions is ungrammatical in Turkish. Therefore, only 26 out of 120 word orders are left to be accounted for.

The word orders that have to be ruled out are given in Table 1. The • shows the positions of the two elements of the matrix clause. In row one, a constituent from a subordinate clause is scrambled to the left, but it is not at the sentence initial position. In row two, a constituent from a subordinate clause is scrambled to the right, but it is not at the sentence final position. In row three, *[NP3 V3]* undergoes local extraposition.

The following section presents an LTAG grammar for the word orders that do not violate the coherence condition *and* involve scrambling with a semantic function. The grammar, through the adjunction operation, reveals a relation between the coherence condition and the semantic function of long distance scrambling. However, local extraposition cannot be related to the coherence condition in the same way, because derivation of local extraposition does not involve the adjunction operation.

Moreover, local extraposition of the subject in a one level embedding sentence is grammatical, as exemplified below. (9) shows the unmarked order. *S1* refers to the subject of the matrix clause, *S2* to the subject of the embedded clause, *O2* to the object of the embedded clause, *V1* and *V2* to the verbs of the matrix and the embedded clauses respectively. *S2* is extraposed in (10). Local extraposition of the subject in a subordinate clause places the subject in the preverbal *focus* position of the matrix clause, therefore it is not semantically vacuous. Local extraposition of a direct object in a subordinate clause, however, may be semantically vacuous because the object is already in a preverbal focus position at its base position.

(9) Elif Ali'nin Ankara'dan geldiğini
Elif Ali-GEN Ankara-ABL come-NOM-P2SG-ACC
biliyor.
know-PROG
S1 [S2 O2 V2] V1
'Elif knows that Ali came from Ankara.'

(10) Elif Ankara'dan geldiğini Ali'nin
Elif Ankara-ABL come-NOM-P2SG-ACC Ali-GEN
biliyor.
know-PROG
S1 [O2 V2 S2] V1
'Elif knows that Ali came from Ankara.'

Since local extraposition is ungrammatical in the structure under investigation, the sentences in row three of Table 1 are omitted in the LTAG grammar developed in the following section.

		SOV	OSV	OVS	SVO	VSO	VOS
V2-0 V3-0	Unmarked	√	√	√	√	√	√
V2-0 V3-1	Topicalization of [NP3]	√	SE	SE	A	A	A
V2-0 R3-1	Extraposition of [NP3]	√	A	A	A	A	A
V2-1 V3-0	Topicalization of [NP3 V3]	√	SE	SE	A	A	A
R2-1 V3-0	Extraposition of [NP3 V3]	√	A	A	A	A	A

