СПРАВОЧНО-ИНФОРМАЦИОННЫЙ РАЗДЕЛ

УДК 81'32:303.714(049.32)

В. В. Борисов, Р. Г. Пиотровский, Е. Б. Соколинская

Лингвостатистика и критерий истинности в языкознании

Рецензия на кн.: Mańczak. Problemy językoznawstwa ogólnego.— Wrocław; Warszawa; Kraków: Ossolineum, 1996 .— 258 с.

Автор рецензируемой книги — один из ведущих представителей современного традиционного языкознания. Его перу принадлежит более пятисот книг, статей, рецензий и тезисов, посвященных вопросам романистики, славистики и германистики, а также проблемам общего языкознания и индоевропеистики [1, с. XI–XXIV]. Профессор Манчак далек от идей и методов структурного языкознания и генеративной лингвистики, лингвостатистики и инженерной лингвистики, а также от тех направлений естественных и физико-математических наук, которые подобно информатике, акустике, теории нечетких множеств и лингвистической переменной активно взаимодействовали с наукой о языке в последние пятьдесят лет. В свете сказанного некоторые ключевые идеи автора, развиваемые в книге, могут, с одной стороны, озадачить лингвиста, придерживающегося традиционных взглядов, а с другой — указать математику, информатику и естественнику на те подводные камни и западни, которые подстерегают последних при их попытках опереться на теорию и практику традиционного языкознания.

Среди эпистемологических идей, развиваемых автором, наиболее важными являются следующие положения:

1. Основной причиной кризисного состояния современного языкознания (238)* является пренебрежение критерием истинности (КИ), который во всех развитых науках реализуется путем применения различных приемов проверки валидности теоретических гипотез и достоверности эмпирических результатов, в первую очередь через контроль практики (14–15).

2. В языкознании КИ повсеместно заменяется критерием персональных авторитетов "*X* сформулировал гипотезу, *X* является авторитетом, — иронизирует В. Манчак, — следовательно гипотеза верна. *Y* также выдвинул гипотезу, но *Y* не является авторитетом, поэтому гипотеза неверна" (13).

3. Основным приемом реализации КИ в языкознании должно быть количественное (в терминологии автора — статистическое) обследование текстов (128 и сл.).

Хотя первые два положения в целом бесспорны, они нуждаются в небольшом комментарии. Проблема КИ и проверки достоверности является болевой точкой большинства гуманитарных дисциплин. Объективных причин здесь несколько. Во-первых, эти науки исследуют нечеткие процессы и объекты, группирующиеся в нечеткие множества и подчиняющиеся нечеткой логике. Удобного и надежного аппарата для объективного и конструктивного описания нечетких объектов и множеств, по признанию самого создателя теории нечетких множеств Л. А. Заде, пока не создано [2, с. 7]. Во-вторых, многие гуманитарные науки, например, история, литературоведение, среди лингвистических дисциплин — сравнительно-историческое языкознание и стилистика художественной речи имеют дело с нестационарными (проще говоря, с уникальными, не повторяющимися) явлениями. В этой ситуации гуманитар не в состоянии осуществить проверочный эксперимент или оценить практикой достоверность своих выводов. Он вынужден иметь дело с фактами, так как они происходят сами по себе, независимо от его контроля. Именно поэтому гуманитарии вынуждены в качестве суррогата КИ использовать мнения авторитетов, политическую или иную конъюнктуру (ср. спор норманистов и антинорманистов по поводу формирования Киевской Руси и происхождения самого термина *Русь* [21]) или опираться на статистику мнений и выбирать ту гипотезу, которая набирает наибольшее число голосов специалистов. Эти традиции настолько сильны в научном менталитете гуманитариев, и лингвистов в частности, что даже в тех случаях, когда новые открытия или практические результаты подтверждают ту или иную теорию, опровергая альтернативную гипотезу, последняя еще долго вращается в научном обиходе.

