Conjunctive Grammars in Greibach Normal Form and the Lambek Calculus with Additive Connectives

Stepan Kuznetsov

Moscow State University

Abstract

We prove that any language without the empty word, generated by a conjunctive grammar in Greibach normal form, is generated by a grammar based on the Lambek calculus enriched with additive ("intersection" and "union") connectives.

1 Conjunctive Grammars

Let Σ be an arbitrary finite alphabet, Σ^* is the set of all words, and Σ^+ is the set of all non-empty words over Σ .

We consider a generalisation of context-free grammars, introduced by Okhotin [9] (and earlier by Szabari [14]).

A conjunctive grammar is a quadruple $\mathcal{G} = \langle \Sigma, N, \mathcal{P}, S \rangle$, where Σ and N are two non-intersecting alphabets (Σ is the alphabet in which the language is being defined, its elements are called *terminal symbols*, and N is an auxiliary alphabet, consisting of *nonterminal symbols*), $S \in N$ (the *start symbol*), and \mathcal{P} is a finite set of *rules* of the form

$$A \to \beta_1 \& \dots \& \beta_m,$$

where $A \in N$, $m \ge 1$, $\beta_1, \ldots, \beta_m \in (\Sigma \cup N)^*$.

We define the language generated by this grammar in terms of a formal deduction system associated with the grammar [10]. This formal system derives pairs of the form [X, w], where $X \in \Sigma \cup N$ and $w \in \Sigma^*$. Axioms are pairs [a, a], for all $a \in \Sigma$, and for every rule $A \to B_{11} \dots B_{1m_1} \& \dots \& B_{k1} \dots B_{km_k} \in \mathcal{P}$,

Published in Proc. Formal Grammar 2012/2013, LNCS vol. 8036, pp. 242–249.

The final publication is available at link.springer.com:

 $http://link.springer.com/chapter/10.1007/978-3-642-39998-5_15$

 $B_{ji} \in \Sigma \cup N$, and for all strings $u_{ji} \in \Sigma^*$, $j \in \{1, \ldots, k\}$, $i \in \{1, \ldots, m_j\}$, that satisfy $u_{11} \ldots u_{1m_1} = \ldots = u_{k1} \ldots u_{km_k} = w$, there is a deduction rule

$$\frac{[B_{11}, u_{11}] \quad \dots \quad [B_{km_k}, u_{km_k}]}{[A, w]}$$

The formal system, associated with the grammar \mathcal{G} , is also denoted by \mathcal{G} . Define $\mathfrak{L}_{\mathcal{G}}(X) \coloneqq \{w \mid \mathcal{G} \vdash [X,w]\}$ and $\mathfrak{L}(\mathcal{G}) \rightleftharpoons \mathfrak{L}_{\mathcal{G}}(S)$ (" \rightleftharpoons " here and further means "equals by definition"). $\mathfrak{L}(\mathcal{G})$ is the *language generated by* \mathcal{G} .

Example 1. Consider the following conjunctive grammar (here small letters stand for terminal symbols, capital stand for nonterminal ones; S is the start symbol):

$$S \rightarrow aAB \& aDC$$

$$A \rightarrow aA$$

$$A \rightarrow a$$

$$B \rightarrow bBc$$

$$B \rightarrow b$$

$$C \rightarrow cC$$

$$C \rightarrow c$$

$$D \rightarrow aDb$$

$$D \rightarrow b$$

This grammar generates the language $\{a^{n+1}b^{n+1}c^n \mid n \ge 1\}$ as an intersection of two context-free languages. For example, the word $aaabbbcc = a^3b^3c^2$ is generated in the following way: first we derive [S, aaabbbcc] from [a, a], [A, aa], [B, bbbcc], [a, a], [D, aabbb], and [C, cc]. The pair [a, a] is an axiom; the others are derived as follows:

For technical reasons we also consider an enlarged version of this deduction system, called \mathcal{G}_{cut} . We allow nonterminal symbols to appear in the second components of the pairs (derivable objects in it are of the form $[X, \omega]$, where $X \in \Sigma \cup N$ and $\omega \in (\Sigma \cup N)^*$) and add new axioms [A, A] for all $A \in N$ and the cut rule:

$$\frac{[B,\tau] \quad [A,\omega_1 B \omega_2]}{[A,\omega_1 \tau \omega_2]}$$

A trivial "cut elimination theorem" holds:

Lemma 1. If $A \in N \cup \Sigma$, $w \in \Sigma^*$, then $\mathcal{G}_{cut} \vdash [A, w]$ if and only if $\mathcal{G} \vdash [A, w]$.

