В данном тексте на стр. 2 (в книге на стр.7), скорее всего, опечатка:
"Особенности построения лингвистического **ЯЗЫКА** приводят к тому, что естественный язык представляет собой нежестко"
следует читать:
"Особенности построения лингвистического **ЗНАКА** приводят к тому, что естественный язык представляет собой нежестко"

Выдержки из учебного пособия «Математическая лингвистика» (Пиотровский Р.Г. и др.)

ВВЕДЕНИЕ

1. Язык и математика. В эпоху научно-технической революции математизация охватывает все сферы человеческой деятельности, в том числе и такие, казалось бы, чисто гуманитарные науки как языкознание. Проникновение математических методов в лингвистику обусловлено двумя причинами.

Во-первых, развитие языковедческой теории и практики требует введения все более точных и объективных методов для анализа языка и текста. Одновременно использование математических приемов при систематизации, измерении и обобщении лингвистического материала в сочетании с качественной интерпретацией результатов позволяет языковедам глубже проникнуть в тайны построения языка и образования текста.

Во-вторых, все расширяющиеся контакты языкознания с другими науками, например с акустикой, физиологией высшей нервной деятельности, кибернетикой и вычислительной техникой, могут осуществляться только при использовании математического языка, обладающего высокой степенью общности и универсальности для различных отраслей знаний. Особенно настойчиво математизируется языкознание в связи с использованием естественного языка в информационных и управленческих системах человек–машина–человек. В действующих системах машинного перевода, автоматического аннотирования, человеко-машинного диалога всякое сообщение на естественном языке перекодируется в математическом языке компьютера.

Говоря об особенностях взаимодействия языкознания и математики, следует иметь в виду, что как естественный язык, так и язык математики являются знаковыми (семиотическими) системами передачи информации.

Основные расхождения между этими языками связаны с различным построением языкового знака и знака математического.

Во-первых, лингвистический знак (слово, словосочетание, предложение) обычно включает в себя четыре компонента – *имя* (материальный носитель информации), *денотат* (отражение предмета из внешнего мира), *десигнат* (понятие о предмете) и *коннотат* (комплекс чувственно-оценочных оттенков, связанных с предметом и понятием о нем); знак математического языка включает только имя и десигнат (математическое понятие); сказанное иллюстрирует рис. 1.

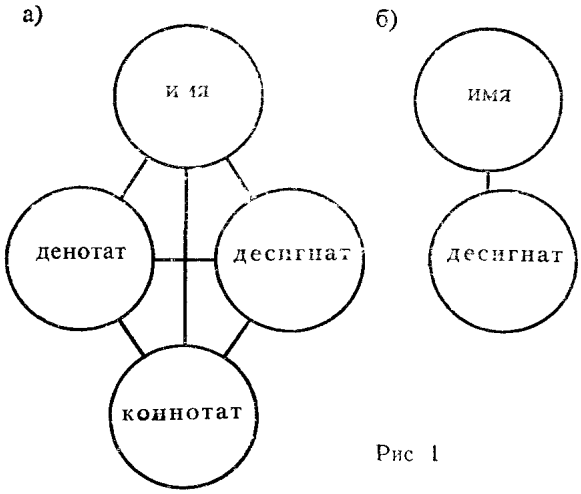

Рис 1

Во-вторых, лингвистический знак многозначен; математический знак имеет, как правило, одно концептуальное значение.

В-третьих, лингвистический знак потенциально метафоричен, у знака математического метафоричность полностью отсутствует.

Все эти свойства лингвистического и математического знаков можно проследить, сравнив значения математического знака 7 и слова *семерка.* Если 7 имеет единственное десигнативное математическое значение – «семь любых объектов», то слово *семерка* имеет несколько значений: «цифра 7», «карта в семь очков», «группа из семи человек» и т. п. При этом в значении слова *семерка* содержатся не только указанные десигнативные понятия, но оно может указывать на конкретный предмет, например на вполне опре-

деленную группу в семь человек. Одновременно это слово несет дополнительные коннотативные метафорические оттенки, связанные с такими словосочетаниями как «великолепная семерка», «семь чудес света», «семь смертных грехов», «семь дочерей Атланта (Плеяды)» и т. д.

Из всего сказанного вытекает еще одно важное различие между десигнативными значениями математического и лингвистического знаков.

Значение каждого математического знака легко представить в качестве множества элементов, причем такое множество имеет вполне четкие границы: значение знака 7 является множеством, охватывающим такие конкретные совокупности, которые включают только семь (не шесть и не восемь!) предметов.

