Соотношение между грамматиками Ламбека и иерархией формальных грамматик Хомского

Александр Грефенштейн

НИУ ВШЭ

1 марта 2023 г.

AB грамматики

Исторически первым математическим формализмом, придуманным для моделирования естественных языков, в которых порядок слов имеет ключевое значение, стали классические категориальные грамматики или AB грамматики.[1]

Выражения (формулы) или синтаксические типы для AB грамматик строятся следующим образом:

$$L ::= P \mid (L\backslash L) \mid (L/L),$$

где P — заранее фиксированный список примитивных синтаксических типов, среди которых есть выделенный тип S.

[1] Bar-Hillel, Y.: A quasi arithmetical notation for syntactic description. Language 29, 47-58 (1953).

AB грамматики

Пусть Σ — конечное множество терминалов (слов). Функция $\mathtt{Lex}$ будет присваивать каждому $\sigma \in \Sigma$ конечное множество синтаксических типов. Итак, AB грамматика — это тройка $(\Sigma, \mathtt{Lex}, S)$.

Будем говорить, что предложение $\sigma_1 \dots \sigma_n$, составленное из слов из Σ имеет тип u, если для каждого терминала σ_i существует тип t_i из $\mathtt{Lex}(\sigma_i)$ такой, что $t_1 \dots t_n \longrightarrow u$ в соответствии со следующими правилам редукции:

$(\backslash_e)$ $u\,(u\backslash v) \longrightarrow v$;

$(/_e)$ $(v/u)\,u \longrightarrow v$.

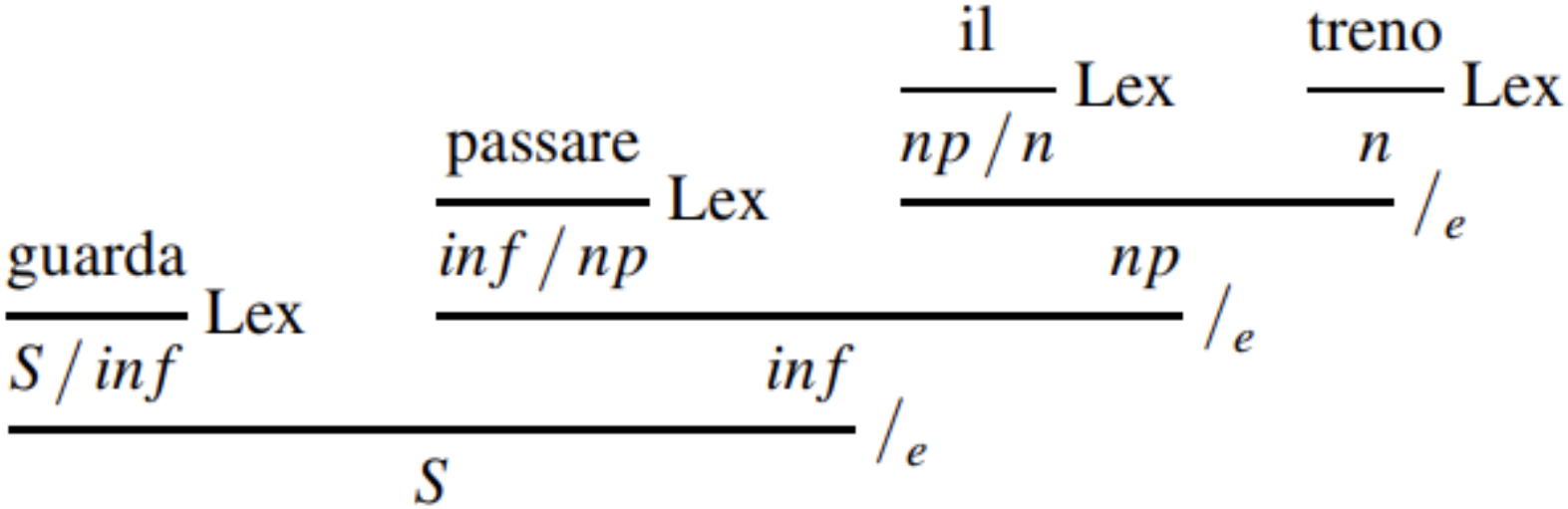
Тогда языком, порождаемым AB грамматикой, называется множество предложений, имеющих тип S.