Table 2: The summary of the 26 legitimate derivations

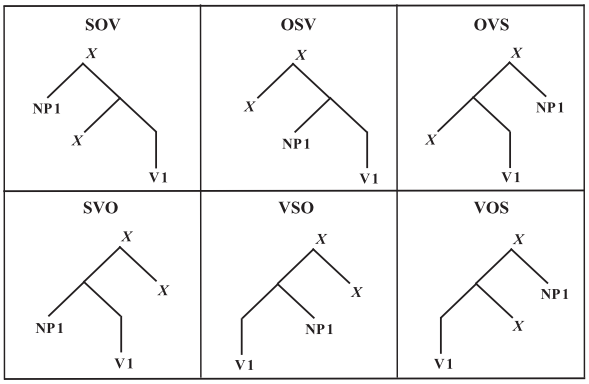

Figure 1: Elementary Matrix Trees

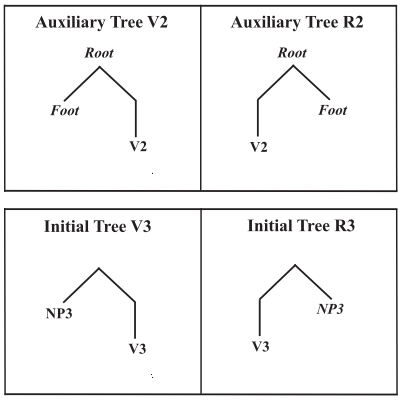

Figure 2: Elementary Trees

3 LTAG Grammar

The elementary structures that participate in the derivation of the two level embedding sentence are the clausal trees shown in Figure 1 and 2. Figure 1 and 2 show all word orders for each clause, only one of which participates in the derivation. The initial tree is the tree of the most subordinate clause, which is headed by *V3*. The two possibilities for the initial tree are shown in Figure 2: the head-initial tree is represented as 'R3', the head-final tree as 'V3'. Likewise, the head-final and head-initial auxiliary trees headed by *V2* are shown in 2. The matrix verb *V1* is a transitive verb, so there are six possible orders on the matrix clause, as shown in Figure 1. An MC-TAG grammar for Turkish local scrambling was demonstrated in (Eyigöz, 2007). Therefore, the elementary trees in Figure 1 and 2 are presumably derived by a set internal merge operation.

Adjoining a tree at a node below the root node may result in topicalization or extraposition of the arguments that are higher than the node of adjunction. The elements above the node of adjunction may be topicalized or extraposed depending on their *directionality* with respect to the node of adjunction. To derive this effect, clausal subcategorization is indicated by a footnode, as opposed to a substitution node.

A matrix *V1* tree adjoins into a tree of its subordinate clause headed by *V2* through its root and foot nodes, labeled *X* in Figure 1. A tree headed by *V2* adjoins into a tree of its subordinate clause headed by *V3* through its *Root* and *Foot* nodes. Since there is no clause that matrix *V1* is subordinate to, nothing adjoins into *V1* trees. As for *V2*, *R2*, *V3*, *R3* trees, it is assumed that adjoining does not take place at a foot node or a substitution node. Therefore, keeping track of the *level* of the node of adjunction is sufficient, as there is at most one possible node of adjunction at each level. As shown in Figure 2, adjunction at level 0 takes place at a root node, adjunction at the level 1 takes place at the sister of the *Foot* node on *V2*/*R2* trees, and at the sister of *NP3* on *V3*/*R3* trees. Finally, there is no possible node of adjunction at the third level. Therefore, there are two nodes of adjunction per tree, one at level 0 and one at level 1.

3.1 Restricting the Derivations

Two possible nodes for adjunction per tree means that there are 16x6 possible TAG derivations that could be

performed with the grammar in Figure 1 and 2. However, some of these derivations result in word orders that violate the coherence condition. Adjunction at the trees of *V2*/*R2* and *V3*/*R3* both at the first level results in word orders that either violate the coherence condition, or word orders that are string equivalent to the word orders derived by other derivations. Likewise, adjoining at the *R3* tree at the root level yields word orders that violate the coherence condition. Ruling out such derivations decreases the number of derivations to 6x6.

An interesting result of eliminating the derivations that violate the coherence condition is that only the derivations that involve long distance scrambling to the sentence initial and the sentence final positions, and local extraposition are left as legitimate derivations.