Иначе организовано десигнативное значение лингвистического знака – оно также может рассматриваться как множество денотатов, однако это множество не всегда имеет четкие границы. Так, например, не удается определить смысловые границы слов *голубой* и *синий, голубой* и *зеленый.* Разные люди в зависимости от особенностей своего хроматического зрения будут называть показываемые им конкретные сине-голубые оттенки то синим, то голубым цветом. Нельзя также указать точную временную границу, разделяющую значения слов *ночь* и *утро.* Иными словами, значения лингвистических знаков представляют собой нечеткие множества с размытыми границами [26, с. 207–214]; [65].

С многозначностью, метафоричностью и нечеткостью смысловых границ лингвистического знака связана также изменчивость его значения. В качестве примера снова возьмем русское прилагательное *голубой.* В 50-е годы это слово, судя по 3-му изданию «Словаря русского языка» С. И. Ожегова (М., 1957), имело в литературном русском языке только одно толкование: «с окраской светло-синего цвета». Однако, словарь-справочник, составленный по материалам прессы и литературы 60-х годов «Новые слова и значения» (М., 1971), указывает для слова *голубой* еще одно значение – «идеализированный», отмечая одновременно такие новые метафорические употребления как «голубое топливо», «голубой экран».

Особенности построения лингвистического ~~языка~~ приводят к тому, что естественный язык представляет собой нежестко организованную диффузную систему, которая воспринимается и используется человеком в значительной мере интуитивно.

Напротив, язык математики является хорошо организованной системой, существующей и функционирующей в виде логического построения, каждый элемент которого имеет осознанную значимость.

Конфронтация естественного языка и языка математики требует, чтобы каждому лингвистическому объекту был поставлен в соответствие некоторый математический объект. Лингвистический знак, например, словосочетание или слово и составляющие этот знак фигуры – фонемы, буквы, слоги – должны интерпретироваться с помощью знаков математических. Эта математическая интерпретация связана с расчленением лингвистического объекта и выделением в нем одного смыслового или сигнального компонента, который становится предметом дальнейшего исследования. Остальные сигнальные и смысловые элементы лингвистического объекта, а также разного рода метафорические коннотативные оттенки из рассмотрения исключаются.

Применение математических методов в языкознании имеет своей целью заменить обычно диффузную, интуитивно сформулированную и не имеющую полного решения лингвистическую задачу одной или несколькими более простыми, логически сформулированными и имеющими алгоритмическое решение математическими задачами. **Такое расчленение сложной лингвистической проблемы на более простые алгоритмизуемые задачи мы будем называть математической экспликацией лингвистического объекта или явления.**

Математическая экспликация интересна не только с чисто познавательной и теоретической точки зрения. Она совершенно необходима при решении прикладных вопросов, связанных с анализом и синтезом устной речи или информационной переработкой текста на ЭВМ. Математическая экспликация лингвистических объектов применяется не только при решении на ЭВМ несложных, хотя и трудоемких задач такого типа как составление частотных и алфавитных словников [3]; [8]; [22] или пословного и пообратного машинного перевода [32 а, с. 286 и ел.], [32 б, с. 107–130], но также при составлении и реализации таких эвристических алгоритмов искусственного интеллекта как семантический машинный перевод [32 в, с. 128–146] или тезаурусное реферирование текста [26, с., 248–268].

2. Комбинаторная и квантитативная лингвистика. Выбор математического аппарата в лингвистических исследованиях – вопрос не простой. Его решение зависит в первую очередь от того, как определяется предмет и основные понятия языкознания и его теоретического ядра – структурно-математической лингвистики.

Некоторые математики и лингвисты считают, что предметом математической и структурной лингвистики должно быть изучение грамматики, порождающей текст. При этом грамматика понимается как конечное множество детерминированных правил, в том числе неграмматических, а язык рассматривается как бесконечное число регулярных цепочек слов, порождаемых этой грамматикой. При этом подходе экс-

пликация лингвистических объектов должна опираться на такие разделы «неколичественной» математики как теория множеств, математическая логика (в особенности, теории рекурсивных функций и бинарных отношений), теория алгоритмов и т. д.

Что же касается «количественных» разделов математики (математическая статистика, теория вероятностей, теория информации, математический анализ), то они считаются либо неприменимыми для экспликации лингвистических явлений, либо играющими вспомогательную роль. На основе применения «неколичественного», или как его иногда называют, «качественного» математического аппарата в теоретическом языкознании сформировалось направление, условно называемое комбинаторной лингвистикой. Это направление противопоставляется квантитативной (количественной) лингвистике [43, с. 273].

