УДК 658.012.011.56

ТЕКСТ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ КАК ОБЪЕКТ СТАТИСТИЧЕСКОГО АНАЛИЗА

М.Ю. Крыгин

Украинский языково-информационный фонд НАН Украины, Киев, Украина, maxus@zeos.net

Статья посвящена применению статистических методов к задачам автоматической обработки текста естественного языка. Определен круг задач по исследованиям текста на естественном языке, решаемых статистическими методами. Сформулированы теоретические основы применения статистических методов к обработке текстов, описаны модели, приведены примеры решения некоторых задач.

СТАТИСТИЧЕСКИЙ ПОРТРЕТ ТЕКСТА, ЦЕПЬ МАРКОВА, ВЕРОЯТНОСТЬ ПЕРЕХОДА, ГЕНЕРАТОР ТЕКСТОВ, ЭЛЕМЕНТ ТЕКСТА

Введение

Статистические методы в науке используются там, где есть возможность доступа к большим объемам данных, а также, если требуется исследовать некоторое достаточно распространенное явление на относительно небольшом массиве данных.

Данная работа посвящена разработке методов, позволяющих получить статистические характеристики языковых явлений для того, чтобы, например, уметь идентифицировать их или дать оценку о наличии и функционировании определенного явления в определенном тексте или ином фрагменте языкового материала. В данной работе естественноязыковой текст рассматривается нами как конечная последовательность символов из некоторого (конечного) алфавита, определенные фрагменты которой получают в ходе анализа лингвистическую интерпретацию.

Целью статистической обработки естественного языка является математическое описание языковых фактов и явлений, связей между ними, получение набора моделей языка для решения определенных лингвистических задач. Здесь ставится задача определить те языковые явления и факты, которые заслуживают внимания – эта задача, возможно, является одной из самых сложных, поскольку от правильного выбора объектов исследования зависит дальнейший результат работы. В наших исследованиях мы выделяем несколько уровней языка: знаковый, фонетический, лексический, грамматический и так далее.

1. Объекты исследования

Исследование языка на уровне знаковой системы преследует цели вычисления распределений символов и их комбинаций в тексте, выявления «запрещенных» сочетаний, нахождение которых в тексте означает ошибку, моделирования языка на символьном уровне.

Исследование фонетической системы дает, в частности, ответ на вопрос, насколько фонетическая система конкретного языка отвечает, например, его орфографической системе.

Достаточно сложными и разнообразными являются статистические исследования лексической и грамматической систем языка. Эти исследования дают возможность не только сгруппировать слова по частотам и получить частотный словарь слов языка, но и исследовать сочетаемость лексических единиц и грамматических форм, приблизить автоматическое выделение устойчивых словосочетаний (например фразеологизмов и терминов) из текста, реализовать (с определенной точностью) автоматический синтаксический анализ и автоматическое снятие омонимии.

Основным массивом данных, на котором мы проводим исследования, является корпус текстов [4]. На массивах корпуса можно исследовать такие явления как грамматическая и лексическая омонимия, установить свойства текста, которые являются характерными, например для конкретного автора. Приведение текста к форме фонетической транскрипции дает возможность провести достаточно полное статистическое исследование фонетической системы языка. В нашей работе мы предполагаем, что тексты корпуса не содержат ошибок, пропусков и полностью совпадают с оригиналами (то есть мы работаем с выправленными текстами). Тексты корпуса подлежат специальной предварительной разметке согласно целям исследования, которому они подвергаются.

Кроме текстов корпуса, в отдельных случаях исследования проводились на Грамматическом словаре украинского языка [1], в котором представлены все словоформы, имеющие место в языке.

2. Статистический портрет текста

В наших исследованиях используется понятие статистического портрета текста, под которым мы понимаем определенным образом сформированную совокупность статистических признаков текста, полученных путем статистической его обработки по неким заранее установленным принципам.

Множество таких признаков состоит из подмножеств, которые относятся к различным уровням (аспектам) языковой системы и предоставляют данные о статистике:

– знаковой системы, в том числе для фонетического (транскрибированного) варианта текста;

– лексической и грамматической систем;

– семантической системы;

– синтаксической системы.

И так далее.

Определим формально понятие статистического портрета текста.

Пусть у нас имеется некоторый текст T, для которого определена некая система отображений:

$$\sigma_\alpha : T \to T_\alpha, \text{ где } T_\alpha = \bigcup_i T_\alpha^i,$$

где T_α^i – набор параметров, отображающих определенное статистическое свойство текста, например распределение символов текста, вероятности перехода или распределение словоформ; T_α – статистический портрет текста T. Он представляет собой объединение наборов статистических параметров для всех свойств текста, которые исследуются.

Нижайший уровень исследования текста производится на его знаковой системе. Исследования статистических свойств языка на уровне знаковой системы явились одними из первых в статистике языка.