Классическим примером этому служит судьба формулы индоевропейского корня, теоретически обоснованная в конце 80-х гг. XIX в. малоизвестным тогда Ф. де Соссюром и оставленная без внимания его современниками. Только после того, как эта гипотеза нашла свое подтверждение на материале хеттского языка [3], она стала входить в арсенал индоевропеистики. Между тем, еще и сегодня встречаются индоевропеисты, сомневающиеся (правда, без достаточных на то оснований) в достоверности открытия Соссюра. Более свежим примером является теория Ельмслева—Хомского, согласно которой естественный язык (ЕЯ), подобно искусственным языкам математики, химии и т. п., представляет собой жесткое исчисление (calculus). На-

* В скобках указываются страницы рецензируемой книги.

чиная с конца 50-х гг. эта концепция стала теоретическим фундаментом при построении десятков систем "алгебраической" переработки текста (АПТ), в том числе и машинного перевода (МП). Альтернативный взгляд, согласно которому язык не является исчислением, но есть открытая, нечетко организованная окказиональная система [4, 249–258; 5], и поэтому для построения систем АПТ и МП необходимо использовать нетрадиционную стратегию, долгое время оставался без внимания. Даже сегодня, когда стало ясно, что на основе концепции Ельмслева—Хомского построить работающую систему АПТ-МП нельзя, а все реально функционирующие системы опираются на альтернативное представление о языке, есть инженерно-лингвистические коллективы, которые, завороженные магией теории языка-исчисления, продолжают безуспешные попытки построения "алгебраических" систем [6].

Как было сказано, центральным приемом, с помощью которого в науку о языке может быть введен КИ является, считает В. Манчак, количественный обсчет лингвистических явлений и, в первую очередь, лексики. Это и понятно, с одной стороны, лексика охватывает большое количество сравнительно легко выделяемых и обсчитываемых единиц. С другой — она содержит от 65 до 80% информации текста [7, с. 236]. Далее, по логике вещей прямолинейное и всеобъемлющее использование частотности в качестве КИ требует от исследователя, отодвинув вопросы системности и соссюрианские дихотомии "язык—речь", "синхрония—диахрония" и т. п. на задний план, рассматривать язык, вслед за В. Гумбольдтом и американскими дескриптивистами, как сумму всех порожденных когда-либо текстов. Именно так поступает проф. Манчак (21–25). К сказанному следует добавить, что автор книги считает частотность основным двигателем в развитии и функционировании языка (32 и сл.). С этих позиций автор рассматривает такие ключевые проблемы современной лингвистики, как вопрос регулярного и нерегулярного фонетического развития (главы IV–V), закономерности в развитии морфологии и словообразовании (главы VII, IX, XII), семантические сдвиги в лексике (гл. XI), проблемы этногенеза, родства языков и их классификации (главы XVI, XVII, XXIII) и другие более частные вопросы. При этом используется не только богатый индоевропейский материал, но также факты иноструктурных языков.

Количественные оценки, в первую очередь лексическая статистика*, являются, разумеется, сильным ходом на пути поиска научной истины в науке о языке. Примером может служить проведенный автором на материале романских и германских языков квантитативный тест (гл. XVIII) по проверке справедливости известного тезиса М. Бартоли [9], согласно которому в группе родственных языков его периферийные члены являются более архаичными, чем центральные, и альтернативного утверждения, по которому периферийные языки под воздействием субстрата и адстрата, наоборот, быстро меняют свое лицо [10, с. 18–19]. Этот тест, хотя и не безупречный со статистической точки зрения, показывает, что центральная или периферийная география языка по всей вероятности существенно не влияет на степень его архаичности или продвинутости (182). К сожалению, выдвигая тезис об универсальности квантитативно-текстовой методики, проф. Манчак, а вместе с ним сотни современных филологов забывают о ряде важных правил организации статистического эксперимента и методики анализа его результатов. Без соблюдения этих правил статистика может стать "опасным орудием в руках неопытного человека" [11, с. 28] или даже "третьим видом лжи". Вот одно из таких правил:

Если речь идет о больших генеральных совокупностях лингвистических единиц, каковыми являются *язык* (в понимании В. Манчака), *диалект, множество лексических* или *грамматических форм языка*, то исследователь, как правило, не может охватить такую совокупность целиком. Он вынужден анализировать лишь отобранную часть ее, т. е. выборочную совокупность. Такая выборка должна быть:

качественно (лингвистически) репрезентативна, т. е. обладать теми основными лингвистическими признаками, которые присущи генеральной совокупности;

репрезентативной в количественном отношении, т. е. иметь такой объем, который обеспечивал бы достаточно высокую уверенность, что полученные на выборке количественные характеристики близки к реальным количественным свойствам генеральной совокупности.

Если выборка не отвечает требованиям качественной и количественной репрезентативности, то получаемые результаты квантитативного эксперимента имеют лишь иллюстративный характер и новой информации, как правило, не несут.

Рассмотрим в этом плане два пассажа из рецензируемой книги. Согласно традиционной точке зрения английский язык является германским, а румынский — романским языками. Однако высказываются мнения, опирающиеся на словарную статистику этих языков, согласно которым английский, содержащий в своем словаре до 85% латинизмов и галлицизмов, является скорее языком романским, а румынский, лексика которого включает до 67% славянизмов, есть язык славянский. Справедливость этих утверждений В. Манчак хочет проверить с помощью не словарной, но речевой статистики (160–162), что вполне разумно. Для этого он рассматривает английский текст из предисловия к Oxford Advanced Learner's Dictionary of Current English (OALDCE) объемом в 50 словоупотреблений и текст румынского классика М. Садовяну длиной в 54 словоформы. Результаты этого теста опровергают, по мнению автора, первоначальные гипотезы и возвращают нас к традиционной точке зрения. Английский текст содержит 32 германских и только 18 романо-латинских словоупотреблений, в румынском же тексте — 44 романских и только 10 славянских словоформ.

Однако продолжим этот тест относительно другой фразы из Оксфордского словаря, непосредственно следующей за тестовым фрагментом проф. Манчака:

As a *consequence*, the length *mark associated* with *certain vowels* has been *restored*, though in *strict phonologocal terms* this *mark* may be *concidered*

* Первые частотные словари появились еще во второй половине прошлого века. К ним относятся: Gamble W. Two lists of selected characters containing all in the Bible and twenty-seven other books.— Shanghai, 1861 [8, с. 116] и Куницкий В. Н. Язык и слог в комедии "Горе от ума".— Киев. 1894, а также словарь Ф. В. Кединга (1897–98), который В. Манчак ошибочно считает первым словарем этого типа (240).

redundant in the *chosen vowel symbols distinguish qualitative difference* (OALDCE.— Moscow—Oxford, 1982 .— C. VI).

Соотношение здесь германизмов и романо-латинизмов (отмечены курсивом) составляет здесь 14:18. Аналогичным образом и в наугад взятой первой фразе одной из повестей того же М. Садовяну

De *Sfânta* Inălţare, *hram* al *ctitoriei Neamţu*, se strânsese în *preajma zidurilor* şi în *ogrăzile călugărilor* mare număr de *norod* (M. Sadoveanu. Fraţii Jderi.— Bucureşti, 1953 .— C. 7),

или в названиях рассказов другого румынского классика — Й. Крянги

"*Popa Duhu*", "*Povestea* lui *Harap*-Alb", "*Ivan Turbinca*", "*Povestea* unui om *leneş*", "*Poveste* (*Prostia* omenească)" (I. Creangă. Opere.— Bucureşti, 1953 .— C. 319)

соотношение романских слов и славянизмов (последние отмечены курсивом) равно 16:17 [12], причем из первых только шесть являются знаменательными, а остальные — служебными словами и грамматическими формантами. Славянизмы же являются полновесными знаменательными словами. Следуя логике В. Манчака, такие результаты говорят в пользу начальной гипотезы.