Методы детерминистского комбинаторного языкознания интенсивно разрабатываются в теории порождающих грамматик Хомского [45], в теоретико-множественных моделях Маркуса [56] и в других лингвистических направлениях.

Однако математическое языкознание не может ограничиться детерминистской, неколичественной экспликацией лингвистически» объектов.

Во-первых, это ограничение затрудняет преобразование нечетких лингвистических множеств, элементы которых имеют вероятностные веса принадлежности, в четкие множества искусственных языков. Между тем указанное преобразование лежит в основе всех видов машинной переработки текста и автоматического распознавания смысла [26, с. 215–228].

Во-вторых, при таком ограничении вне сферы применения математических методов остается акустико-физиологическая и психолингвистическая проблематика речеобразования, а также стилистика и история языка, при изучении которых широко применяются не столько комбинаторные, сколько количественные измерения [18]; [21]; 127]; [32 в, с. 361–400].

Для того чтобы правильно оценить соотношение комбинаторных и количественных математических методов при описании языка и текста, рассмотрим обитую схему речевой деятельности и текстообразования.

Порождение текста определяется, с одной стороны, системой языка и ограничивающей ее действие нормой, а, с другой – совершенно независимой от языка внешней ситуацией (рис. 2).

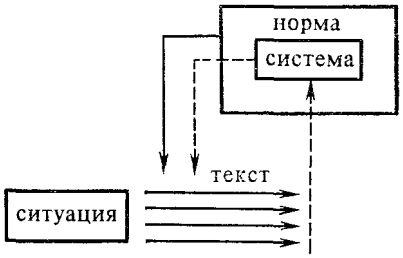

Рис. 2

Если согласиться с тем, что система языка есть механизм, порождающий тексты без каких-либо вероятностных ограничений [45]; [59], то станет ясным, что экспликация этой системы должна осуществляться с помощью тех неколичественных разделов математики, о которых мы говорили выше.

Рассматривая язык как неколичественную систему, комбинаторная лингвистика пытается описать механизм перехода от языка к речи с помощью тех же приемов «неколичественной» математики. Такие описания представляют собой контекстно-свободные грамматики, т. е. грамматики, не учитывающие контекстных ограничений на употребление отдельных лингвистических единиц и их сочетаний. В связи с этим контекстно-свободные грамматики порождают много цепочек, не являющихся реальными предложениями данного языка. Чтобы добиться порождения реальных текстов, необходимо перейти от контекстно-свободных грамматик к более сильным контекстно-зависимым грамматикам. Такие грамматики можно построить при условии, что к элементам системы языка применяются вероятностные оценки, а сам язык рассматривается как неколичественная производящая система, функционирование которой регулируется вероятностными ограничениями, заложенными в норме [32а, с. 5–46]; [47].

Что же касается текста (речи), то он представляет собой линейною цепочку отграниченных друг от друга (дискретных) символов, (фонем, букв, слогов, слов). Каждый из символов встречается в тексте с определенной частотой и обладает особыми валентностями, т. е. лингвистическими способностями сочетаться с другими символами. Эти свойства лингвистических единиц в тексте эксплицируются в терминах теории вероятностей и математической статистики. К результатам вероятностно-статистического описания, взятым в сочетании с данными лингво-психологического эксперимента, может быть применен аппарат теории информации, с помощью которого удается количественно оценить как структурную организацию текста, так и заключенною в нем смысловую информацию.

Из всего сказанного следует, что ***математическая экспликация*** центральной проблемы современного языкознания «система языка – норма – текст» может быть осуществлена на основе ***применения методов как «качественной», так и «количественной» математики***.

В связи с разработкой лингвистических аспектов искусственного интеллекта возникает необходимость формального описания внешних ситуаций, стимулирующих порождение текста. Для описания этих ситуаций используются как количественная, так и комбинаторная методика.

Что же касается моделирования непрерывных изменений языка во времени (диахроническая лингвистика), географическом пространстве (диалектология), а также в специально-профессиональном и художественном континууме (социолингвистика и стилистика), то целесообразно использовать понятия бесконечного множества, предельного перехода, непрерывности, т. е. понятия, составляющие основу математического анализа.