Пример AB грамматики и дерева разбора

Word	Type(s)	Translation
cosa	$(S / (S / np))$	*what*
guarda	(S / inf)	*he/she watches*
passare	(inf / np)	*passing by*
il	(np / n)	*the*
treno	n	*train*

Пример *AB* грамматики и дерева разбора

Word	Type(s)	Translation
cosa	$(S / (S / np))$	*what*
guarda	(S / inf)	*he/she watches*
passare	(inf / np)	*passing by*
il	(np / n)	*the*
treno	n	*train*



Контекстно-свободные грамматики

Другим способом описания синтаксиса естественных языков является контекстно-свободная грамматика, т.е. формальная грамматика (T, NT, P, S), в которой правила продукции имеют следующий вид:

$$X \longrightarrow W,$$

где $X \in NT$, $W \in (NT \cup T)^+$.[2]

[2] В рамках данного доклада мы не рассматриваем грамматики, допускающие пустое слово.

Контекстно-свободные грамматики

Другим способом описания синтаксиса естественных языков является контекстно-свободная грамматика, т.е. формальная грамматика (T, NT, P, S), в которой правила продукции имеют следующий вид:

$$X \longrightarrow W,$$

где $X \in NT$, $W \in (NT \cup T)^+$.[2]

Простой пример:

$$S \to (S \wedge S) \mid \neg S \mid p \mid q \mid r.$$

[2] В рамках данного доклада мы не рассматриваем грамматики, допускающие пустое слово.

AB грамматики v.s. контекстно-свободные

Оказывается, что два рассмотренных выше формализма по существу описывают одно и то же:

Теорема(от CFG к AB)

Каждая контекстно-свободная грамматика без пустого слова слабо эквивалентна некоторой AB грамматике, содержащей только типы вида X, X/Y или $(X/Y)/Z$.

Схема доказательства

Сначала приведём грамматику к сильной нормальной форме Грейбах. Далее построим по этой нормальной форме G_C подходящую AB грамматику.

AB грамматики v.s. контекстно-свободные

AB грамматика G будет такая:

- Терминалы G будут такие же, как и у G_C;
- Примитивными типами G будут нетерминальные символы G_C.
- Если в G_C было правило $X \longrightarrow aX_1 \dots X_n$, то $((\dots((X/X_n)/X_{n-1})/\dots)/X_2)/X_1 \in \mathtt{Lex}(a)$.

Можно убедиться в том, что деревья разбора у G_C и G изоморфны и действительно используются только заявленные типы.

Обратное утверждение так же будет верно, но уже не понадобится нам вдальнейшем.

От AB грамматик к грамматикам Ламбека

Забегая вперёд, известно, что AB грамматики и грамматики Ламбека порождают один и тот же класс языков.

Однако у грамматик Ламбека есть важное преимущество: они позволяют выражать некоторые языковые феномены более естественно и компактно.

Так, в лексиконе, приведенном несколько слайдов назад, рассмотрим предложение «cosa guarda passare?». Оно имеет тип

$$(S/(S/np))\,(S/inf)\,(inf/np)\,,$$

и, очевидно, не лежит в порожденном такой AB грамматикой языке. Однако в исчислении Ламбека с тем же лексиконом это предложение будет иметь тип S. За это ответственно наличие выводов из гипотез — правые правила для каждого из делений.

Категориальные грамматики Ламбека

Синтаксические типы для грамматик Ламбека расширяют типы для AB грамматик следующим образом:

$$Lp ::= P \mid (Lp \backslash Lp) \mid (Lp / Lp) \mid (Lp \cdot Lp).$$

Как и AB грамматики, грамматики Ламбека определяются как соответствующая тройка $(\Sigma, \mathrm{Lex}, S)$. Однако порождающий язык определяется уже совсем иначе.