Proof. The "if" part is obvious. For the "only if" part, we prove that every pair, derivable in \mathcal{G}_{cut} , is derivable without applying the cut rule (therefore, as w does not contain nonterminal symbols, they do not occur in the derivation, thus this derivation is valid in the original system). Let $[B, \tau]$ and $[A, \omega_1 B\omega_2]$ be derivable without applying the cut rule. Prove that $[A, \omega_1 \tau \omega_2]$ also has a cut-free proof. Proceed by induction on the derivation of $[A, \omega_1 B\omega_2]$. If it is an axiom, then ω_1 and ω_2 is empty, B = A, and our goal coincides with the left premise, $[B, \tau]$. If $[A, \omega_1 B\omega_2]$ is derived using an inference rule, then we can perform the substitution of τ for B in the premises of this rule, and apply the induction hypothesis.

2 Greibach Normal Form

Consider only languages without the empty word.

A conjunctive grammar is in *Greibach normal form* (a generalisation of Greibach normal form for context-free grammars [3]), if all the rules are of the form $A \to a\beta_1 \& \ldots \& a\beta_k, a \in \Sigma, \beta_j \in N^+$ or of the form $A \to a, a \in \Sigma$.

The question remains open, whether every conjunctive grammar can be transformed into this form. However, it is true for languages over the one-letter alphabet, as shown by Okhotin and Reitwießner [11]. Therefore, conjunctive grammars in Greibach normal form can capture some languages that are not context-free or even finite intersections of those, since the language $\{a^{4^n} \mid n \ge 1\}$ is generated by a conjunctive grammar found by Jeż [4].

Example 2. The grammar from Example 1 can be easily transformed into Greibach normal form: $S \rightarrow a AB \& a DC$

$$S \rightarrow aAB \otimes aD$$

$$A \rightarrow aA$$

$$A \rightarrow a$$

$$B \rightarrow bBU$$

$$B \rightarrow b$$

$$U \rightarrow c$$

$$C \rightarrow cC$$

$$C \rightarrow cC$$

$$D \rightarrow aDV$$

$$D \rightarrow b$$

$$V \rightarrow b$$

3 Multiplicative-Additive Lambek Calculus

In this section we define an extension of the Lambek calculus (introduced in [7]) with two new connectives, *additive conjunction* and *disjunction*. The additive

(intersective) conjunction was already introduced by Lambek [8], and the whole calculus was considered by Kanazawa [5]. We shall call this calculus **MALC**, as in [6], but use the Lambek-style notation for connectives.

A countable set $\Pr = \{p_1, p_2, p_3, ...\}$ is called the set of *primitive types*. *Types* of **MALC** are built from primitive types with five binary connectives: \cdot (multiplication, product conjunction), \setminus (left division), / (right division), \cap (intersection, additive conjunction), \cup (union, additive disjunction). We denote types with capital Latin letters and their finite sequences (possibly empty) with capital Greek ones; Λ stands for the empty sequence. *Sequents* (derivable objects) of **MALC** are of the form $\Pi \rightarrow C$.

Axioms: $A \to A$.

Rules of inference:

$$\begin{array}{ll} \frac{A \, \Pi \rightarrow B}{\Pi \rightarrow A \setminus B} \ (\rightarrow \setminus), \ \Pi \neq \Lambda; & \frac{\Pi \rightarrow A \quad \Gamma B \, \Delta \rightarrow C}{\Gamma \, \Pi \ (A \setminus B) \, \Delta \rightarrow C} \ (\setminus \rightarrow); \\ \frac{\Pi A \rightarrow B}{\Pi \rightarrow B / A} \ (\rightarrow /), \ \Pi \neq \Lambda; & \frac{\Pi \rightarrow A \quad \Gamma B \, \Delta \rightarrow C}{\Gamma \ (B / A) \, \Pi \, \Delta \rightarrow C} \ (/ \rightarrow); \\ \frac{\Gamma \rightarrow A \quad \Delta \rightarrow B}{\Gamma \Delta \rightarrow A \cdot B} \ (\rightarrow \cdot); & \frac{\Gamma A B \, \Delta \rightarrow C}{\Gamma \ (A \cdot B) \, \Delta \rightarrow C} \ (- \rightarrow); \\ \frac{\Gamma \rightarrow A_1 \quad \Gamma \rightarrow A_2}{\Gamma \rightarrow A_1 \cap A_2} \ (\rightarrow \cap); & \frac{\Gamma A_i \, \Delta \rightarrow C}{\Gamma \ (A_1 \cap A_2) \, \Delta \rightarrow C} \ (\cap \rightarrow)_i, i = 1, 2; \\ \frac{\Gamma \rightarrow A_i}{\Gamma \rightarrow A_1 \cup A_2} \ (\rightarrow \cup)_i, i = 1, 2; & \frac{\Gamma A_1 \, \Delta \rightarrow C}{\Gamma \ (A_1 \cup A_2) \, \Delta \rightarrow C} \ (\cup \rightarrow); \\ \frac{\Pi \rightarrow A \quad \Gamma A \, \Delta \rightarrow C}{\Gamma \, \Pi \, \Delta \rightarrow C} \ (\cup \rightarrow); \\ \frac{\Pi \rightarrow A \quad \Gamma A \, \Delta \rightarrow C}{\Gamma \, \Pi \, \Delta \rightarrow C} \ (\cup). \end{array}$$

The cut rule is eliminable using the standard technique [7].

The fragment without \cap and \cup is the ordinary (multiplicative) Lambek calculus, called **MLC** or **L**. We also consider fragments of **MALC** with other restrictions of the set of connectives: **MALC**(/, \cap), **MALC**(/, \cap), **MLC**(/).

4 Categorial Grammars

A **MALC**-grammar is a triple $\mathscr{G} = \langle \Sigma, H, \rhd \rangle$, where Σ is a finite alphabet, $H \in \text{Tp}$, and \triangleright is a finite correspondence between Tp and Σ ($\triangleright \subset \text{Tp} \times \Sigma$). The language generated by \mathscr{G} is the set of all nonempty words $a_1 \dots a_n$ over Σ for which there exist types B_1, \dots, B_n such that **MALC** $\vdash B_1 \dots B_n \to H$ and $B_i \triangleright a_i$ for all $i \in \{1, \dots, n\}$. We denote this language by $\mathfrak{L}(\mathscr{G})$.

The notions of $MALC(/, \cap)$ -, $MALC(/, \cdot, \cap)$ -, MLC-, and MLC(/)-grammar are defined similarly.

As shown by Gaifman [1] and Buszkowski [2], any context-free language without the empty word is generated by an $\mathbf{MLC}(/)$ -grammar. On the other hand, any language generated by an \mathbf{MLC} -grammar is context-free (Pentus [12]). Kanazawa [5] proved that any finite intersection of context-free languages is generated by a $MALC(/, \cap)$ -grammar (therefore such grammars go beyond context-free). No generalisation of Pentus' theorem for MALC is yet known.

Theorem 1. If a language without the empty word is generated by a conjunctive grammar in Greibach normal form, then this language is generated by a $MALC(/, \cdot, \cap)$ -grammar.

5 The Construction

Given a conjunctive grammar $\mathcal{G} = \langle N, \Sigma, \mathcal{P}, S \rangle$ in Greibach normal form, we shall construct a **MALC** $(/, \cdot, \cap)$ -grammar \mathcal{G} , such that $\mathfrak{L}(\mathcal{G}) = \mathfrak{L}(\mathcal{G})$.

In order to avoid notation collisions, further we shall use the following naming convention (all these letters can also be decorated with numerical or other indices):

Letter	Range
A, B, S	N (nonterminal symbols of \mathcal{G})
a	Σ (terminal symbols)
x	$N\cup\Sigma$
w, u	Σ^* (strings of terminal symbols)
β	N^+ (strings of nonterminal symbols)
$ au, \omega$	$(N\cup\Sigma)^*$
p	$\Pr(\text{primitive types of } \mathbf{MALC})$
E, F, G, P	Tp (types of \mathbf{MALC})
Γ, Φ, Ψ	Tp^* (sequences of types)

With every $A \in N$ we associate a distinguished primitive type p_A . For $\beta = B_1 \dots B_m$ let $P_\beta \rightleftharpoons p_{B_1} \dots p_{B_m}$ (multiplication is associative, so we can omit the brackets).

