ОКРЕСТНОСТНЫЕ ГРАММАТИКИ И MODEL-THEORETIC SYNTAX

Владимир Борщев

ВИНИТИ РАН, Москва

borschev@linguist.umass.edu

Лет тридцать пять тому назад мы с М.В. Хомяковым предложили представлять формальные грамматики как системы аксиом, описывающие синтаксические структуры языка. Работа наша была благополучно забыта. Некоторое время тому назад на Западе появилось направление “Model-theoretic syntax”, развивающее аналогичные идеи. Цель доклада – сопоставить наши работы с этими новыми исследованиями. в частности с использованием модальной логики для описания синтаксических структур.

1. Введение

1.1. Тема

Аксиомы, *теории*, *модели*, *теоретико-модельный подход* – эти слова, а главное, понятия и методы, давно уже используются не только в логике, но и в лингвистике, в языках программирования и других областях, главным образом для описания семантики (Model-Theoretic Semantics).

Лет тридцать пять тому назад мы с моим соавтором М.В.Хомяковым предложили использовать этот подход для описания синтаксиса, т.е. представлять формальную грамматику как систему аксиом, описывающую синтаксические структуры языка. Опубликовав несколько работ и выступив на нескольких конференциях (Борщев, Хомяков 1970, 1973а, 1973б, Borščev, Xomjakov 1973, 1976, 1977), мы не встретили большого интереса и вскоре сами стали заниматься другими вещами. Работы наши были благополучно забыты. Но некоторое время тому назад направление “Model-Theoretic Syntax” появилось на Западе, и там развиваются аналогичные идеи (назову только несколько работ: Blackburn et al 1993, Blackburn and Meyer-Viol 1997, Rogers 1997, Pullum and Scholz 2001). Цель доклада – напомнить о наших работах и сопоставить их с этими новыми исследованиями, в частности, с использованием модальной логики для описания синтаксических структур.

Из-за ограниченности объема я ограничусь здесь изложением основных идей и примерами, не приводя формальных определений и результатов.

1.2. Исходный импульс и «выбор парадигмы»

Исходным импульсом для нас была работа Ю.А.Шрейдера по окрестностной модели языка (Шрейдер 1967), прежде всего его простые окрестностные грамматики.

В этих грамматиках, как и в грамматиках Хомского, язык понимается как множество слов в некотором алфавите. У каждого вхождения букв в слово рассматривались их окрестности. Простая окрестностная грамматика определялась как конечный набор допустимых окрестностей. Слово принадлежит языку, задаваемому такой грамматикой, если у каждого вхождения каждой буквы в это слово существует по крайней мере одна допустимая окрестность (принадлежащая данному набору).

Простые окрестностные грамматики действительно очень просты. Но они содержали важную идею, отличающую их от порождающих грамматик Хомского. Окрестностная грамматика не является *порожающей* процедурой, определяющей язык. Это *статическое* описание задаваемого языка, набор условий (ограничений), которому подчиняются слова этого языка. Способ описания этих условий, как ограничений на систему окрестностей, также представлялся очень естественным.

Мы обобщили идею Шрейдера и развили ее в рамках логико-алгебраической парадигмы. Мы предложили рассматривать язык как множество *конечных моделей* в некоторой сигнатуре[1], а грамматику – как множество *аксиом*, описывающих это множество моделей, т.е использовать *теоретико-модельный подход* для описания синтаксиса.

1.3. Наши основные идеи

Тексты. В классической математической лингвистике язык понимается как множество слов в некотором алфавите. Мы предложили рассматривать не только слова, но и синтаксические структуры, такие, как деревья непосредственных составляющих и деревья зависимостей, графы и т.п. Структуры такого типа мы называли *текстами*[2]. Большинство таких структур могут быть предствлены как конечные модели в подходящей сигнатуре.

Язык понимается как множество текстов.

Грамматика языка – это множество аксиом, каждая из которых выполняется для всех текстов языка. Так что язык, описываемый такой грамматикой, это *аксиоматизированный класс моделей*.

Мы раделили аксиомы в таких грамматиках на две главные части.

