Lambek Calculus and Formal Languages

(Extended abstract)

Mati Pentus

Department of Mathematical Logic
Faculty of Mechanics and Mathematics
Moscow State University
Moscow, Russia, 119899
pentus@lpcs.math.msu.ru

Introduction

The systematic study of generating grammars was started by N. Chomsky in the 50s (cf. [10]). He defined several classes of generating grammars, which are interesting for both linguists and mathematicians, e. g. context-sensitive grammars, context-free grammars, and linear grammars. On the other hand, categorial grammars were studied by Y. Bar-Hillel, J. Lambek and others. The notion of a basic categorial grammar was introduced in [1]. In the same paper it was proved that the languages recognized by basic categorial grammars are precisely the context-free ones.

Another kind of categorial grammar was introduced by J. Lambek [15]. These grammars are based on a syntactic calculus, presently known as the Lambek calculus. Chomsky [11] conjectured that these grammars are also equivalent to context-free ones. In [12] Cohen proved that every basic categorial grammar (and, thus, every context-free grammar) is equivalent to a Lambek grammar. He also proposed a proof of the converse. However, as pointed out in [6], this proof contains an error. Buszkowski proved that some special kinds of Lambek grammars are context-free [6, 8, 9]. These grammars use weakly unidirectional types or types of order at most two.

The first result of this paper (Theorem 1) says that Lambek grammars generate only context-free languages. Thus they are equivalent to context-free grammars and also to basic categorial grammars. This fact (sometimes called *the Chomsky Conjecture*) was proved in [16] and [17].

The inteneded syntactic string models, i.e., *free semigroup models* (also called *language models* or *L-models*) for the Lambek calculus were considered in [3], [4], and [5]. The more general class of *groupoid models* has been studied in [7], [13], and [14]. In [4] W. Buszkowski established that the product-free fragment of the Lambek calculus is L-complete (i.e., complete w.r.t. free semigroup models), using the canonical model. The question of L-completeness of the full Lambek calculus remained open (cf. [2]).

The second result of this paper (Theorem 2) gives a positive answer to this question. The proof has been publised in [18] and [19].

1 Preliminaries

For any set $\mathcal{M}$ we denote by $\mathcal{M}^+$ the set of all finite non-empty strings consisting of elements of $\mathcal{M}$. The set of all subsets of $\mathcal{M}$ is denoted by $\mathbf{P}(\mathcal{M})$.

We consider the syntactic calculus introduced in [15]. The types of the Lambek calculus are built of primitive types $p_1, p_2, \ldots$ and three binary connectives $\bullet, \backslash, /$. We shall denote the set of all types by Tp. Capital letters A,B, ... range over types. Capital Greek letters range over finite (possibly empty) sequences of types. Sequents of the Lambek calculus are of the form $\Gamma \to A$, where Γ is a nonempty sequence of types.

Axioms: $A \to A$

Rules:

$$\frac{\Gamma \to A \quad \Delta \to B}{\Gamma\Delta \to A\bullet B}\ (\to\bullet) \qquad \frac{\Gamma A B \Delta \to C}{\Gamma(A\bullet B)\Delta \to C}\ (\bullet\to)$$

$$\frac{A\Pi \to B}{\Pi \to A\backslash B}\ (\to\backslash) \text{ where } \Pi \text{ is non-empty} \quad \frac{\Pi \to A \quad \Gamma B \Delta \to C}{\Gamma\Pi(A\backslash B)\Delta \to C}\ (\backslash\to)$$

$$\frac{\Pi A \to B}{\Pi \to B/A}\ (\to/) \text{ where } \Pi \text{ is non-empty} \quad \frac{\Pi \to A \quad \Gamma B \Delta \to C}{\Gamma(B/A)\Pi\Delta \to C}\ (/\to)$$

$$\frac{\Pi \to B \quad \Gamma B \Delta \to A}{\Gamma\Pi\Delta \to A}\ (CUT)$$

The cut-elimination theorem for this calculus is proved in [15].

2 Lambek grammars recognize context-free languages

Definition. We assume that a finite alphabet $\mathcal{T}$ and a distinguished type D are given. A *Lambek grammar* is a mapping f such that, for all $t \in \mathcal{T}$, $f(t) \subset$ Tp and $f(t)$ is finite.

The *language generated by the Lambek grammar* is defined as the set of all expressions $t_1 \ldots t_n$ over the alphabet $\mathcal{T}$ for which there exists a derivable sequent $B_1 \ \ldots \ B_n \to D$ such that $B_i \in f(t_i)$ for all $i \leq n$.

Definition. We assume that two disjoint alphabets $\mathcal{T}$ and $\mathcal{W}$ are given. The elements of $\mathcal{T}$ are called *terminal symbols* and those of $\mathcal{W}$ are *auxiliary symbols.*

A *context-free rewrite rule* is of the form $X \Rightarrow e$, where X is an auxiliary symbol and e is a non-empty word in the alphabet $\mathcal{T} \cup \mathcal{W}$.