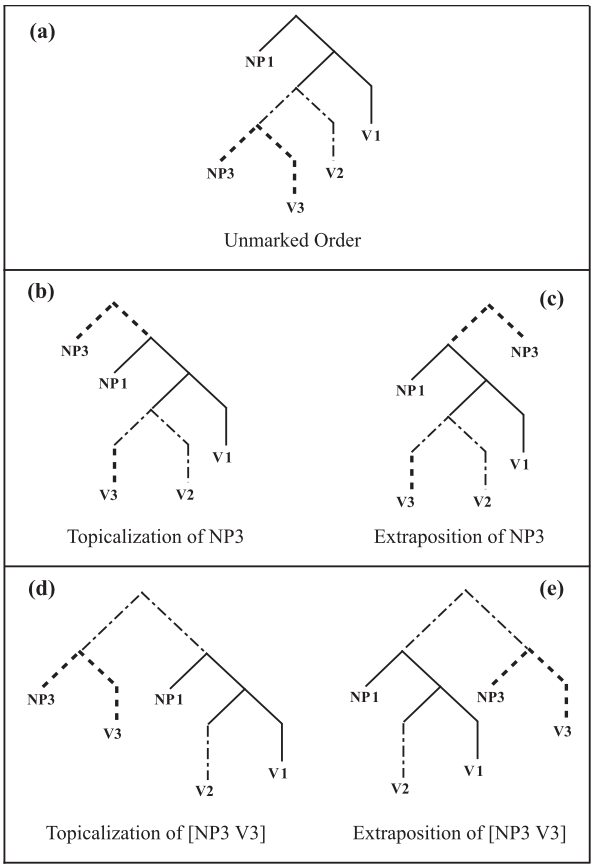

Figure 3: Derivation Examples

As argued in section 2, local extraposition in subordinate clauses has to be ruled out on grounds independent of the coherence condition. Adjoining at the root of the head-initial tree headed by *V2* (R2 in Figure 2) results in the local extraposition of its argument *[NP3 V3]*. Therefore, this derivation is also eliminated, which decreases the number of derivations from 6x6 to 5x6.

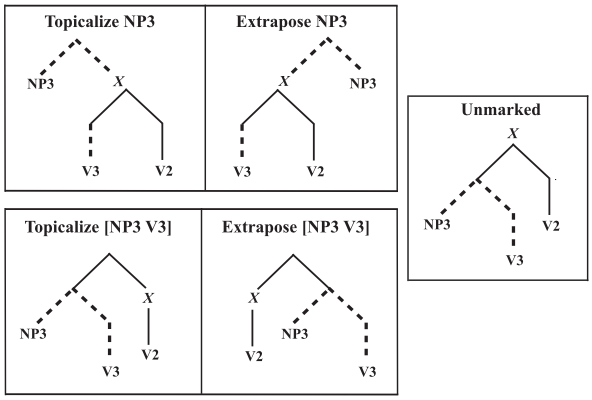

Figure 4: Revised Initial Trees

Figure 3 shows the results of the five legitimate derivations on the SOV order of the matrix clause. The trees in Figure 3 yield grammatical sentences. Figure (a) shows the unmarked order. Comparing (b) with the unmarked order in (a), we can see that adjoining the tree headed by *V2* into the tree headed by *V3* at level 1 results in topicalization of its argument *N3*. Figures (b) and (c) illustrate the derivation of topicalization and extraposition based on the directionality of the tree headed by *V3* (head-initial vs. head-final). Likewise, the trees in (d) and (e) show topicalization and extraposition of the argument *[NP3 V3]* based on the directionality of the tree headed by *V2*.

Table 2 summarizes the 5x6 legitimate derivations and acceptability judgments associated with them. Not all 30 possibilities are realized because topicalization out of a topicalized constituent is string vacuous topicalization. Therefore, topicalization does not apply to OSV and OVS word orders on the matrix clause, because the foot node is already at the sentence initial topic position in these trees. Accordingly, *SE* in Table 2 stands for sentences that are string equivalent to sentences derived by other derivations. $\surd$ in Table 2 stands for the grammatical sentences. Finally, *A* stands for sentences that are not grammatical but acceptable.

In section 2.1, the number of permutations that do not violate the coherence condition and involve scrambling with a semantic function was determined to be 26. Table 2 shows the linguistically appropriate derivations of these 26 word orders.

4 TAG Grammar Revisited

The coherence condition is enforced on the LTAG grammar developed in section 3 by restricting the set of possible derivations. In order to move from restrictions placed on derivations to restrictions placed on elementary trees, there are alternative paths to pursue. Motivated by the

grammaticality judgments listed in Table 2 and the coherence condition, the revised TAG grammar comprises of the revised initial trees in Figure 4 and the auxiliary matrix trees in Figure 1. Adjunction takes place at the nodes with the label *X* on the initial trees, through the nodes with the same label on the auxiliary trees.

The revised grammar comprises of two sets of trees to be combined. The first set -the initial trees in Figure 4- corresponds to the five rows of Table 2. The second set -the auxiliary matrix trees in Figure 1- corresponds to the six columns of Table 2. The combination of the unmarked SOV tree with any tree in Figure 4 results in a grammatical sentence. Similarly, the combination of the unmarked *[NP3 V3 V2]* tree with any tree in Figure 1 results in a grammatical sentence. The combination of the unmarked SOV tree with the unmarked *[NP3 V3 V2]* tree derives the unmarked word order at the upper left corner of Table 2.