Первая часть аксиом определяет широкий класс текстов в подходящей сигнатуре, обладающих одинаковыми *общими* (*глобальными*) свойстами. Мы назвали такой класс текстов *знаковой системой*. Упоминавшиеся выше типы текстов являются примерами знаковых систем: слова в некотором алфавите (точнее, множество моделей, представляющих эти слова), деревья непосредственных составляющих, деревья с некоторыми дополнительными отношениями на их вершинах, формулы органической химии (химические графы) и т.п.

Чтобы определить конкретную знаковую систему, нужна подходящая сигнатура и набор аксиом, описывающих свойства соответствующих отношений из этом множестве.

Каждый язык – это подмножество некоторой знаковой системы. *Вторая часть аксиом* (составляющих грамматику) выделяет язык в некоторой знаковой системе, описывая *локальные* (*специфические*) свойства текстов этого языка.

Для описания этих локальных свойств мы обобщили идею *окрестностных грамматик* Шрейдера. Мы рассматривали *окрестности* элементов текста (скажем, окрестности вершин в деревьях непосредственных составляющих). Грамматика в узком смысле описывала возможную систему окрестностей для

---

[1] *Сигнатура* – это множество символов *отношений*. *Модель* в данной сигнатуре – это некоторое множество объектов (*несущее множество модели*), на котором каждому символу сигнатуры сопоставлено некоторое отношение.

[2] Я боюсь, что термин *текст* мог вызывать ложные ассоциации, связанные с «обычными» текстами – последовательностями предложений, параграфов, глав. Но латинское *textus* переводится как *структура*, *сплетение*, *ткань*.

каждого такого элемента. Т.е. грамматика – это формула, которая выполняется на такой системе окрестностей.

Мы распространили такого рода подход на описание системы переводов между языками.

В разделах 2 и 3 я опишу наш подход чуть подробнее и рассмотрю некоторые примеры, а в разделе 4 сравню его с идеями, предлагаемыми в недавних работах по теоретико-модельному синтаксису.

2. Примеры знаковых систем

2.1. Деревья непосредственных составляющих

Пример дерева непосредственных составляющих приведен на Рис. 1 и чуть более абстрактный пример – на рис. 2.