Обозначим через U_M обобщенный алфавит языка M, то есть конечное множество всех символов, которые встречаются при графическом (письменном) отображении текстов данного языка. Это множество является более широким, чем алфавит языка за счет того, что в него входят, кроме букв алфавита, еще и пробел, знаки пунктуации, дефис, специальные символы (включая символы чисел). Таким образом, U_M имеет такую структуру:

$$U_M = B_M \cup P_M \cup S_M \cup \pi,$$

где B_M – обычный алфавит языка M; P_M – множество знаков пунктуации:

$P_M = \{p_1 = '.', p_2 = ',', p_3 = ';', p_4 = ':', p_5 = '...', p_6 = '!', p_7 = '?',$
$p_8 = '-', p_9 = '"', p_{10} = ')', p_{11} = '('\}$; S_M – множество специальных знаков; через π обозначен пробел.

Для украинского языка B_M имеет вид:

$$B_{УКР} = \{b_1 = 'А', b_2 = 'Б', b_3 = 'В', b_4 = 'Г', b_5 = 'Ґ',$$
$$b_6 = 'Д', \ldots, b_{30} = 'Щ', b_{31} = 'Ь', b_{32} = 'Ю',$$
$$b_{33} = 'Я', b_{34} = 'апостроф'\}.$$

Объединение B_M, дефиса и пробела обозначим через A_M:

$$A_M = B_M \cup \alpha \cup \{-\}.$$

Будем рассматривать текст как линейную последовательность элементов: $T = x_1 x_2 \ldots x_N$, N – длина текста в определенных элементах. Элементом текста T в зависимости от уровня его рассмотрения может быть символ алфавита языка, лексема, граммема, семантическое состояние, фразеологическая или синтаксическая конструкция и тому подобное. На тексте T определим распределение его элементов как вероятность встретить определенный элемент в тексте T, то есть отношение количества вхождений данного *элемента* в текст к числу *всех элементов данного типа* в тексте; аналогично зададим распределение сочетаний по j элементов текста как количество вхождений данного сочетания в текст к $N - j$. Определим вероятность перехода из элемента x в элемент y как отношение количества сочетаний этих элементов в тексте к количеству всех сочетаний элементов в тексте по два, первый из которых является элементом x. Аналогично определим вероятность перехода из последовательности элементов $\{x_1 x_2 \ldots x_n\}$ в элемент y как отношение количества вхождений в текст сочетаний элементов $\{x_1 x_2 \ldots x_n y$ (элемент y – фиксирован)} к количеству вхождений сочетаний элементов $\{x_1 x_2 \ldots x_n x$ (x пробегает все множество возможных элементов текста)} На множестве текстов языка M зададим семейство отображений $\sigma_{el}^\lambda : T \to T_{el}, \lambda = 1,2,\ldots 10$ следующим образом:

$$\sigma_{el}^1 : T \to T_{el}^1 = \{(a); P(a)\}$$

$$\sigma_{el}^2 : T \to T_{el}^2 = \{(a_1 a_2); P(a_1 a_2)\}$$

...

$$\sigma_{el}^{10} : T \to T_{el}^{10} = \{(a_1 a_2 \ldots a_{10}); P(a_1 a_2 \ldots a_{10})\}$$

и семейство отображений $\sigma_{mark}^\lambda : T \to T_{mark}, \lambda = 1,2,\ldots 10$ — следующим образом:

$$\sigma_{mark}^1 : T \to T_{mark}^1 = \{P(x_i = a_k / x_{i-1} = a_m)\},$$

i – порядковый номер элемента x в тексте;

$$\sigma_{mark}^2 : T \to T_{mark}^2 = \{P(x_i = a_k / x_{i-1} = a_m, x_{i-2} = a_n)\};$$

i – порядковый номер элемента x в тексте;

...

$$\sigma_{mark}^{10} : T \to T_{mark}^{10} = \{P(x_i = a_k / x_{i-1} = a_m, x_{i-2} = a_n, \ldots,$$
$$x_{i-10} = a_l)\},$$

i – порядковый номер элемента x в тексте}.

Число 10 было выбрано как максимальная длина исследуемых последовательностей элементов текста с тем, чтобы покрыть среднюю длину слова (при рассмотрении знаковой системы текста), а также длину словосочетаний, синтаксических и грамматических конструкций.

Статистический портрет текста представляется в виде объединения:

$$T_\alpha = \bigcup_i (T_{el}^i \cup T_{mark}^i),$$

где T_{el}^i представляют собой распределения сочетаний элементов текста по i; T_{mark}^i представляют собой вероятности перехода, описанные выше.