В области комбинаторной лингвистики наряду с фундаментальными исследованиями появилось уже немало работ типа учебных пособий, в которых систематизируются и популяризируются основные ее идеи [13]; [45]; [56]. В ином положении находится квантитативная лингвистика. Здесь можно указать лишь несколько книг и сборников, в которых исследуются или описываются отдельные вопросы приложения математического анализа, теории вероятностей и статистики в языкознании – см. [4]; [6]; [7]; [15].

Однако систематического изложения основных идей квантитативной лингвистики до сих пор нет. Предлагаемая читателю книга имеет своей целью восполнить этот пробел.

В первой части книги рассматриваются элементы математического анализа и их лингвистические приложения. С помощью этого аппарата строятся математические модели, описывающие: изменение лингвистических объектов во времени (гл. 1–4); распределение информации в письменном тексте (гл. 1, 2, 4), акустическую структуру устной речи (гл. 1).

Во второй части к лингвистическому материалу прилагается аппарат комбинаторики, теории вероятностей и математической статистики. Эта методика используется для: измерения смысловой информации слов и избыточности текста (гл. 5); описания функций распределения в тексте слогов, слов, словосочетаний и грамматических классов (гл. 6); построения статистических моделей текста и вероятностных характеристик норм языка (гл. 8, 9).

Математический аппарат, необходимый для построения всех этих моделей, чаще всего дается в виде определений без строгих математических доказательств, которые читатель всегда может найти в вузовских учебниках и пособиях по математическому анализу [28], теории вероятностей [10]; [14]; математической статистике [30]; [36] и лингво-статистике [б], [7].

Авторы приносят благодарность рецензентам проф. Б. Н. Головину и проф. А. С. Длину, а также доц. П. М, Алексееву и канд. техн. наук К. А. Разживину, замечания которых способствовали улучшению книги. Авторы благодарны В. В. Колесниковой, С. А. Моисеевой, П. В. Садчиковой и коллегам по группе «Статистика речи» за помощь при подготовке рукописи к печати. Кроме того, авторы выражают признательность редактору А. М. Суходскому, проделавшему большую работу по редактированию книги.

ЧАСТЬ ПЕРВАЯ

ИССЛЕДОВАНИЕ ЛИНГВИСТИЧЕСКИХ ПРОЦЕССОВ МЕТОДАМИ КВАНТИТАТИВНОЙ ЛИНГВИСТИКИ

ГЛАВА 1. ИСХОДНЫЕ ПОНЯТИЯ КВАНТИТАТИВНОЙ ЛИНГВИСТИКИ

§ 1. Множество лингвистических объектов

1. Понятие множества. Одно из основных понятий современной математики – понятие *множества.* Оно является первичным, т. е. не поддается определению через другие, более простые понятия. С понятием множества мы встречаемся довольно часто: буквы русского алфавита образуют множество, то же можно сказать о словоупотреблениях*, содержащихся в данном предложении, на данной странице и т. д.

Приведенные примеры обладают одним существенным свойством:

все эти множества состоят из определенного конечного числа объектов, которые мы будем называть *элементами множества.* При этом каждый из объектов данного вида либо принадлежит, либо не принадлежит рассматриваемому множеству. Так, например, буква *ф* вне всякого-сомнения принадлежит множеству букв, образующих русский алфавит, в то время как буква / этому множеству не принадлежит. Множества, включающие только такие объекты, принадлежность или непринадлежность которых к тому или иному множеству не вызывает сомнения, называются *четкими множествами.* Поскольку каждый рассматриваемый объект либо принадлежит, либо не принадлежит к рассматриваемому четкому множеству, эти множества всегда имеют ясно очерченные границы.

---

* В дальнейшем мы будем различать следующие лексикологические понятия: ***словоупотребление, форма слова (словоформа), слово***, а также ***исходная форма слова***. Под *словоупотреблением* понимается цепочка букв, заключенная между двумя пробелами в тексте и имеющая одно значение (омонимические словоупотребления рассматриваются как различные). Полностью совпадающие словоупотребления представляют одну *словоформу*. Слово выступает как некоторый класс (сумма) семантически и грамматически связанных между собой словоформ. Словоупотребление является единицей текста, слово – единицей двуязычного, толкового, энциклопедического и т. д. словаря. В этих словарях слово представлено в так называемой *исходной форме*, в качества которой в русском языке выступает обычно именительный падеж единственного числа – для именных форм и инфинитив – для глагольных форм. Что же касается словоформы, то она используется обычно в качестве единицы частотного словаря.