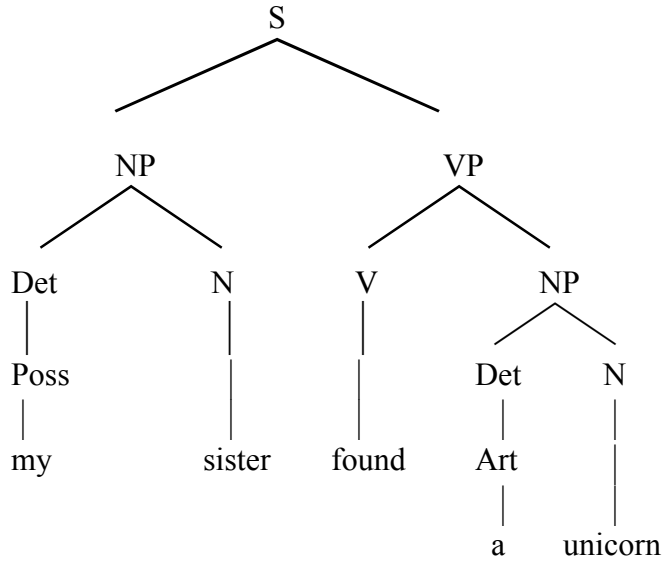

Рис. 1

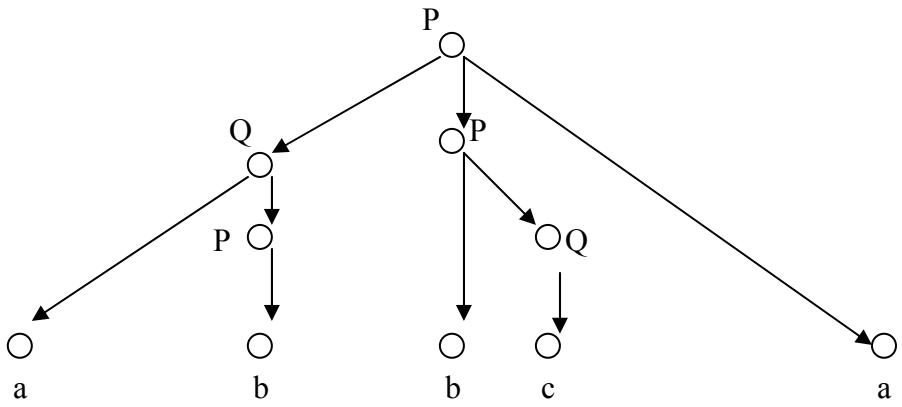

Рис. 2.

Чтобы представить такие деревья как конечные модели (т.е *тексты* в нашей терминологии), нужно описать их несущие множества и сигнатуру Ω_{ICTr} отношений, определенных на этих множествах.

Несущее множество дерева – это множество его вершин. Отношения удобно разделить на три части: $\Omega_{\mathrm{ICTr}} = \Omega_{\mathrm{Tr}} \cup \Omega_{\mathrm{Ord}} \cup \Omega_{\mathrm{Lbl}}$. Отношения из Ω_{Tr} представляют структуру «собственно» дерева, отношения из Ω_{Ord} отвечают за

порядок слева-направо, а отношения из Ω_{Lbl} представляют ярлыки (терминальные и нетерминальные символы), метящие вершины дерева.

Ω_{Tr} = {↓,⇓, **Root**, **Leaf**}[3], где ↓ и ⇓ – бинарные отношения, ↓ – отношение непосредственной доминации, ⇓ – отношение строгого частичного порядка, транзитивное замыкание отношения ↓ (и, конечно, ↓ – транзитивная редукция отношения ⇓). Дерево непосредственных составляющих должно быть деревом по отношению ↓ в смысле теории графов. **Root** и **Leaf** – унарные отношения, **Root** метит корень дерева, а листья помечены **Leaf**.

Ω_{Ord} = {→, ⇒}состоит из двух бинарных отношений: ⇒ – отношение строгого частичного порядка (слева направо) на вершинах дерева и → – транзитивная редукция ⇒, соотносящая «непосредственных соседей».

Два частичных порядка ⇓ и ⇒ соотносятся естественным образом. Использованная здесь нотация отражает «перпендикулярность» этих отношений, их точное соотношение описывается обычно аксиомами «дополнительного распределения» и «несмешения составляющих» (exclusivity and non-tangling conditions).

Как уже говорилось, отношения из Ω_{Lbl} представляют терминальные и нетерминальные символы.

На «картинках» при изображении деревьев непосредственных составляющих (см. Рис.1 и 2 выше) обычно линиями или стрелочками изображается только отношение ↓.

Аксиомы, выделяющие знаковую систему деревьев непосредственных составляющих (множество всех деревьев) из множества всех текстов в сигнатуре Ω_{ICTr}, легко записать на языке логики первого порядка[4]. Я не буду здесь этого делать (неформально и не очень полно они описаны выше).

2.2. Слова, деревья зависимости и другие примеры

Приведу с минимальными комментариями примеры текстов, представляющих слова в некотором алфавите (рис. 3), и деревья зависимостей (рис. 4).

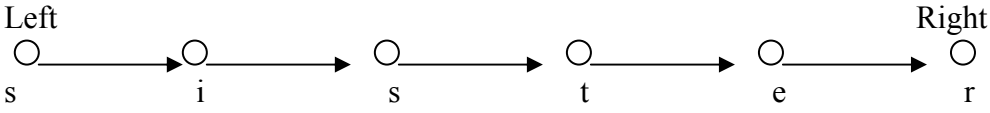

Рис.3. Слово “sister”

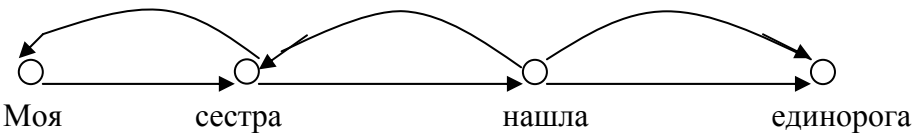

Рис. 4.