При рассмотрении знаковой системы текста элементами x являются символы алфавита языка A_M. Для лексического же уровня элементами текста являются уже определенные словоформы, а множеством значений элементов текста – словарь словоформ D_M языка M. $T = w_1 w_2 ... w_{i-1} w_i w_{i+1} ... w_N$, $w_i \in D_M$, N – длина текста в словах, D_M – расширенный словарь языка M, который содержит все возможные в языке словоформы. Для уровня грамматической системы элементами текста являются так называемые грамматические состояния, формальными репрезентантами которых служат грамматические значения, маркированные соответствующими грамматическими категориями. При этом текст представляется в виде:

$$T = g_1 g_2, ..., g_{i-1}, g_i, g_{i+1}, ..., g_N ,$$

$g_i \in GD_M$, N – длина текста в словах, GD_M – грамматический словарь языка M, который для каждой словоформы задает ее грамматическое значение $g = (p, f)$, где p – часть речи, f – грамматическое значение слова.

Для получения грамматических значений слов применяются специальные функции [5], разработанные в Украинском языково-информационном фонде НАН Украины. Основную проблему в данном исследовании представляет явление грамматической омонимии и, как следствие, невозможность автоматически однозначно присвоить словоформе ее грамматическое значение. Неомонимичные же словоформы получают грамматическое значение автоматически. Омонимичным словоформам присваивают грамматические значения в полуавтоматическом режиме: для каждой такой словоформы предлагается выбрать из списка возможных вариантов часть речи, которой они принадлежат, а затем грамматическое значение внутри части речи.

Использование статистических методов позволяет также разработать модель семантики и предложить методику построения семантических структур. Эталоном, который используется для построения семантических структур, в данном исследовании является словарь украинского языка (СУЯ) [2]. Для построения семантической модели текста каждому слову приписывается признак $s = (w_0, w_{num})$, w_0 – начальная форма слова, w_{num} – номер семантического значения в СУЯ. Таким образом, текст T можно представить в виде последовательности признаков: $T^{sem} = s_1 s_2 ... s_N$.

Статистические исследования проводились на материале корпуса текстов [4] и грамматического словаря украинского языка [1] объемом 3,2 млн словоформ. Ввиду того, что полученные данные довольно громоздки и получение их не являлось целью исследования, приводить их здесь не будем. Приведем лишь несколько заключений, которые были сделаны по результатам статистического анализа. Грамматический словарь для статистических исследований был представлен в виде последовательности всех зафиксированных в нем словоформ, разделенных пробелами, например: «*... абазин абазина абазинові абазину абазина абазином ...*». В приведенных ниже результатах учтены лишь те сочетания символов, которые не содержат пробела нигде, кроме начальной и конечной позиции. Исключение последовательностей символов с пробелом посередине связано с тем, что они не отображают свойство языка в целом, так как являются частями соседних словоформ грамматического словаря и не встречаются в естественных текстах в таком виде.

Таблица 1

Статистика буквосочетаний в Грамматическом словаре украинского языка

Длина сочетания (в символах)	Количество уникальных сочетаний, найденных в словаре	Комбинаторно возможное количество уникальных сочетаний	% реализации от теоретически возможного
1	35	35	100
2	1037	1225	84,65
3	13781	42875	32,14
4	93774	1500625	6,25
5	354544	52521875	0,68
6	850049	1838265625	0,05
7	1524535	64339296875	0,0024
8	2222461	2251875390625	0,000099
9	2740430	78815638671875	0,0000035
10	2921764	2758547353515625	0,0000001

Из приведенной таблицы видно, что количество уникальных буквосочетаний, найденных в грамматическом словаре, существенно зависит от ранга (длины) буквосочетания: при увеличении ранга буквосочетаний растет количество уникальных буквосочетаний, причем наиболее существенный рост наблюдается при переходах от ранга 1 до 7, то есть до достижения средней длины слова украинского языка, далее рост незначителен. Количество уникальных буквосочетаний совпадает с комбинаторно возможным только для ранга 1, также оно остается достаточно высоким для ранга 2, для ранга 3 – существенно ниже, а далее наблюдается резкий спад.

В грамматическом словаре почти полностью представлена вся возможная лексика украинского языка и все возможные комбинации символов в пределах слов, однако он не в полной мере отражает статистические свойства речи вследствие того, что там представлены абсолютно все допустимые грамматической системой словоформы, в том числе и редкоупотребляемые. То есть на основе грамматического словаря можно получить все возможные в естественном языке комбинации символов,

но нельзя достоверно оценить их частотность. Для исследования символьной системы на частотность использовался массив текстов из корпуса [4].

Таблица 2

Зависимость между длиной сочетаний и их количеством с абсолютной частотой > 1 (для Грамматического словаря)

Длина сочетания									
1	2	3	4	5	6	7	8	9	10
100	99,8	99,4	98,9	96,6	92,6	86,4	78	66,8	53,7
Среди них с абсолютной частотой больше 1 (в %)									

Таблица 3

Зависимость между длиной сочетаний и их количеством с абсолютной частотой > 1 (для текста)

Длина сочетания									
1	2	3	4	5	6	7	8	9	10
100	98,72	95,96	86,64	60,49	38,80	21,28	10,04	3,8	0,59
Среди них с абсолютной частотой больше 1 (в %)									

Из приведенных таблиц видно, что в грамматическом словаре повторяемость комбинаций символов выше. Это объясняется тем, что в словаре представлены полные парадигмы для всех слов языка, обеспечивающие наличие последовательностей символов, отвечающих корню слова, от 7 до 20 и даже более раз; кроме того, разные слова могут иметь одинаковые словоизменительные морфемы.

