

LEXICALIZATION AND CATEGORIAL GRAMMARS – A STORY BAR-HILLEL MIGHT HAVE LIKED

Aravind K. Joshi
Department of Computer and Information Science
and
Institute for Research in Cognitive Science
University of Pennsylvania
Philadelphia, PA 19104
USA
joshi@linc.cis.upenn.edu

EXTENDED ABSTRACT

In the early 60's Professor Bar-Hillel and his co-workers published a series of seminal papers on formal grammars, their mathematical properties and linguistic adequacy. Many of these papers appear in Bar-Hillel 1964. One of their papers was on categorial grammars and phrase structure grammars (Bar-Hillel, Gaifman, and Shamir 1960). In this paper they showed that categorial grammars are weakly equivalent to context-free grammars. This was a period when any formal system that was shown to be equivalent or conjectured to be equivalent to CFG's was put aside, as CFG's were shown by Chomsky to be inadequate for the description of language. I suspect this was the reason why Bar-Hillel himself did not pursue the development of categorial grammars (see the last paragraph of Bar-Hillel, Gaifman and Shamir 1960).

Since mid-70's there has been renewed interest in the development of grammars that either achieve the effects of transformations by non-transformational methods, or significantly reduced the role of transformations in a transformational grammar. Generalized Phrase Structure Grammars (GPSG), Lexical Functional Grammars (LFG), Categorial Grammars of various kinds, Combinatory Categorial Grammars (CCG), in particular, and (Lexicalized) Tree-Adjoining Grammars ((L)TAG) are some of examples of the first kind. Various versions of Government and Binding Theory (GB), including the most recent proposals by Chomsky (Minimalist Theory) are examples of the second kind. LTAGs, in a sense, are 'transformational' because the composition operations in LTAGs are reminiscent of 'generalized transformations' of Chomsky in *Syntactic Structures*, a fact noted by him in his paper on Minimalist Theory.

Mathematical, computational, and linguistic properties of LTAGs, their extensions and other related systems have been extensively studied, All these properties follow from two key properties of LTAGs:

- **Extended Domain of Locality (EDL):** The elementary trees of LTAG provided an extended domain (as compared to CFG's or CFG-based grammars) for the specification of syntactic and related semantic dependencies.
- **Factoring Recursion from the Domain of Dependencies (FRD);** Recursion is factored away from the domains over which dependencies are specified.

LTAGs are more powerful than CFGs both weakly and, more importantly, strongly, in the sense that even if a language is context-free, LTAGs can provide structural descriptions not available in a CFG. LTAGs belong to the so-called class of 'mildly context-sensitive' grammars. LTAGs have proved useful also in establishing equivalences among various classes of grammars, Head Grammars, Linear Indexed Grammars, and CCGs for example.

The particular results concerning LTAGs for our current purpose concern lexicalization of grammars. A grammar G is said to be lexicalized if it consists of

- a finite of structures (strings, trees, dags, for example), each structure being associated with a lexical item, called its 'anchor', and
- a finite set of operations for composing these structures.

A grammar G is said to strongly lexicalize another grammar G' if G is a lexicalized grammar and if the structural descriptions (trees, for example) of G and G' are exactly the same.

The following results are easily established.

- CFGs cannot strongly lexicalize CFGs. Although for every CFG there is an equivalent CFG in the Greibach Normal Form (GNF), it only weakly lexicalizes the given CFG as only a weak equivalence is guaranteed by GNF.
- Tree Substitution Grammars (TSG), i.e., grammars with a finite set of elementary trees together with the operation of substitution, cannot strongly lexicalize CFGs.
- TSGs with substitution and another operation called *adjoining* can strongly lexicalize CFGs. But these grammars are exactly LTAGs. Thus LTAGs strongly lexicalize CFGs.

These results show how LTAGs arise naturally in the course of strong lexicalization of CFGs. Strong lexicalization was achieved by working with trees rather than strings, hence the property EDL, and by introducing adjoining, which results in the property FRD. Thus both EDL and FRD are crucial for strong lexicalization.

Now the categorial grammars of the kind studied by Ajdukiewicz and Bar-Hillel, CG(AB) are weakly equivalent to CFGs. CGs in general are, of course, lexicalized by definition. It is also easy to show that CG(AB)s cannot lexicalize CFGs (ignoring, for now, the question of relabeling the nodes).

As we have observed before, LTAGs arose in the process of strong lexicalization of CFGs. What would be an analogous process if we begin with CG(AB)? This is exactly the question I will discuss in this paper. The main idea is to work with a finite set of basic partial proof trees (basic PPTs) as the building blocks of a CG together with inference rules from proof trees to proof trees, the resulting system called CG (PPT), tentatively. CG (PPT) has properties which are remarkably similar to the properties of LTAGs. As we have said before, LTAGs implicitly capture the essential idea of 'generalized transformations' (albeit in a highly restricted form). Hence, CG(PPT) can also be thought of as capturing the idea of 'generalized transformations' (albeit in a nontransformational and highly constrained manner) in a categorial framework. The finite set of basic PPTs is essentially built out of CG(AB). This is somewhat similar to the idea of building the structural descriptions of the so-called 'kernel' sentences¹ by a CFG (and by implication CFG-equivalent grammars such as CG(AB) in the very early formulations of transformational grammars by Chomsky in *Syntactic Structures*, an idea suggested by Bar-Hillel et al. towards the end of their paper (Bar-Hillel, Gaifman and Shamir 1960). This suggestion and Bar-Hillel's strong interest in a comparative study of formal grammars are the basis for the second half of the title of this paper.

References

- Bar-Hillel, Y., Gaifman, C. and Shamir, E. 1960 On categorial and phrase structure grammars. *The Bulletin of the Research Council of Israel*, 9F:1-16.
- Bar-Hillel, Y. 1964 *Language and Information* Addison-Wesley Publishing Company, Palo Alto and The Jerusalem Academic Press Ltd. Jerusalem.

¹These are related but not equivalent to the kernel sentences of Harris.