A *context-free grammar* is a finite set $\mathcal{R}$ of context-free rewrite rules, with one auxiliary symbol S designated as its *start symbol.*

By $\bar{\mathcal{G}}(\mathcal{T}, \mathcal{W}, S, \mathcal{R})$ we denote the set of all expressions over the alphabet $\mathcal{T} \cup \mathcal{W}$ that arise through some finite sequence of rewritings of the start symbol S via the rules of $\mathcal{R}$.

The *language generated by the context-free grammar* is defined as

$$\bar{\mathcal{G}}(\mathcal{T}, \mathcal{W}, S, \mathcal{R}) \cap \mathcal{T}^+.$$

Theorem 1. *For any Lambek grammar there exists a context-free grammar such that the languages generated by these grammars coincide.*

3 L-completeness of the Lambek calculus

Definition. We define *L-model* (also called *language model* or *free semigroup model*) to be a triplet $\langle \mathcal{W}^+, \circ, w \rangle$, where $\mathcal{W}$ is an arbitrary alphabet, $\circ$ denotes concatenation of words from $\mathcal{W}^+$, and w is a function $w\colon \mathrm{Tp} \to \mathbf{P}(\mathcal{W}^+)$ such that

(1) $w(A \bullet B) = w(A) \circ w(B)$;
(2) $w(A \backslash B) = \{\gamma \in \mathcal{W}^+ \mid w(A) \circ \{\gamma\} \subseteq w(B)\}$;
(3) $w(B/A) = \{\gamma \in \mathcal{W}^+ \mid \{\gamma\} \circ w(A) \subseteq w(B)\}$.

Here for any two sets $\mathcal{A} \subseteq \mathcal{W}^+$ and $\mathcal{B} \subseteq \mathcal{W}^+$ by $\mathcal{A} \circ \mathcal{B}$ we denote the set $\{\alpha \circ \beta \mid \alpha \in \mathcal{A} \text{ and } \beta \in \mathcal{B}\}$.

Definition. A sequent $A_1 \ldots A_n \to B$ is *true* in a model $\langle \mathcal{W}^+, \circ, w \rangle$ iff

$$w(A_1) \circ \ldots \circ w(A_n) \subseteq w(B).$$

Theorem 2. *A sequent is derivable in the Lambek calculus if and only if it is true in every L-model.*

Theorem 3. *A sequent is derivable in the Lambek calculus if and only if it is true in every L-model over an alphabet $\mathcal{W}$ consisting of two symbols.*

References

1. Y. Bar-Hillel, C. Gaifman, and E. Shamir. On categorial and phrase-structure grammars. *Bull. Res. Council Israel Sect. F*, 9F:1–16, 1960.
2. J. van Benthem. *Language in Action: Categories, Lambdas and Dynamic Logic.* North-Holland, Amsterdam, (Studies in Logic 130), 1991.
3. W. Buszkowski. Undecidability of some logical extensions of Ajdukiewicz-Lambek calculus. *Studia Logica*, 37:59–64, 1978.
4. W. Buszkowski. Compatibility of categorial grammar with an associated category system. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 28:229–238, 1982.
5. W. Buszkowski. Some decision problems in the theory of syntactic categories. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 28:539–548, 1982.
6. W. Buszkowski. The equivalence of unidirectional Lambek categorial grammars and context-free grammars. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 31:369–384, 1985.
7. W. Buszkowski. Completeness Results for Lambek Syntactic Calculus. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 32:13–28, 1986.
8. W. Buszkowski. Generative power of categorial grammars. In R.T. Oehrle, E. Bach, and D. Wheeler, editors, *Categorial Grammars and Natural Language Structures*, pages 69–94, Reidel, Dordrecht, 1988.
9. W. Buszkowski. On generative capacity of the Lambek calculus. In J. van Eijck, editor, *Logics in AI*, pages 139–152, Springer, Berlin, 1991.

10. N. Chomsky. *Syntactic Structures*. Mouton, The Hague, 1957.
11. N. Chomsky. Formal properties of grammars. In R.D. Luce et al., editors, *Handbook of Mathematical Psychology*, vol. 2, pages 323–418, Wiley, New York, 1963.
12. J.M. Cohen. The equivalence of two concepts of categorial grammar. *Information and Control*, 10:475–484, 1967.
13. K. Došen. A Completeness Theorem for the Lambek Calculus of Syntactic Categories. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 31:235–241, 1985.
14. K. Došen. *A brief survey of frames for the Lambek calculus*. Bericht 5-90, Konstanzer Berichte zur Logik und Wissenschaftstheorie, Universität Konstanz, 1990.
15. J. Lambek. The mathematics of sentence structure. *American Mathematical Monthly*, 65(3):154–170, 1958.
16. M. Pentus. *Lambek grammars are context free*. Preprint No. 8 of the Department of Math. Logic, Steklov Math. Institute, Series Logic and Computer Science, Moscow, 1992.
17. M. Pentus. Lambek grammars are context free. *Proceedings of the 8th Annual IEEE Symposium on Logic in Computer Science*, 429–433, 1993.
18. M. Pentus. Lambek calculus is L-complete. ILLC Prepublication Series LP–93–14, Institute for Logic, Language and Computation, University of Amsterdam, 1993.
19. M. Pentus. Models for the Lambek calculus. *Annals of Pure and Applied Logic*, 75:179–213, 1995.