На базе полученных T_{zn} был построен генератор текстов, работающий по законам распределения символов в украинском языке:

$$G : T_{zn} \to T^{gen}, G \in \{G_0, G_1, G_2, G_N\}. \qquad (7)$$

В этой формуле символом G обозначен генератор текстов, принимающий значения G_i, из множества различных моделей его работы; T^{gen} – текст, сгенерированный посредством G.

Простейшая модель работы генератора текстов G_0 представляется дискретным стационарным процессом с независимыми значениями и равновероятными состояниями. Следующая модель – G_1 представляет собой дискретный стационарный процесс с независимыми значениями и состояниями, которые имеют вероятности, описанные в T^1_{zn}. На следующих этапах генераторы текстов базируются на модели цепи Маркова [5]. Работа генератора G_2 моделируется односвязной цепью Маркова с вероятностями перехода T^1_{mark}. Пример таким образом экспериментально сгенерированного текста:

“ПЕНИВІДЕНОВЖЕВСТ ВАПІЙТЬОМИТРЦЯ ЗАСІТОМ’Ю Д ПЕМИЙ ДЖКОМЕЙ А ПОДИРДКИ ГОБРОГА”.

Аналогично, генератор G_3 использует модель двусвязной цепи Маркова с вероятностями перехода T^2_{mark}, а генератор G_4 – модель трехсвязной цепи Маркова с вероятностями перехода T^3_{mark}.

Текст:

“ВЖДЕНЕХИТИТТЯМ ВУ МАТНО ВИВ ЯКУ МОВАВРИЖА ВОПІШКАЛА ЛЮБКАЗ СТУТІЙСЬ ДОРЧУ АРИЗНА”

был экспериментально получен в результате работы G_3, а текст:

“ДУМАЛИ ДЯКУСЬ НУТИ НЕ ПОДИНІ СЬОГОМУ КИНУ ХАЙ НЕЇ ТУТ СТАНА ПРИТИ У ВИЛИПА”

был сгенерирован с помощью G_4.

Как видно, последний текст наиболее близок к украинскому тексту. Таким образом, чем выше связность цепи Маркова, используемой в модели работы генератора текстов, тем ближе к реальному генерируемый им текст.

Используя статистические методы можно проводить исследования текста на предмет того, каким языком (или какими языками) написан текст,

Таблица 4

Характеристика работы генератора текстов (на 1000 символов)

Модель работы генератора текстов	Общее количество «слов»	Средняя длина «слова»	Средняя длина реальных слов	Количество реальных слов	% реальных слов
Дискретный стационарный случайный процесс с независимыми значениями и равновероятными состояниями	29	33,51	0	0	0
Дискретный стационарный случайный процесс с независимыми значениями, у которого распределение состояний совпадает с распределением символов текста	137	6,3	1,25	4	2,92
Однородная односвязная цепь Маркова	116	7,62	2,36	11	9,48
Однородная двухсвязная цепь Маркова	119	7,39	2,77	26	21,85
Однородная трехсвязная цепь Маркова	133	6,52	5,05	65	48,87

формальное сравнение текстов с целью выявления плагиата, сравнение стилей (определение авторства), определение темы текста и многое другое. Вид статистического портрета можно выбрать исходя из задач исследования. Например, для исследования фонетической системы не обязательно строить статистический портрет лексической или грамматической систем. Далее приведем примеры наших подходов к решению конкретных задач.

3. Анализ студенческих работ на плагиат

Этот анализ является одним из относительно простых видов анализа текста. Мы полагаем, что при списывании студент не изменяет текст исходной работы, поскольку с одной стороны, сложность редактирования можно сравнить со сложностью написания новой работы, а с другой – если студент отредактировал текст работы, то этот новый текст можно уже рассматривать как новую работу, так как студент не только внимательно ее прочитал, но и понял, и, возможно, высказал некоторые свои мысли.

Для проведения анализа была создана тематически сгруппированная база данных студенческих работ. Текст тестируемой работы сравнивался со всеми работами из базы по данной тематике.

Статистический портрет в данном исследовании состоит из распределения слов и их комбинаций длиной до 3. Пример анализа студенческой работы на плагиат изображен на рисунке ниже:

На рис. 1 в левом окне – текст тестируемой студенческой работы, в правом – текст реферата из базы данных, с которым обнаружено сходство, совпадающие части выделены синим цветом. Программа автоматически проводит анализ текста реферата, результатом которого является вывод о том написана ли работа самостоятельно или списана с другой работы; в случае обнаружения плагиата, программа выдает список работ из базы, с которыми обнаружено сходство, для того чтобы преподаватель смог увидеть степень сходства текстов и показать, из какого источника работа списана. Существует возможность настройки точности работы программы.