Предложение $\sigma = \sigma_1 \dots \sigma_n$ лежит в порожденным грамматикой Ламбека языке, если для каждого терминала σ_i существует тип t_i из $\mathrm{Lex}(\sigma_i)$ такой, что секвенция $t_1, \dots, t_n \vdash S$ выводима в секвенциальном исчислении Ламбека.

Секвенциальное исчисление Ламбека

$$\frac{\Gamma, B, \Gamma' \vdash C \quad \Delta \vdash A}{\Gamma, \Delta, A \setminus B, \Gamma' \vdash C} \setminus_h \qquad \frac{A, \Gamma \vdash C}{\Gamma \vdash A \setminus C} \setminus_i \quad \Gamma \neq \varepsilon$$

$$\frac{\Gamma, B, \Gamma' \vdash C \quad \Delta \vdash A}{\Gamma, B / A, \Delta, \Gamma' \vdash C} /_h \qquad \frac{\Gamma, A \vdash C}{\Gamma \vdash C / A} /_i \quad \Gamma \neq \varepsilon$$

$$\frac{\Gamma, A, B, \Gamma' \vdash C}{\Gamma, A \bullet B, \Gamma' \vdash C} \bullet_h \qquad \frac{\Delta \vdash A \quad \Gamma \vdash B}{\Delta, \Gamma \vdash A \bullet B} \bullet_i$$

$$\frac{\Gamma \vdash A \quad \Delta_1, A, \Delta_2 \vdash B}{\Delta_1, \Gamma, \Delta_2 \vdash B}\, cut \qquad \frac{}{A \vdash A}\, axiom$$

Секвенциальное исчисление Ламбека: немного фактов

- Можно рассматривать только те выводы, где аксиомы состоят только из примитивных типов;
- Число позитивных и негативных вхождений всякого примитивного типа p в доказуемую секвенцию одинаково;
- Сечение устранимо;
- Выполнено свойство подформульности;
- При отказе от конкатенции типов и ограничения единицей порядка присваемых типов словам, выводимость секвенции $t_1, \ldots, t_n \vdash u$ будет эквивалентна $t_1, \ldots, t_n \to u$ в смысле AB грамматики. Следовательно, язык, задаваемый грамматикой Ламбека $G = (\Sigma, \mathrm{Lex}, S)$ будет совпадать с языком, генерируем грамматикой G в смысле AB грамматики.

Грамматики Ламбека v.s. контекстно-свободные

По модулю последнего факта с предыдущего слайда от контекстно-свободных грамматик прийти к грамматикам Ламбека довольно просто: поскольку мы уже знаем, что контекстно-свободные грамматики сводятся к AB грамматикам, используя даже ограниченный вид типов, то заметим, что эти типы порядка не больше 1. Следовательно, имеет место

Теорема(от *CFG* к Ламбеку)

Каждая свободная от пустого слова контекстно-свободная грамматика G слабо эквивалетна некоторой грамматике Ламбека.

Грамматики Ламбека v.s. контекстно-свободные

Идея доказательства обратной теоремы примерно следующая: пусть мы можем сформулировать исчисление Ламбека используя только конечный набор аксиом и правило cut. Поскольку мы имеем свойство подформульности, то для выводимости секвенции $t_1, \ldots, t_n \vdash S$ нам стоит заботиться только о типах, которые появляются в аксиомах.

Переписывая каждую аксиому $X_1, \ldots, X_n \vdash X$ в виде правила продукции

$$X \longrightarrow X_1, \ldots X_n,$$

мы относительно легко получим, что этой грамматикой задаётся тот же самый язык. Причём о cut даже думать не надо — это просто правило подстановки, которое и так неявно присутствует в формальных грамматиках. Однако так просто это не сработает.

Грамматики Ламбека v.s. контекстно-свободные

Пусть A — формула, через $|A|$ будем обозначать количество примитивных типов, входящих в A и называть $|A|$ размером формулы A.

Наша главная цель теперь в том, чтобы показать, что для фиксированного натурального m существует конечное множество доказуемых секвенций $\mathtt{AX}(m)$ такое, что всякая доказуемая секвенция, в которой формулы имеют размер меньше m, выводимы из $\mathtt{AX}(m)$ используя только правило $\mathtt{cut}$.