[3] Нотация, которую я здесь использую, отличается от нотации, используемой в наших работах (Борщев,Хомяков 1973а и др).

[4] Кроме, конечно, аксиомы конечности, выделяющей тексты в этой сигнатуре из всех моделей, не обязательно конечных.

Несущее множество каждой модели (текста), представляющей некоторое слово в данном алфавите, состоит из вершин, представляющих вхождения букв в это слово. Сигнатуру Ω_{Str} отношений таких моделей естественно разделить на две части: $\Omega_{Str} = \Omega_{LOrd} \cup \Omega_{TLbl}$

Отношения из Ω_{Lord}, подобно отношениям из Ω_{Ord} выше, представляют порядок «слева-направо», $\Omega_{Lord} = \{\rightarrow, \Rightarrow,$ **Left**, **Right**$\}$, только здесь $\Rightarrow$ – это отношение линейного порядка, а $\rightarrow$ – его редукция; унарные отношения **Left** и **Right** метят «крайние» вхождения. Отношения из Ω_{TLbl} представляют буквы алфавита.

Деревья зависимостей (графы Теньера) представляют грамматические зависимости на словах предложения. В таких деревьях нет нетерминальных вершин. Сигнатура Ω_{DT} этой знаковой системы – это, грубо говоря, смесь отношений из рассмотренных выше сигнатур Ω_{ICTr} and Ω_{Str}, $\Omega_{DT} = \Omega_{Tr} \cup \Omega_{LOrd} \cup \Omega_{TLbl}$. Отношения из Ω_{Tr} и Ω_{LOrd} обдают свойствами, описанными выше, но их соотношения несколько иные, я не буду здесь на этом останавливаться (см., например, Тестелец 2001).

Мы рассматривали и другие типы текстов (другие знаковые системы), представляющие контекстные грамматики Хомского, уже упоминавшиеся химические графы (Пантюхина и др. 1972) и ряд других.

3. Окрестностные грамматики и языки

3.1. Окрестности и типы окрестностей

Как уже говорилось, мы выделяем языки внутри знаковой системы, описывая локальные свойства их текстов. Мы рассматриваем *окрестности* вершин текстов, разные *типы* таких окрестностей и о*крестностные грамматики*.

Говоря неформально, окрестность вершины текста – это некоторый подтекст, содержащий эту вершину. Окрестностная грамматика накладыват ограничения на возможные окрестности вершин текста. Текст принадлежит языку, задаваемому грамматикой, если для каждой его вершины выполняются эти ограничения. В самом простом случае, окрестностная грамматика – это набор допустимых окрестностей и текст принадлежит языку, если у кажой вершины есть окрестность из этого набора.

Я ограничусь здесь неформальным описанием примеров для деревьев непосредственных составляющих.

3.2. «Кусты» и другие окрестности для деревьев

Простейшие окрестности для деревьев непосредственных составляющих – это «кусты». Такая окрестность для каждой вершины x содержит эту вершину (*центр* окрестности, на рисунках черный кружок) и все ее непосредственные составляющие (если они есть), т.е каждую вершину y, такую, что $\downarrow(x, y)$ – см. рис. 5.

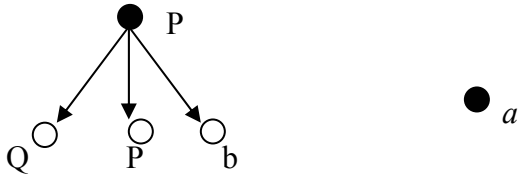

Рис.5.

Рассматривались и другие типы окрестностей, например, произвольные поддеревья, содержащие данную вершину (см. рис. 6).

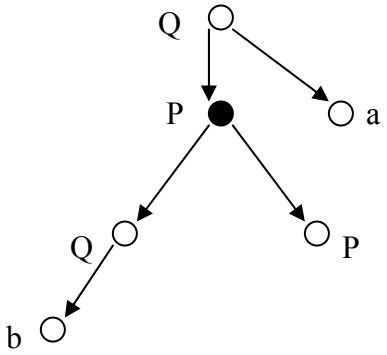

Рис. 6.