4. Сравнение предвыборных программ политических партий и блоков

Статистический портрет для этого анализа включает в себя распределение слов и их комбинаций длиной до 3, основным компонентом является распределение слов по частоте. Для анализа были выбраны предвыборные программы политических

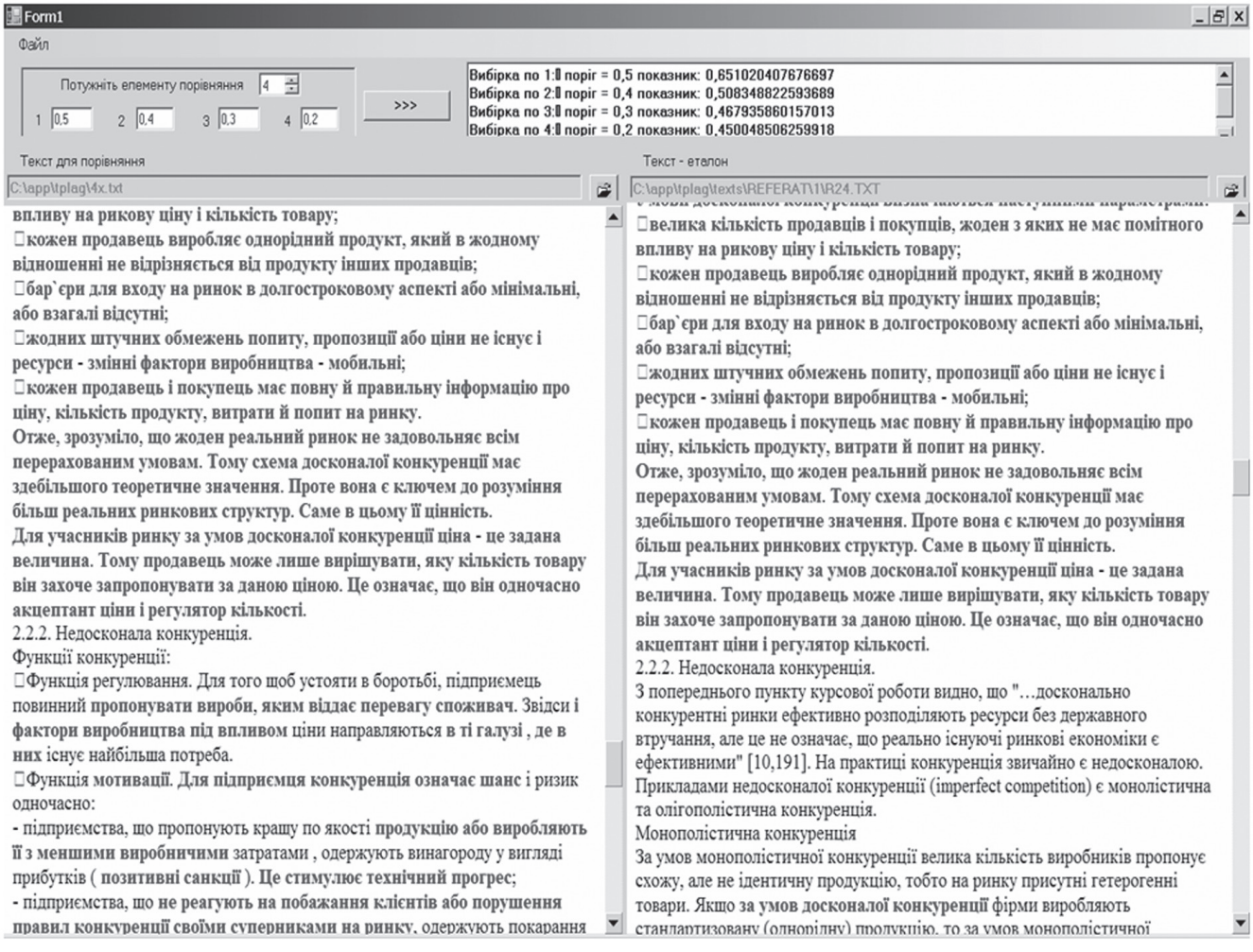

Рис. 1. Результат анализа студенческих работ

партий и блоков к парламентским выборам 2002 года с их официальных сайтов, в сравнении также участвовал текст Конституции Украины.