Неформально говоря, доказав нашу главную цель мы сразу получим, что язык, задаваемый некоторой грамматикой Ламбека G генерируется подходящей контекстно-свободной грамматикой с правилами редукции из $\mathtt{AX}(m)$.

Грамматики Ламбека v.s. контекстно-свободные

Обозначим через $\rho_p(\Delta)$ число вхождений примитивного типа p в формулы из Δ.

Первым шагом к реализации этой идеи будет

Интерполяция

Пусть секвенция $\Gamma, \Delta, \Theta \vdash C$ доказуема и $\Delta \neq \varepsilon$. Тогда существует такая формула I, называемая интерполянтом, что:

1. $\Delta \vdash I$;
2. $\Gamma, I, \Theta \vdash C$;
3. $\rho_p(I) \leqslant \rho_p(\Delta)$;
4. $\rho_p(I) \leqslant \rho_p(\Gamma, \Theta, C)$,

для любого примитивного типа p.

Грамматики Ламбека v.s. контекстно-свободные

Набросок доказательства

Индукция по выводу без сечений. Случай аксиомы тривиален. Далее рассматриваем последнее применённое правило. Ограничимся рассмотрением правила

$$\frac{\Pi \vdash X \quad \Phi, Y, \Psi \vdash C}{\Phi, \Pi, X \setminus Y, \Psi \vdash C} \setminus_h$$

- Пусть Δ включено либо в Π, либо в Φ, либо в Ψ и по И.П. имеем интерполянт I для Δ в соответствующей посылке. Тогда он нам и подходит.
- Пусть Δ включено в Φ, Π, т.е. $\Delta = \Delta', \Delta''$; $\Phi = \Phi', \Delta'$; $\Pi = \Delta'', \Pi''$. По И.П. I' будет интерполянтом для Δ' в $\Phi', \Delta', Y, \Psi \vdash C$, а I'' для Δ'' в $\Delta'', \Pi'' \vdash X$. Тогда $I' \cdot I''$ искомый интерполянт для Δ.

Грамматики Ламбека v.s. контекстно-свободные

- Пусть Δ включено в $\Pi, X\backslash Y, \Psi$, т.е. $\Pi = \Pi', \Pi''$; $\Psi = \Psi', \Psi''$; $\Delta = \Pi'', X\backslash Y, \Psi'$. По И.П. I' будет интерполянтом для Π' в $\Pi', \Pi'' \vdash X$, а I'' для Y, Ψ' в $\Phi, Y, \Psi', \Psi'' \vdash C$. Тогда $I'\backslash I''$ искомый интерполянт для Δ.
- Пусть Δ включено в $\Phi, \Pi, X\backslash Y, \Psi$, т.е. $\Phi = \Phi', \Phi''$; $\Psi = \Psi', \Psi''$; $\Delta = \Phi'', \Pi, X\backslash Y, \Psi'$. По И.П. I будет интерполянтом для Φ'', Y, Ψ' в $\Phi', \Phi'', Y, \Psi', \Psi'' \vdash C$. Тогда сам I и является искомым интерполянтом для Δ.

Грамматики Ламбека v.s. контекстно-свободные

Для второго шага нам понадобится интерпретация типов исчисления Ламбека в свободной группе и корректность доказуемых секвенций относительно такой интерпретации.

Пусть $G = (M, \circ)$ — свободная группа над P, т.е. множество несократимых слов в алфавите P. Стандартная интерпретация $[\cdot]$ Lp в G определяется индукцией по построению типа:

- $[p] = p$;
- $[A \cdot B] = [A] \circ [B]$;
- $[A \backslash B] = [A]^{-1}[B]$;
- $[A / B] = [A][B]^{-1}$.

Теорема

Если секвенция $\Gamma \vdash \Delta$ доказуема, то $[\Gamma] = [\Delta]$.

Грамматики Ламбека v.s. контекстно-свободные

Пора начинать второй шаг!