Since intersection in **MALC** is commutative and associative, we can use intersections of nonempty sets of types, not bothering about order and brackets: $\bigcap_{j=1}^{k} E_j$ stands for $E_1 \cap \ldots \cap E_k$, and if $\mathcal{M} = \{E_1, \ldots, E_k\}$, then $\bigcap \mathcal{M} \rightleftharpoons E_1 \cap \ldots \cap E_k$. If $\mathcal{M} = \{E\}$, then $\bigcap \mathcal{M} \rightleftharpoons E$.

For every $a \in \Sigma$ let

$$\mathcal{M}_a \coloneqq \{ p_A / \left(\bigcap_{j=1}^k P_{\beta_j} \right) \mid (A \to a\beta_1 \& \dots \& a\beta_k) \in \mathcal{P} \} \cup \{ p_A \mid (A \to a) \in \mathcal{P} \}.$$

Let $G_a \cong \bigcap \mathcal{M}_a$. For $A \in N$ let $G_A \cong p_A$. The following holds due to the $(\cap \rightarrow)$ rule:

Lemma 2. If $E \in \mathcal{M}_a$ and $\mathbf{MALC} \vdash \Phi E \Psi \rightarrow F$, then $\mathbf{MALC} \vdash \Phi G_a \Psi \rightarrow F$.

For $\omega = x_1 \dots x_n \in (N \cup \Sigma)^+$ let $\Gamma_{\omega} \rightleftharpoons G_{x_1} \dots G_{x_n}$.

Lemma 3. If $\mathcal{G} \vdash [A, w]$, then $\mathbf{MALC} \vdash \Gamma_w \to p_A$.

Proof. We proceed by induction on the length of w. The base case (w = a) corresponds to an application of a rule of the form $A \to a$ to the [a, a] axiom (this is the only way to derive [A, a]). In this case we have $p_A \in \mathcal{M}_a$, therefore by Lemma 2 we get **MALC** $\vdash G_a \to p_A$, and $\Gamma_w = G_a$.

Now let w contain at least two symbols and the last step of the derivation of [A, w] be an application of the rule $A \to a\beta_1 \& \ldots \& a\beta_k$. Then w = aw', and for every $j \in \{1, \ldots, k\}$, if $\beta_j = B_{j1} \ldots B_{jm_j}$, then $w' = u_{j1} \ldots u_{jm_j}$ and for every $i = \{1, \ldots, m_j\}$ we have $\mathcal{G} \vdash [B_{ji}, u_{ji}]$. Therefore, by induction hypothesis, $\mathbf{MALC} \vdash \Gamma_{u_{ji}} \to p_{B_{ji}}$, whence $\mathbf{MALC} \vdash \Gamma_{w'} \to P_{\beta_j}$ for every j. Applying the $(\to \cap)$ rule k times we get

$$\mathbf{MALC} \vdash \Gamma_{w'} \to \bigcap_{j=1}^k P_{\beta_j},$$

and, finally, by $(/ \rightarrow)$,

$$\mathbf{MALC} \vdash p_A / \left(\bigcap_{j=1}^k P_{\beta_j}\right) \Gamma_{w'} \to p_A$$

Since $p_A / (\bigcap_{j=1}^k P_{\beta_j}) \in \mathcal{M}_a$, by Lemma 2 we have $\mathbf{MALC} \vdash G_a \Gamma_{w'} \to p_A$, and $G_a \Gamma_{w'} = \Gamma_w$.

Before proving the inverse statement, we shall prove two technical lemmata:

Lemma 4. MALC $\vdash \Phi \rightarrow \bigcap_{j=1}^{k} P_{\beta_j}$ if and only if MALC $\vdash \Phi \rightarrow P_{\beta_j}$ for every $j \in \{1, \ldots, k\}$.