Разумеется, тесты автора и рецензентов не имеют объяснительной силы, поскольку в обоих случаях выборки не являются репрезентативными, ни с точки зрения статистики, так как они слишком малы, ни лингвистически, поскольку они не представляют всего разнообразия румынских стилей и подъязыков.

Из сказанного следует, что применение лингвостатистических методов позволяет получать новую информацию, отвечающую критерию истинности, во-первых, при условии лингвистически грамотного построения выборки, во-вторых, при определении ее репрезентативного объема и, наконец, при выяснении степени достоверности полученных количественных результатов. К сожалению, два последних условия не слишком часто выполняются филологами. Причину этого наш автор видит, во-первых, в излишней усложненности и поэтому недоступности для гуманитария математического аппарата, применяемого для их выполнения, а, во-вторых, в том, что использование и развитие этого аппарата становится самоцелью для математической лингвистики (239–241). Эта очень популярная среди традиционных лингвистов идея требует особого комментария.

Выше уже говорилось о том, что лингвистика имеет дело с нестационарными процессами, а также с нечеткими объектами, группирующимися в нечеткие множества и подчиняющимися нечеткой логике. Поэтому лобовое применение вероятностно-статистического аппарата, выработанного на естественнонаучном материале, здесь не всегда дает правильные решения. Например, некорректно оценивать с помощью традиционного среднего квадратического отклонения (а тем более линейного) достоверность численных оценок той информации, которую несут буквы, морфемы, текстовые слова и словоупотребления, не только потому, что большинство лингвистических объектов не дает нормального распределения, но также из-за того, что такие оценки получены с помощью логарифмической меры. Другая сложность лингво-статистических исследований заключается в том, что ученый вынужден работать с лингвистическими выборками малого объема, такими, как древние памятники, короткие рассказы и документы. Именно поэтому математическим лингвистам приходится заново вырабатывать достаточно сложные и изощренные процедуры для адекватного моделирования таких важных для науки о языке зависимостей как *длина текста и объем его словаря, частота слова и его номер в частотном словаре* (так называемый закон Ципфа, который В. Манчак считает (32–43) краеугольным камнем своей квантитативной концепции), закон Менцерата—Альтманна, описывающий отношение между объемом единицы текста (словом, словосочетанием и предложением) и длиной образующих эту единицу элементов и другими связанными с этими законами зависимостями [13, с. 66–95; 14, с. 1–24]. Поэтому, если языковед хочет понять сущность глубинных синергетических закономерностей языка, которые стоят за внешними количественными отношениями в тексте, он должен овладеть основами математической грамотности.

При всей важности и актуальности лингвостатистики, абсолютизация квантитативных приемов приводит иногда к заведомо сомнительным выводам. Рассмотрим в этом плане проблему прародины индоевропейцев.

Этот вопрос проф. Манчак стремится решить путем прямолинейных лексических сопоставлений. Логика рассуждений, навеянная лингвостатистической таксономией А. Л. Кребёра [15], выглядит здесь следующим образом. Наиболее близким к языку-основе (латинскому, прагерманскому, праславянскому, индоевропейскому и т. п.) является тот из живых языков, лексика которого дает наибольшее число схождений с другими языками данной семьи или группы. Отсюда следует, по мнению автора (151–160, 176), что территория, занимаемая таким языком, и есть искомая прародина тех народов, которые являются носителями данной группы или семьи языков. В частности прародина славян локализуется автором в междуречьи Одера и Вислы, германцев — в Южной Германии, а индоевропейцев — на Балканах (225–229). Возможность вторичных контактов родственных языков и их хронология, древние миграции народа, лучше всего сохранившего старый лексический фонд, в расчет не принимаются.