Говорим, что секвенция $\Gamma \vdash C$ тонкая, если она доказуема и для всякого примитивного типа p $\rho_p(\Gamma, C)$ равно либо 0, либо 2.
Следующее утверждение очевидно по модулю устранения сечения и использованию примитивных аксиом.

Утверждение 1

Каждая доказуемая секвенция может быть получена из тонкой путём замены примитивных типов.

Грамматики Ламбека v.s. контекстно-свободные

Теперь докажем, что для тонких секвенций мы можем найти интерполянт фиксированного размера.

Утверждение 2

Пусть $\Gamma, \Delta, \Theta \vdash C$ — тонкая секвенция. Тогда существует формула B такая, что

1. Секвенции $\Delta \vdash B$ и $\Gamma, B, \Theta \vdash C$ тонкие;
2. $|B| = |||[\Delta]|||$.

Доказательство

Проводя обычную интерполяцию, мы имеем интерполянт B такой, что

1. $\Delta \vdash B$ и $\Gamma, B, \Theta \vdash C$ доказуемы;
2. $\rho_p(B) \leqslant \min(\rho_p(\Gamma, \Theta, C), \rho_p(\Delta))$.

Грамматики Ламбека v.s. контекстно-свободные

Докажем оценки на количество примитивных типов в секвенциях. Имеем:

$$0 \text{ или } 2 = \rho_p(\Gamma, \Delta, \Theta, C) = \rho_p(\Gamma, \Theta, C) + \rho_p(\Delta).$$

Тогда $\rho_p(B)$ равно 0 или 1 и мы имеем:

$$\rho_p(\Delta, B) = \rho_p(\Delta) + \rho_p(B) \leqslant \rho_p(\Gamma, \Delta, \Theta, C) + \rho_p(B) \leqslant 2 + 1.$$

Но так как $\Delta \vdash B$ доказуема, то $\rho_p(\Delta, B)$ чётно. Значит, не равно 3 и меньше 2.

Аналогичное рассуждение для $\rho_p(\Gamma, B, \Theta, C)$:

$$\rho_p(\Gamma, B, \Theta, C) = \rho_p(\Gamma, \Theta, C) + \rho_p(B) \leqslant \rho_p(\Gamma, \Delta, \Theta, C) + \rho_p(B).$$

Последнее не превосходит 3, откуда всё и получается.

Грамматики Ламбека v.s. контекстно-свободные

Докажем, что $|B| = ||[\Delta]||$.

Если p не входит в Δ, то его очевидно нет ни в B, ни в $[\Delta]$.

Если p входит в Δ один раз, то p также входит в $[\Delta]$ один раз, ибо ни с кем не может сократиться. Поскольку секвенция $\Delta \vdash B$ тонкая, p входит в B единожды.

Если p входит дважды в Δ, то p не входит в Γ, Θ, C, а, значит, и в B. По корректности в свободной группе $[\Gamma, \Delta, \Theta] = [C]$. Отсюда получаем, что $[\Delta] = [\Gamma]^{-1}[C][\Theta]^{-1}$. Но так как p не входит в Γ, Θ, C, p не входит в $[\Delta]$.

Грамматики Ламбека v.s. контекстно-свободные

Наконец, из утверждения 2 можно получить

Следствие 1

Пусть секвенция $A_1, \ldots, A_n \vdash A_{n+1}$ тонкая и для каждого i $|A_i| \leqslant m$. Тогда выполнено одно из следующих двух условий.

1. существует k и формула B такие, что $|B| \leqslant m$ и следующие секвенции тонкие:
   - $A_1, \ldots, A_{k-1}, B, A_{k+2}, \ldots, A_n \vdash A_{n+1}$;
   - $A_k, A_{k+1} \vdash B$.
2. существует формула B такая, что $|B| \leqslant m$ и следующие секвенции тонкие:
   - $B, A_n \vdash A_{n+1}$;
   - $A_1, \ldots, A_{n-1} \vdash B$.

Грамматики Ламбека v.s. контекстно-свободные

Теперь пора доказывать основную лемму и теорему Пентуса об эквивалентности всякой грамматики Ламбека некоторой контекстно-свободной грамматике.