Proof. The "if" part is just k applications of $(\rightarrow \cap)$. The "only if" part is proved using the cut rule (for every j_0):

$$\frac{\Gamma \to \bigcap_{j=1}^k P_{\beta_j} \quad \bigcap_{j=1}^k P_{\beta_j} \to P_{\beta_{j_0}}}{\Gamma \to P_{\beta_{j_0}}} \quad (\text{cut})$$

Lemma 5. If $\omega \in (N \cup \Sigma)^+$, $\beta = B_1 \dots B_m \in N^+$, and **MALC** $\vdash \Gamma_{\omega} \to P_{\beta}$, then there exist such $\tau_1, \dots, \tau_m \in (N \cup \Sigma)^+$, that $\omega = \tau_1 \dots \tau_m$ and **MALC** $\vdash \Gamma_{\tau_i} \to p_{B_i}$ for every $i \in \{1, \dots, m\}$.

Proof. We can rearrange the derivation, so that the applications of $(\rightarrow \cdot)$ will be in the bottom (they are interchangeable with $(\cap \rightarrow)$ and $(/ \rightarrow)$, and these two are the only ones that can be applied below $(\rightarrow \cdot)$). Now the statement of the lemma is obvious.

Lemma 6. If MALC $\vdash \Gamma_{\omega} \rightarrow p_A$, then $\mathcal{G}_{cut} \vdash [A, \omega]$.

Proof. Induction by the length of ω . If $\omega = a$, then the only possible case is $p_A \in \mathcal{M}_a$. Then $(A \to a) \in \mathcal{P}$, and $\mathcal{G}_{\text{cut}} \vdash [A, a]$.

Now let ω contain at least two letters. Consider the lowest application of $(/ \rightarrow)$ in the derivation of $\Gamma_{\omega} \rightarrow p_A$. Beneath this application there are only applications of $(\cap \rightarrow)$ —the ones that open the type to which $(/ \rightarrow)$ is applied, and the ones that deal with other types in Γ_{ω} . We can transform the derivation so that the latter will be applied before the application of $(/ \rightarrow)$. Then we have $\omega = \omega_1 a \tau \omega_2, p_{A'} / (\bigcap_{i=1}^k P_{\beta_i}) \in \mathcal{M}_a$, and the derivation step looks as follows:

$$\frac{\Gamma_{\tau} \to \bigcap_{j=1}^{k} P_{\beta_j} \quad \Gamma_{\omega_1} p_{A'} \Gamma_{\omega_2} \to p_A}{\Gamma_{\omega_1} p_{A'} / (\bigcap_{j=1}^{k} P_{\beta_j}) \Gamma_{\tau} \Gamma_{\omega_2} \to p_A} \ (/ \to)$$

Then, by Lemma 4, **MALC** $\vdash \Gamma_{\tau} \rightarrow P_{\beta_j}$ for every $j \in \{1, \ldots, k\}$. By Lemma 5, if $\beta_j = B_{j1} \ldots B_{jm_j}$, $\tau = \tau_{j1} \ldots \tau_{jm_j}$, and **MALC** $\vdash \Gamma_{\tau_{ji}} \rightarrow p_{B_{ji}}$ (for every j and i in the ranges). By induction hypothesis, $\mathcal{G}_{\text{cut}} \vdash [B_{ji}, \tau_{ji}]$, and, adding [a, a], we can apply the rule for $A' \rightarrow a\beta_1 \& \ldots \& a\beta_k$, therefore $\mathcal{G}_{\text{cut}} \vdash [A', a\tau]$.

By induction hypothesis for the right premise of the $(/ \rightarrow)$ rule, $\mathcal{G}_{\text{cut}} \vdash [A, \omega_1 A' \omega_2]$. Finally, applying the cut rule to $[A', a\tau]$ and $[A, \omega_1 A' \omega_2]$, we get $[A, \omega_1 a \tau \omega_2] = [A, \omega]$.

Now we are ready to define $\mathscr{G} = \langle \Sigma, \rhd, H \rangle$. Let $H = p_S$, and $E \rhd a$ if and only if $E = G_a$. If $w \in \mathfrak{L}(\mathcal{G})$, then $\mathcal{G} \vdash [S, w]$, and, by Lemma 3, **MALC** $\vdash \Gamma_w \to p_S$, whence $w \in \mathfrak{L}(\mathscr{G})$. Conversely, if $w \in \mathfrak{L}(\mathscr{G})$, then **MALC** $\vdash \Gamma_w \to p_S$. By Lemma 6 we get $\mathcal{G}_{cut} \vdash [S, w]$, and by Lemma 1 $\mathcal{G} \vdash [S, w]$. Hence, $w \in \mathfrak{L}(\mathcal{G})$.