Одним из наиболее известных примеров, опровергающих лобовое лексико-статистическое решение этногенеза, является судьба угорских языков. Следуя концепции В. Манчака, прародину угров следовало бы поместить либо в Венгрии, либо в низовьях Оби, где бытуют хантыйский и мансийский языки. Решение зависело бы от того, в каком из трех языков обнаруживалось бы больше искони угорских слов. Разумеется, такое решение бессмысленно, поскольку на основании других лингвистических, палеоботанических и зоолингвистических, а отчасти и археологических данных установлено, что прародиной угров являются районы между Волгой и Уралом [16, с. 33–36].

История науки показывает, что при изучении сложных систем критерий истинности вырабатывается путем сопоставления результатов, полученных с помощью разных критериев и методических приемов, в этом же ключе работает и известный принцип дополнительности. Именно поэтому, возвращаясь к затронутой в книге В. Манчака проблеме этногенеза, коротко рассмотрим комплекс исследовательских приемов, которыми пользовались де-

сятки лингвистов, историков и археологов, начиная с А. Неринга и В. Бранденштейна [17, с. 9–275, 231–277] и кончая М. Гимбутас [18], для обоснования гипотезы южно-уральской прародины индоевропейцев. Поиск истины (X) можно условно представить здесь в виде последовательного пересечения (наложения) таких наиболее существенных для решения указанной задачи множеств-территорий как:

область курганных погребений (A') и районы приручения лошади (A''), простирающаяся от предгорий Карпат до Южного Урала;

три лесостепных предгорных района, т. е. Прикарпатье (B'), Северный Кавказ (B''), Южный Урал (B'''), для которых характерна флора и фауна, оставившая свои следы как в западных, так и в восточных индоевропейских языках;

области наиболее вероятных непосредственных и культурно-экономических контактов индоевропейцев с: 1) семитами — на юг от Урала, Кавказа и на восток от Балкано-карпатского района (C'), 2) финноуграми — на север от Южного Урала (C''), тюрками — на восток от Южного Урала (C'''), ср. [19].

Иными словами

$$X = A' \cap A'' \cap B' \cap B'' \cap B''' \cap C' \cap C'' \cap C''',$$

и в результате величина X указывает на Южный Урал.

Особо отметим, что при локализации прародины индоевропейцев большинство специалистов не использует количественные характеристики сходства и различий в лексике отдельных индоевропейских языков, предлагаемые А. Л. Крёбером и В. Манчаком, поскольку в них неразъединимо суммированы такие генетические и культурно-экономические связи народов-носителей, которые относятся к совершенно различным эпохам и территориям.

Теоретико-множественную процедуру, неосознанно реализованную авторами южно-уральской гипотезы, можно было бы формализовать, усиливая этим мотивацию принятия одной из альтернативных гипотез. Для этого пришлось бы выполнить трудную задачу, предусмотренную теорией принятия решений при нечеткой исходной информации, которая заключается в присвоении каждому из взаимодействующих множеств и их элементам весов принадлежности [20, с. 11–12, 20–22], в нашем случае — правдоподобия. В такой процедуре, вероятно, нашлось бы место и лексико-статистическим гипотезам.

Наш комментарий к основным положениям книги проф. Манчака и некоторые сомнения по поводу некоторых идей, развиваемых автором, нисколько не умаляют ее значения для развития теории и методологии языкознания. Она представляет собой пока еще, к сожалению, редкий пример рассмотрения теоретической концепции на богатом многоаспектном и многоязычном материале. В этом смысле книга стоит в одном ряду с такими трудами как "Язык" Л. Блумфилда или "Основы фонологии" Н. С. Трубецкого. Основная же заслуга проф. Манчака состоит в том, что он впервые четко сформулировал давно вызревавшую в лингвистике проблему критерия истинности и приемов его подтверждения.