Программы, участвующие в исследовании, были пронумерованы. Для каждого текста программы политической партии были построены следующие множества: частотный список слов, встречающихся в тексте (первые 100 слов за исключением так называемых стоп-слов – часто употребляемых и служебных слов, не несущих смысловой нагрузки), все пары и тройки слов. Далее частотный список слов для каждой программы политической партии сравнивался с частотными списками слов всех остальных программ. В результате сравнения мы получили квадратную симметричную $k \times k$ матрицу P^1, где k – количество программ партий, участвующих в исследовании. Каждый элемент этой матрицы $p^1[i,j]$ соответствует проценту совпадающих слов в программах с номерами i и j.

$$p^1[i,j] = p^1[j,i],\ p^1[i,i] = 1. \quad (8)$$

Аналогичным образом сравнивались множества пар и троек слов для каждой программы партии со всеми остальными программами. В результате сравнения мы получили квадратные $k \times k$ матрицы P^2, P^3, где k – количество программ партий, участвующих в исследовании. Каждый элемент матрицы P^2

$$p^2[i,j] = (\mathrm{N}^2(i,j)/\mathrm{N}^2(i))*100\%, \quad (9)$$

где $\mathrm{N}^2(i,j)$ – количество пар слов в программе партии с номером i, совпадающих с парами слов партии j, $\mathrm{N}^2(i)$ – общее количество пар слов в программе партии с номером i.

$$p^2[i,j] \neq p^2[j,i],\ p^2[i,i] = 1. \quad (10)$$

Аналогично для матрицы P^3:

$$p^3[i,j] = (\mathrm{N}^3(i,j)/\mathrm{N}^3(i))*100\%, \quad (11)$$

$$p^3[i,j] \neq p^3[j,i],\ p^3[i,i] = 1. \quad (12)$$

Степень сходства программ партий с номерами i и j мы определяем таким образом:

$$p(i,j) = (p^1[i,j] + 2p^2[i,j] + 3p^3[i,j])/6. \quad (13)$$

В результате получаем $k \times k$ матрицу P, которая является результатом сравнения программ партий:

$$P = (P^1 + 2P^2 + 2P^3)/6. \quad (14)$$

Программы политических партий, участвующих в исследовании, были пронумерованы согласно списку:

1. Аграрная партия
2. Народно-демократическая партия
3. Конституция Украины
4. Партия регионов
5. Партия промышленников и предпринимателей
6. Народный Рух Украины
7. Социалистическая партия
8. Трудовая Украина
9. СДПУ(О)
10. УРП Собор
11. БЮТ
12. Наша Украина
13. Коммунистическая партия

В приведенных ниже матрицах элементы округлены до целой части.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	100	48	34	45	52	55	53	46	64	35	28	30	40
2	48	100	29	42	46	51	45	41	61	41	37	39	44
3	34	29	100	27	27	31	29	25	36	33	37	34	38
4	45	42	27	100	45	44	42	39	56	41	38	41	44
5	52	46	27	45	100	43	44	38	53	44	46	46	49
6	55	51	31	44	43	100	31	41	42	51	47	48	54
7	53	45	29	42	44	31	100	38	45	44	44	44	46
8	46	41	25	39	38	41	38	100	52	41	41	44	46
9	64	61	36	56	53	42	45	52	100	58	57	58	63
10	35	41	33	41	44	51	44	41	58	100	34	30	35
11	28	37	37	38	46	47	44	41	57	34	100	34	19
12	30	39	34	41	46	48	44	44	58	30	34	100	31
13	40	44	38	44	49	54	59	46	63	35	19	31	100

Рис. 2. Матрица P^1

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	100	12	7	9	11	13	11	10	17	6	1	4	2
2	6	100	5	9	10	11	7	8	15	6	1	4	0
3	2	3	100	3	3	5	4	3	7	2	0	2	0
4	4	8	5	100	10	9	7	7	13	5	1	4	1
5	4	7	4	7	100	8	7	6	13	4	1	3	1
6	3	6	4	5	6	100	5	5	10	3	0	3	0
7	3	4	4	5	6	6	100	5	11	2	0	2	1
8	4	6	4	6	7	8	6	100	12	4	1	4	0
9	3	5	3	5	6	6	6	5	100	3	0	3	0
10	5	10	6	9	10	12	7	8	14	100	1	4	0
11	4	6	5	6	7	8	7	7	12	6	100	5	0
12	3	7	5	7	8	9	7	8	13	4	1	100	0
13	7	6	6	7	9	7	12	5	14	3	1	31	100

Рис. 3. Матрица P^2

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	100	3	1	1	1	2	2	1	2	0	0	0	0
2	1	100	0	1	1	2	0	1	2	1	0	0	0
3	0	0	100	0	0	1	0	0	0	0	0	0	0
4	0	1	0	100	1	1	0	0	2	0	0	0	0
5	0	1	0	1	100	1	0	0	2	0	0	0	0
6	0	1	0	0	0	100	0	0	1	0	0	0	0
7	0	0	0	0	0	0	100	0	1	0	0	0	0
8	0	0	0	0	1	1	0	100	1	0	0	0	0
9	0	0	0	0	1	0	0	0	100	0	0	0	0
10	0	1	1	1	1	2	0	1	1	100	0	0	0
11	0	0	0	0	0	0	0	1	1	0	100	0	0
12	0	0	0	0	1	1	0	1	2	0	0	100	0
13	1	0	0	0	1	0	2	0	1	0	0	0	100