Note that in \mathscr{G} every $a \in \Sigma$ is associated with only one type (such grammars are called *grammars with single type assignment* or *deterministic grammars*). Having the intersection connective, it is usually easy to make our grammar deterministic (cf. [5]); for the pure Lambek calculus the fact that any context-free language is generated by a deterministic **MLC**-grammar is not obvious, but still valid, as shown by Safiullin [13].

Example 3. This construction gives the following **MALC**-grammar equivalent to the grammar from Example 2:

$$\begin{aligned} a &\triangleright p_A \cap (p_A / p_A) \cap (p_D / (p_D \cdot p_V)) \cap (p_S / ((p_A \cdot p_B) \cap (p_D \cdot p_C))) \\ b &\triangleright p_B \cap p_D \cap p_V \cap (p_B / (p_B \cdot p_U)) \\ c &\triangleright p_C \cap p_U \cap (p_C / p_C) \end{aligned}$$

Acknowledgements

I am grateful to Prof. Mati Pentus and Alexey Sorokin for fruitful discussions. I am also grateful to Ivan Zakharyashchev for bringing my attention to conjunctive grammars.

This research was supported by the Russian Foundation for Basic Research (grants 11-01-00281-a and 12-01-00888-a) and by the Presidential Council for Support of Leading Scientific Schools (grant NŠ 5593.2012.1).

References

- Bar-Hillel, Y., Gaifman, C., Shamir, E.: On the categorial and phrasestructure grammars. Bull. of the Research Council of Israel, Sect. F, 9F, 1–16 (1960)
- [2] Buszkowski, W.: The equivalence of unidirectional Lambek categorial grammars and context-free languages. Zeitschr. f
 ür math. Logik und Grundl. der Math., 31, 369–384 (1985)
- [3] Greibach, S.: A new normal-form theorem for context-free phrase structure grammars. Journal of the ACM, 12, 42–52 (1965)
- [4] Jeż, A.: Conjunctive grammars can generate non-regular unary languages. International Journal of Foundations of Computer Science, 19(3), 597–615 (2008)
- [5] Kanazawa, M.: The Lambek calculus enriched with additional connectives. Journal of Logic, Language and Information, 1, 141–171 (1992)
- [6] Kanazawa, M.: Lambek calculus: recognizing power and complexity. In: Gerbrandy, J., Marx, M., de Rijke, M., Venema, Y. (eds.) JFAK. Essays Dedicated to Johan van Benthem on the Occasion of his 50th Birthday. Vossiuspers, Amsterdam University Press (1999)
- [7] Lambek, J.: The mathematics of sentence structure. American Math. Monthly, 65(3), 154–170 (1958)
- [8] Lambek, J.: On the calculus of syntactic types. In: Jakobson, R. (ed.) Structure of Language and its Mathematical Aspects. Amer. Math. Soc. (1961)
- [9] Okhotin, A.: Conjunctive grammars. Journal of Automata, Languages and Combinatorics, 6(4), 519–535 (2001)
- [10] Okhotin, A.: The dual of concatenation. Theor. Comput. Sci., 345 (2–3), 425–447 (2005)
- [11] Okhotin, A., Reitwießner, C.: Conjunctive grammars with restricted disjunction. Theor. Comput. Sci., 411 (26–28), 2559–2571 (2010)
- [12] Pentus, M.: Lambek grammars are context-free. In.: 8th Annual IEEE Symposium on Logic in Computer Science, pp. 429–433. IEEE Computer Society Press, Los Alamitos (1993)
- [13] Safiullin, A. N.: Derivability of admissible rules with simple premises in the Lambek calculus. Moscow University Math. Bull., 62(4), 72–76 (2007)
- [14] Szabari, A.: Alternujúce Zásobníkové Automaty (Alternating Pushdown Automata), in Slovak, diploma work (M. Sc. thesis), University of Košice, Czechoslovakia, 45 pp. (1991)