3.3. Окрестностные грамматики и окрестностные языки

Как уже говорилось, простейшая окрестностная грамматика – это конечный набор окрестностей некоторого типа. Окрестностный язык, задаваемый такой грамматикой, – это множество деревьев, у каждой вершины которой найдется окрестность из данного набора.

Например, каждой бесконтекстной грамматике можно сопоставить набор описанных выше «кустов»: например, правилу типа $P \rightarrow QPb$ сопоставляется куст, изображенный в левой части рис. 5, а каждому терминальному символу – тривиальную окрестность (см. правую часть рис. 5). Такая грамматика задает окрестностный язык, содержащий все синтаксические структуры, определяемые исходной бесконтекстной грамматикой.

В общем случае, окрестностная грамматика – это булевская формула из окрестностей некоторого типа (описанные выше простейшие грамматики можно понимать, как дизъюнктивные формулы).

Таким образом, окрестностные грамматики – это тоже аксиомы. выполняющиеся на текстах окрестностного языка. Их можно представить и в виде «обычных» формул языка первого порядка. Но окрестностное представление более структурировано и наглядно.

Итак, аксиомы, задающие знаковую систему, вместе с каждой окрестностной грамматикой, определяют окрестностный язык, как аксиоматический класс моделей.

3.4. Что было сделано

1) Мы получили целый ряд формальных результатов, прежде всего по характеризации окрестностных языков в знаковой системе деревьев непосредственных составляющих.

2) Были подробно исследовны некоторые другие знаковые системы, в частности системы, соответствующие контекстым грамматикам Хомского, именным окрестностным грамматикам (Борщев 1967) и др.

3) Как уже говорилось, идея окрестностных грамматики была распространена на описание переводов текстов. Перевод описывлся как пара текстов с бинарным отношением «переводимости» между вершинами этих текстов. Окрестностная система переводов – это множество таких пар. задаваемое окрестностной грамматикой.

4) Идея окрестностых грамматик была использована для описания функций и отношений, которое можно назвать окрестностной версией логичесого

программирования (Борщев, Хомяков 1972, 1974). При таком описании отношений рассматривались окрестности не для вершин, а для предикатных символов. Эта идея могла бы быть полезной и для описания окрестностных языков.

4. Недавние работы по Теоретико-Модельному Синтаксису

Из-за недостатка места только несколько слов о двух таких работах (Blackburn et al 1993 и Rogers, James 1997). Обе они имеют дело с деревьями непосредственных составляющих, вершины которых помечены «признаковыми структурами» (feature structure decorated trees), т.е. наборами грамматических признаков и их значений (например, CASE со значениями типа nominative, genitive, etc или PERSON со значениями *1d, 2d, 3d*). Такие структуры рассматривались в Generalized Phrase Structured Grammar (GPSG), см. (Gazdar *et al* 1985).

В обеих работах такие деревья рассматриваются как модели, на которых выполняются формулы некоторого логического языка, т.е. используется теоретико-модельный подход. В работе (Rogers, James 1997) используется так называемый monadic second-order language, а в работе (Blackburn et al 1993) – язык пропозициональной модальной логики.

Остановлюсь на второй работе. Сами по себе деревья непосредственных составляющих (без признаков) представляются примерно так же, как в разделе 2 выше. Но поскольку для их описания используется модальная логика, они интерпретируются, как модели Крипке: вершины рассматриваются как «возможные миры», а отношение непосредственной доминации как отношение «доступности» между «возможными мирами». Грамматика задается, как формула модальной логики. Формула истинна в модели (т.е. дереве), если она истинна в каждой вершине дерева (в каждом возможном мире).

Язык пропозициональной модальной логики содержит кроме констант и булевских операций несколько модальных операторов, в частности унарные операторы $\downarrow$,$\uparrow$. Так, формула $\downarrow\phi$ выполняется на вершине *и* дерева, если среди ее непосредственных составляющих существует вершина *и*', на которой выполняется формула ϕ.