Рис. 4. Матрица P^3

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	100	14	9	11	13	15	14	12	17	8	5	6	7
2	11	100	7	11	12	13	10	10	16	9	7	8	7
3	6	6	100	6	6	7	6	5	8	6	6	6	6
4	9	10	6	100	11	11	9	9	15	9	7	8	8
5	10	11	6	10	100	10	10	8	14	9	8	9	9
6	10	11	7	9	9	100	7	9	11	10	8	9	9
7	10	9	6	9	9	7	100	8	12	8	7	8	8
8	9	9	6	9	9	10	8	100	13	8	7	9	8
9	12	12	7	11	11	9	10	10	100	11	10	11	11
10	8	11	8	10	11	14	10	10	15	100	6	6	6
11	6	8	8	8	8	11	10	10	14	8	100	7	3
12	6	9	7	9	11	12	10	11	15	6	6	100	5
13	10	9	8	10	12	11	15	9	16	7	4	6	100

Рис. 5. Матрица P

5. Сравнение текстов словарей

В последнее время в связи со значительной активизацией лексикографической деятельности, появлением на рынке словарной продукции, имеющей явно выраженные признаки «клонов» известных словарей, возникает необходимость установить степень отличия между текстами такого рода изданий. Для решения таких задач также успешно применяются статистические методы. При этом возможно сравнение как по отдельным словарным статьям, так и в целом по всему корпусу словаря.

В Украинском языково-информационном фонде НАН Украины выпускается серия словарей «Словники України». В рамках этой серии – ежегодно издается орфографический словарь украинского языка. Составляется новый большой (в 20 томах) толковый «Словник української мови» (СУМ), прототипом которого является изданный в 70-80 годы прошлого столетия 11-томный «Словник української мови». Для ответа на вопрос, насколько отличаются тексты словарей разных изданий, был разработан инструментарий, позволяющий исследовать тексты словарей на предмет сходства и отличия. В основе методики сравнения лежит сравнение множеств цепочек символов, ограниченных пробелами. При исследовании текстов СУМа проводилось сравнение троек словоформ, входящих в 11- и 20-томник. Ниже приведены фрагменты словарных статей с заголовочными словами «АНАРХІЧНО» – «АНАТОМ» из 11- и 20-томника и результат их сравнения, произведенный составленной нами программой. В левом окне – текст 11-томного СУМ, в правом – 20-томного (красным цветом выделены совпадающие части).

В результате анализа выяснилось, что текст первого тома 20-томного СУМа совпадает с соответствующим текстом 11-томника примерно на 37% согласно разработанной нами методике. Таким образом, можно вполне определенно утверждать, что 20-томник – совершенно новый лексикографический труд.

Выводы

Имея лишь текст на естественном языке, с помощью статистических методов, используя только формальные математические модели, можно проводить разнообразные исследования текстов. Процесс исследования текстов целесообразно разбить на три стадии: выбор параметров статистического портрета, статистическая обработка текста, в результате которой получаем статистический портрет с заданными параметрами и непосредственное исследование текстов через их статистические портреты. Получение статистических портретов текстов позволяет, абстрагировавшись от содержания текста, проводить дальнейшие исследования