СПИСОК ЛИТЕРАТУРЫ

1. Munus Amicitiae. Studia linguistica in honorem Witoldi Mańczak septuagenarii / Ed. curaverunt A. Bochnakowa et S. Widlak.— Cracoviae: Universitas Iagellonica, 1995.

2. Нечеткие множества и теория возможности. Последние достижения / Под ред. Р. Р. Ягера: Пер. с англ.— М.: Радио и связь, 1986.

3. Kuryłowicz J. indoeuropéen et ḫ hittite // Symbolae in grammaticae in Honorem Joannis Rozwadowski, 1 .— Kraków, 1927 .— С. 92–104.

4. Дрейфус Г. Чего не могут вычислительные машины. Критика искусственного разума: Пер. с англ.— М.: Прогресс, 1978.

5. Мельников Г. П. Системология и языковые аспекты кибернетики.— М.: Сов. радио, 1978.

6. Piotrowski R. Machine translation in new Russia // Limbaj şi Tehnologie / Ed. D. Tufiş.— Bucureşti: Editura Academiei Române, 1996 .— С. 85–92.

7. Piotrowski R. Text — Computer — Mensch.— Bochum: Brockmeyer, 1984.

8. Алексеев П. М. Статистическая лексикография (типология, составление и применение частотных словарей): Уч. пособие.— Л.: ЛГПИ им. А. И. Герцена, 1975.

9. Bartoli M. Caratteri fondamentali delle lingue neolatine // Archivio Glottologico Italiano.— 1937 .— Vol. 29, fasc. 2 .— С. 1–20.

10. Kuhn A. Romanische Philologie. Erster Teil. Die romanischen Sprachen.— Bern: Franke, 1951.

11. Юл Дж. Э. и Кендэл М. Дж. Теория статистики.— 14-е изд., пересмотр. и расшир.: Пер. с англ.— М.: Госстатиздат ЦСУ, 1960.

12. Gioranescu A. Diccionario Etimológico Rumano. Fasc. 1–7 // La Laguna Tenerife: Biblioteca filológica. Universidad de La Laguna. 1959–66.

13. Орлов Ю. К. Статистическое моделирование речевых потоков // Вопр. кибернетики. Вып. 41. Статистика речи и автоматический анализ текста. М.— Л.: АН СССР. Научный совет по комплексной проблеме "Кибернетика", 1978 .— С. 66–99.

14. Hřebíček L. Text levels. language constructs, constituents and the Menzerath—Altmann law // Quantitative Linguistics. Vol. 56 .— Trier: Wissenschaftlicher Verlag Trier, 1995 .— С. 1–24.

15. Kröber A. L. Statistics, Indo-European, and taxonomy // Language .— 1960 .— Vol. 36, N. 1 .— С. 1–21.

16. Гуя Я. Прародина финно-угров и разделение финно-угорской этнической общности // Основы финно-угорского языкознания (вопросы происхождения финно-угорских языков).— М.: Наука, 1974 .— С. 28–41.

17. Die Indogermanen- und Germanenfrage. Neue Wege zu ihrer Lösung: Wiener Beiträge zur Kulturgeschichte und Linguistik. B. 4 .— Salzburg—Leipzig, 1936.

18. Gimbutas M. Primary and secondary homeland of the Indo-European // The J. of Indo-European Studies.— 1985 .— Vol. 13, N. 1–2 .— С. 10.

19. Леман В. П. Новое в индоевропеистических исследованиях: Пер. с англ. // Вопр. языкознания.— 1991, N 4 .— С. 5–30; N 5 .— С. 5–26.

20. Орловский С. А. Проблемы принятия решения при нечеткой информации.— М.: Наука. Гл. ред. физ.-мат. лит-ры, 1981.

21. Мавродин В. В. Борьба с норманизмом в русской исторической науке. Стенограмма публичной лекции, прочитанной в 1949 году в Ленинграде. Л.: Всесоюзное общество по распространению политических и научных знаний. ЛО, 1949.

Материал поступил в редакцию 10.03.99.