Так как деревья «обогащены» признаковыми структурами, то используется «двуслойный» язык модальной логики (один слой для описания структуры дерева. а другой для описания структуры признаков).

В заключение нужно заметить, что хотя направление это (Model-Theoretic Syntax) в настоящее время активно развивается, оно до сих пор, увы, достаточно маргинально в общем потоке синтаксических работ.

Список литературы:

Борщев В.Б. 1967 Окрестностные грамматики *НТИ*, № 11, 39-41.

Борщев В.Б., Хомяков М.В. 1970 Окрестностные грамматики и перевод. *НТИ*, Серия 2, №3, 39-44.

Борщев В.Б., Хомяков М.В. 1972 Схемы для функций и отношений. *Препринт доклада на семинаре стран-членов СЭВ «Автоматическая обработка текстов на естественных языках»*, Ереван.

Борщев В.Б., Хомяков М.В. 1973а Аксиоматический подход к описанию формальных языков. В сб. *Математическая лингвистика*, под ред. С.К. Шаумяна. М., Наука, 5-47.

Борщев В.Б., Хомяков М.В. 1973б Окрестностные переводы. В сб. *Математическая лингвистика*, под ред. С.К. Шаумяна. М., Наука, 48-62.
Пантюхина М.Е., Борщев В.Б., Хомяков М.В. 1972 Об одном способе описания языка химических структурных формул. *НТИ*, Серия 2, № 5, 34-36.
Борщев В.Б., Хомяков М.В. 1974 Схемы для функций и отношений. В сб. *Исследования по формализованным языкам и неклассическим логикам*, под ред. Д.А. Бочвара. Москва, «Наука», 23-49.
Тестелец Я.Г. 2001 *Введение в общтий синтаксис*. Москва, изд. РГГУ.
Шрейдер Ю.А. 1967 Окрестностная модель языка. *Труды симпозиума по применению порождающих грамматик.* Тарту, сентябрь 1967.
Blackburn, Patrick, Claire Gardent, and Wilfried Meyer-Viol 1993 Talking about trees. *Proceedings of the 1993 Meeting of the Europeen Chapter of the Association for Computational Linguistics*, 21-29.
Blackburn, Patrick and Wilfried Meyer-Viol 1997 Modal Logic and Model-Theoretic Syntax. In M. deRijke (ed.), *Advances in Intensional Logic*, 29-60. Dodrecht, Kluwer Academic.
Borščev, V.B. and M.V. Xomjakov 1973 Axiomatic approach to a description of formalized languages and translation. Neighbourhood Languages, in F.Kiefer, ed., *Linguistische Forschungen*,18, Soviet Papers in Formal Linguistics, vol 3, Athanaeum, 37-114.
Borščev, V.B. and M.V. Xomjakov 1976 Neighbourhood Grammars and Translation. An axiomatic Approach to the Description of Formal Languages. In Ferenc Papp and Győrgy Szėpe, ed. *Papers in Computational Linguistics* (Proceedings of the 3rd International Meeting on Computational Linguistics, Debrecen, 1971). Budapest, 427-432.
Borščev, V.B. and M.V.Chomjakov 1977 Neighbourhood Description of Formal Languages. In Leo S. Olschki ed. *Computational and Mathematical Linguistics. Proceedings of the International Conference on Computational Linguistics*, Pisa, 27/VIII-1/IX 1973, Firenze 3-7.
Gazdar, G., Klein,E., Pullum, G., and Sag, I. 1985 *Generalized Phrase Structure Grammar.* Basil Blackwell; Harvard University Press.
Pullum, Geoffrey K. and Barbara C. Scholz 2001 On the Distinction between Model-Theoretic and Generative-Enumerative Syntactic Frameworks. In Philippe de Groote,Glyn Morill and Christian Retore (eds.) *Logical Aspects of Computational linguistics, Lecture Notes in Artificial Intelligence 2099*. Springer 17-43.
Rogers, James 1997 “Grammarless” phrase structure grammar. *Linguistics and Philosophy* 20, 721-746.