As argued in section 2, the coherence condition is reminiscent of the 'clause union' hypothesis for German and Dutch verb constructions, in that the coherence condition seems to collapse the two level embedding structure into a one level embedding structure by either combining the *[V3 V2]* into one complex verb, or freezing *[NP3 V3]* as one complex object. The trees in Figure 4 reflect the merger expressed by the coherence condition.

5 Conclusion

The LTAG grammar proposed in this work has two levels: the first level assumes lexicalized and linguistically appropriate elementary trees, where as the second level assumes elementary trees that are derived from the elementary trees of the first level, and are not lexicalized. The choice of the grammar proposed in this work, especially the introduction of the second level, is motivated mainly by how conveniently the grammar expresses the special status of the unmarked order and how the grammar relates the unmarked order to the other grammatical word orders of the same sentence. Moreover, the coherence condition, which is a filter on the acceptable permutations of the two level embedding sentence, seems to express the merger that results in the second level of the LTAG grammar.

References

Tilman Becker, Aravind K. Joshi, and Owen Rambow. 1991. Long-distance scrambling and tree adjoining grammars. In *ACL*, pages 21–26.

Tilman Becker, Owen Rambow, and Michael Niv. 1992. The derivational generative power of formal systems or scrambling is beyond lcfrs. Technical Report IRCS-92-38, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA.

Emine E. Ergüvanlı. 1984. *The Function of Word Order in Turkish Grammar*. University of California Press, Los Angeles, California.

Arnold Evers. 1975. *The Transformational Cycle in Dutch and German*. Ph.D. thesis, University of Utrecht.

Elif Eyigöz. 2007. Tag analysis of turkish scrambling. Master's thesis, UCLA, Los Angeles, CA.

Robert Evan Frank. 1992. *Syntactic locality and Tree Adjoining Grammar: grammatical, acquisition and processing perspectives*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA.

Robert Frank. 2000. *Phrase Structure Composition and Syntactic Dependencies*. MIT Press.

Aravind K. Joshi, Leon S. Levy, and Masako Takahashi. 1975. Tree adjunct grammars. *J. Comput. Syst. Sci.*, 10(1):136–163.

Antony Kroch and Mark Baltin, editors, 1989. *Asymmetries in Long Distance Extraction in a Tree Adjoining Grammar*. University of Chicago Press.

Anthony Kroch and Aravind Joshi. 1985. Linguistic relevance of tree adjoining grammar. Technical Report MS-SC-85-16, Department of Computer and Information Sciences, University of Pennsylvania.

Antony Kroch and Aravind Joshi, 1987. *Analyzing Extraposition in a Tree Adjoining Grammar*, page 107149. Syntax and semantics. Academic Press, New York.

Murat Kural. 1992. Properties of Turkish scrambling. Master's thesis, UCLA.

Owen Rambow and Young-Suk Lee. 1994. Word order variation and tree-adjoining grammar. *Computational Intelligence*, 10:386–400.

Owen Rambow. 1994. *Formal and Computational Aspects of Natural Language Processing*. Ph.D. thesis, University of Pennsylvania.

Yves Schabes and Stuart M. Shieber. 1992. An alternative conception of tree-adjoining derivation. In *ACL*, pages 167–176.

Yves Schabes, Anne Abeille, and Aravind K. Joshi. 1988. Parsing strategies with 'lexicalized' grammars: application to tree adjoining grammars. In *Proceedings of the 12th conference on Computational linguistics*, pages 578–583, Morristown, NJ, USA. Association for Computational Linguistics.

Yves Schabes. 1990. *Mathematical and computational aspects of lexicalized grammars*. Ph.D. thesis, University of Pennsylvania.

Stuart M. Shieber, Yves Schabes, and Fernando C. N. Pereira. 1995. Principles and implementation of deductive parsing. *J. Log. Program.*, 24(1&2):3–36.

David Weir. 1988. *Characterizing Mildly Context-Sensitive Grammar Formalisms*. Ph.D. thesis, University of Pennsylvania, Philadelphia.