Основная лемма

Пусть $A_1, \ldots, A_n \vdash A_{n+1}$ такая доказуемая секвенция, что $|A_i| \leqslant m$ для каждого i. Тогда эта секвенция доказуема, используя только правило cut, из доказуемых секвенций вида $U, V \vdash X$ или $U \vdash X$ таких, что $|U|, |V|, |X| \leqslant m$.

Доказательство

Индукция по n. При $n \leqslant 2$ доказывать нечего. По утверждению 1 наша секвенция получена из некоторой тонкой секвенции $A_1', \ldots, A_n' \vdash A_{n+1}'$. При этом очевидно, что $|A_i'| \leqslant m$ для всех i.

Грамматики Ламбека v.s. контекстно-свободные

Обозначим эту подстановку примитивных типов через σ. По следствию 1 существует $|B'| \leqslant m$ такая, что:

1. Секвенции $A'_1, \ldots A'_{k-1}, B', A'_{k+2}, \ldots, A'_n \vdash A'_{n+1}$ и $A'_k, A'_{k+1} \vdash B'$ тонкие. Пусть $B = \sigma(B')$. Тогда $|B| \leqslant m$. Возвращаясь к предыдущим примитивным типам, мы получаем, что секвенции $A_1, \ldots A_{k-1}, B, A_{k+2}, \ldots, A_n \vdash A_{n+1}$ и $A_k, A_{k+1} \vdash B$ доказуемы. По И.П. получаем, что секвенция

$$A_1, \ldots A_{k-1}, B, A_{k+2}, \ldots, A_n \vdash A_{n+1}$$

доказуема только с помощью правила `cut` из доказуемых секвенций правильного вида. Поскольку секвенция $A_k, A_{k+1} \vdash B$ уже имеет правильный вид, применяя `cut`, мы получаем заключение леммы.

2 Секвенции $B', A'_n \vdash A'_{n+1}$ и $A'_1, \ldots, A'_{n-1} \vdash B'$ тонкие. Далее рассуждения аналогичны пункту 1.

Теорема(от грамматик Ламбека к *CFG*)

Пусть $\mathtt{Lex}$ — лексикон некоторой грамматики Ламбека G_L и пусть m максимальный размер типов, присваемых словам G_L. Тогда язык, генерируемый G_L, совпадает с языком, генерируемым контекстно-свободной грамматикой G_C следующего вида:

Грамматики Ламбека v.s. контекстно-свободные

- Терминалы — все слова G_L;
- Нетерминалы — все типы A грамматики G_L такие, что $|A| \leqslant m$;
- Стартовый символ такой же, как и у G_L;
- $X \longrightarrow a$, если $a \in \mathtt{Lex}(a)$;
- $X \longrightarrow A$, если секвенция $A \vdash X$ доказуема;
- $X \longrightarrow AB$, если секвенция $A, B \vdash X$ доказуема.

Доказательство

Пусть $a_1 \dots a_n$ допускается грамматикой G_C. Тогда существуют типы $X_i \in \mathtt{Lex}(a_i)$ такие, что $S \longrightarrow X_1 \dots X_n$. Легко видеть, что вывод в G_C можно перестроить в вывод в G_L (обратить $\longrightarrow$ в $\vdash$), используя только правило `cut`.

Грамматики Ламбека v.s. контекстно-свободные

Пусть $a_1 \dots a_n$ допускается грамматикой G_L. Тогда существуют типы $X_i \in \mathtt{Lex}(a_i)$ такие, что секвенция $X_1, \dots X_n \vdash S$ доказуема. По основной лемме эта секвенция доказуема из доказуемых секвенций, соответствующих правилам продукции грамматики G_C, используя только правило `cut`. Индукцией по данному «`cut`-only» выводу можно убедиться, что он соответствует выводу в G_C.

Заключительные комментарии

- Сведение грамматики Ламбека к подходящей контекстно-свободной не полиномиально и особых надежд на полиномиальное сведение нет. Однако есть полиномиальное сведение для случая одного деления.
- Для грамматик Ламбека с пустым словом аналогичный теореме Пентуса результат также верен.