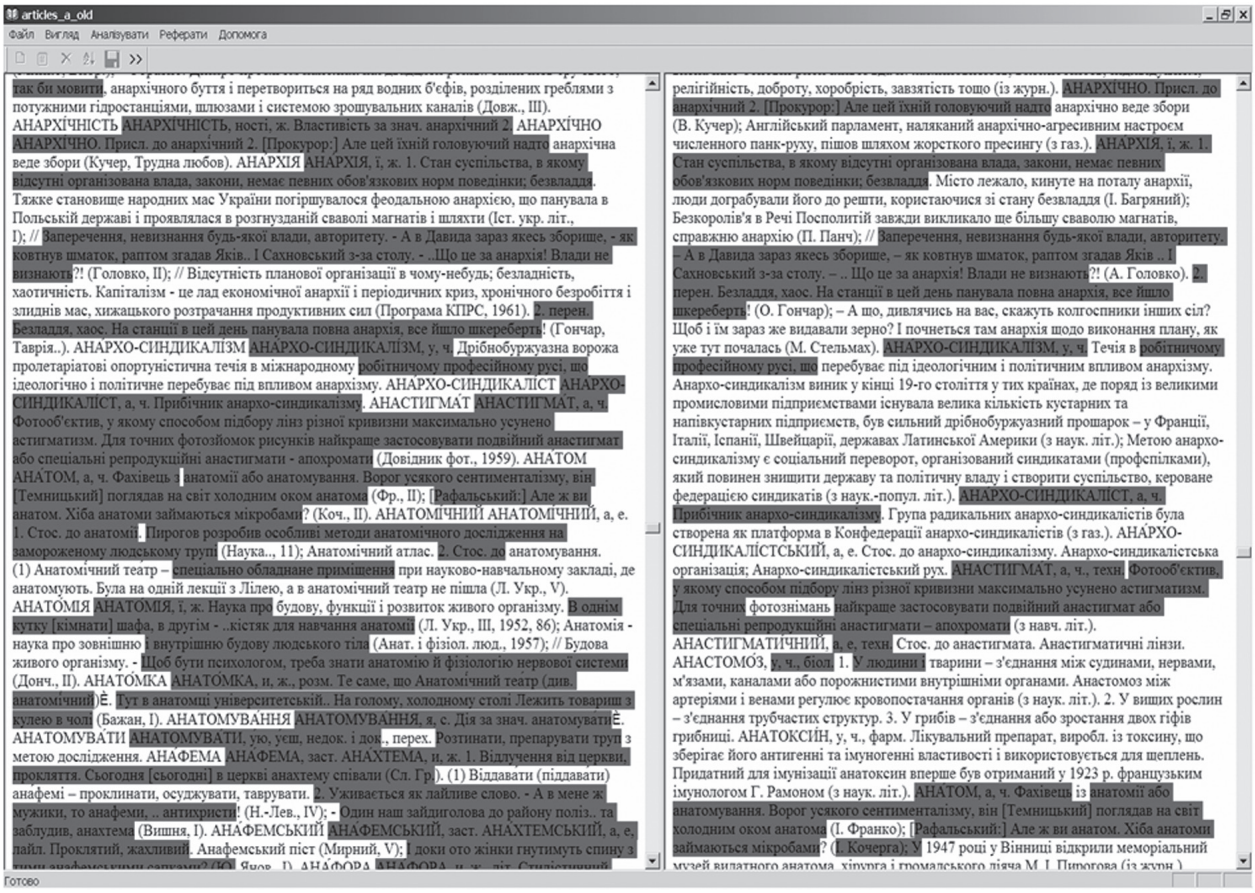

Рис. 6. Сравнение фрагментов 11-томного и 20-томного СУМ

на множествах различных параметров. Параметры, входящие в статистический портрет, выбираются в зависимости от цели исследования. Полагаем, что статистические методы являются наиболее эффективными при проведении различного рода лингвистических экспертиз.

Список литературы: 1. *Шевченко, И. В.* Электронный грамматический словарь украинского языка [Текст] / И.В. Шевченко, А.Г. Рабулец, В.А. Широков // Труды международной конференции «Megaling'2005. Прикладная лингвистика в поиске новых путей». 27 июня – 2 июля 2005 года. Меганом, Крым, Украина. – 2005. – С. 124–129. **2.** *Словник* української мови в 11 т. [Текст]. – К.: Наук. думка, 1970–1980. **3.** *Широков, В. А.* Корпусна лінгвістика [Текст] / В.А. Широков, О.В.Бугаков, Т.О.Грязнухіна та ін. – К.: Довіра, 2005. – С. 126-169. **4.** *Рабулець, О. Г.* Лінгвістичний корпус як середовище обробки та збереження інформаційних ресурсів [Текст] / О.Г. Рабулець, Н.М. Сидорчук // Прикладна лінгвістика та лінгвістичні технології. – К.: Довіра, 2008. – с. 316-328. **5.** *Вентцель, Е. С.* Теория случайных процессов и ее инженерные приложения [Текст] / Е.С. Вентцель, Л. А. Овчаров. – М.: Наука, 1991. – С. 98-127.

Поступила в редколлегию 23.03.2010 г.

УДК 658.012.011.56

Природномовний текст як об'єкт статистичного аналізу / М. Ю. Кригін // Біоніка інтелекту: наук.-техн. журнал. – 2010. – № 1 (72). – С. 75–82.

У статті описано статистичний підхід до розв'язання деяких задач, пов'язаних з обробкою природного тексту. Представлено теоретичне обгрунтування застосованого підходу, що базується на понятті статистичного портрета тексту. Відображено результати дослідження статистичних властивостей текстів, на які спирається наш підхід.

Табл. 4. Іл. 6. Бібліогр.: 5 найм.

UDK 658.012.011.56

Natural language text as an object of statistic analysis / M. Yu. Krygin // Bionics of Intelligence: Sci. Mag. – 2010. – № 1 (72). – P. 75–82.

In the article statistic approach to solving problems concerned with natural texts processing is described. Theoretical basis for used approach based on concept of statistic portrait of text is represented. Results of research of statistic properties of texts that underlie our approach are described.

Fig.: 6. Tab.: 4. Ref.: 5 items.