Mildly Context-Sensitive Grammars

Aravind K. Joshi

Final version

MILDLY CONTEXT-SENSITIVE GRAMMARS and languages (MCSG, MCSL) arose out the study of formal grammars adequate to model natural language structures, which are as restricted as possible in their formal power when compared to the unrestricted grammars which are equivalent to Turing machines. In the early 1980's, a grammatical formalism called generalized phrase structure grammar (GPSG) was proposed to describe various syntactic phenomena previously described in transformational terms, see Gazdar et al (1985). GPSG is weakly equivalent to context-free grammars (CFG). In the late 1980's, some clear examples of natural language phenomena were discovered that required formal power beyond CFG, a well-known example is the crossing dependencies in subordinate clauses in Dutch. Hence, the question of how much power beyond CFG is necessary to describe these phenomena became important.

Tree Adjoining Grammars (TAG) and their lexicalized versions (LTAG) are tree generating systems. TAGs factor recursion and the domain of dependencies in a novel way, leading to 'localization' of dependencies, their long distance behavior following from the operation of composition, called 'adjoining' (Joshi 1985, Joshi and Schabes 1997). TAGs have more power than CFGs and this extra power is a corollary of factorization of recursion and the domain of dependencies. This extra power appears to be adequate for the various phenomena requiring formal power more than CFG. Based on the formal properties of TAGs it was proposed in Joshi (1985) that the class of grammars that is necessary for describing natural languages might be characterized as *mildly context-sensitive grammars* (MCSG, MCSL for the corresponding languages) possessing at least the following properties: 1) context-free languages (CFL) are properly contained in MCSL; 2) languages in MCSL can be parsed in polynomial time; 3) MCSGs capture only certain

kinds of dependencies, e.g., nested dependencies and certain limited kinds of crossing dependencies (e.g., in the subordinate clause constructions in Dutch or some variations of them, but perhaps not in the so-called MIX (or Bach) language, which consists of equal numbers of a's, b's, and c's in any order, and 4) languages in MCSL have linear growth property, i.e., if the strings of a language are arranged in increasing order of length then two consecutive lengths do not differ by arbitrarily large amounts, in fact, any given length can be described as a linear combination of a finite set of fixed lengths.

It should be noted that these properties do not precisely define MCSG but rather give only a rough characterization, as the properties are only necessary conditions, and further some of the properties are properties of structural descriptions rather than the languages, hence, difficult to characterize precisely. This characterization of MCSG, obviously motivated by the formal properties of TAGs would have remained only as a remark, if it were not for some subsequent developments.

During mid to late 80's several grammar formalisms were proposed, which are more powerful than CFGs. These are (1) Head Grammars (HG), which introduced some wrapping operations beyond the concatenation operation in CFG, see Pollard (1985), (2) Combinatory Categorial Grammars (CCG), which are categorial grammars with composition operations of function application, function composition (both harmonic and non-harmonic) and some limited type raising, see Steedman (1985,1996), and (3) Linear Indexed Grammars (LIG) which are like CFGs but in each rule the left hand symbol and no more than one symbol on the right hand side have stacks associated with them, see Gazdar (1985). These stacks share information are are manipulated in essentially the same way as push down stacks.

It turns out that these four systems TAG, HG, LIG, and CCG are all weakly equivalent, a very important result in the theory of formal grammars, see Joshi et al (1991), Vijay-Shanker and Weir (1994), Weir (1988). Since these relationships are established in a constructive manner, they help to develop insights into the relationships between the various aspects of these formalisms, for example, the domain of locality for each formalism, the operation of adjoining, head wrapping, manipulation of stack valued features, function application, function composition, etc, i.e., we develop an understanding of 'strong' relationships (i.e., relationship between structural descriptions provided by different formalisms) and not just the 'weak' relationships (i.e., relationships between the string languages generated by different formalisms). Understanding these strong relationships is critical to the de-

scription of natural language structures. More recently, MCSG have found applications for the description of secondary and higher structures of some biological sequences.

References Gazdar G., Klein E., Pullum G. K., and Sag I. A. (1985) *Generalized Phrase Structure Grammar*, Basil Blackwell, Oxford.

Gazdar G. (1985) Applicability of indexed grammars to natural languages. Technical Report, CSLI-85-34, Center for Study of Language and Information, Stanford University, Stanford, CA, USA

Joshi A. K. (1985). Tree-adjoining grammars: How much context sensitivity is required to provide reasonable structural descriptions? In D. Dowty, L. Karttunen, and A. Zwicky, editors, *Natural Language Parsing*, 206-250, Cambridge University Press.

Joshi A. K., Vijay-Shanker K., and Weir, D. (1991) *The Convergence of Mildly Context-Sensitive Grammar Formalisms*, In P. Sells, S.M. Shieber, and T. Wasow, The MIT Press, Cambridge, MA.

Joshi A. K. and Schabes, Y. (1997). Tree-Adjoining Grammars. In G. Rosenberg and A. Salomaa, editors, *Handbook of Formal Languages*, 69-123. Springer, Berlin.

Pollard C. (1985) *Generalized phrase structure grammars, head grammars and natural language*, Ph. D. Dissertation, Stanford University, Stanford, CA, USA.

Steedman M. (1985) Combinators and grammars, In R. Oherle, E. Bach, and D. Wheeler, editors, *Categorial Grammars and Natural Language Structures*, Foris, Dordrecht, Holland.

Steedman M. (1996). *Surface Structure and Interpretation.* MIT Press, Cambridge, MA.

Vijay-Shanker K. and Weir D. J. (1994) The equivalence of four extensions of context-free grammars, *Mathematical Systems Theory* 27, 511-546.

Weir D. J. (1988). *Characterizing mildly context-sensitive grammar formalisms.* Ph.D. Dissertation, University of Pennsylvania, Philadelphia. PA, USA.

Index Mild context sensitivity, tree adjoining grammar, lexicalized tree adjoining grammar, head grammar, linear indexed grammar, combinatory

categorial grammar, generalized phrase-structure grammar, generative capacity, MCSG, MCSL, TAG, LTAG, HG, LIG, CCG, GPSG