чения при помощи аналогии с песчаной насыпью в сочетании с мембранной аналогией [3]. Найденные области контакта при заданных в соответствии с аналогией начальных данных определяют пластические зоны сечения.

ЛИТЕРАТУРА

1. Бреббия К., Теллес Ж., Вроубел Л. *Методы граничных элементов.* - М.: Мир. – 1987. – 524 с.

2. Черноусько Ф.Л., Баничук Н.В. *Вариационные задачи механики и управления.* - М.: Наука. – 1973. – 236 с.

3. Малинин Н.Н. *Прикладная теория пластичности и ползучести.* - М.: Машиностроение. – 1975. – 400 с.

**А. Ф. Ахметова, Р. С. Якушев (Казань)**

## МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ СЛОГА ПРИ СТАТИСТИЧЕСКОМ АНАЛИЗЕ ТЕКСТА

Одной из интересных задач компьютерной технологии по обработке информации является сканирование и оцифровка текстов, представленных на твердых носителях. При распознавании текстов, собственно слов, которые представляют совокупность букв, возникает необходимость в предварительной лингвистической обработке воспринимаемых и оцифровываемых знаков, которые подчиняются определенным грамматическим правилам языка.

Первые работы по теории распознавания имели практический характер. Одними из первых для распознавания печатных букв были предложены методы, основанные на следующей идее: изображение буквы сравнивается путем наложения на маски – трафареты, определенные для всех символов-знаков алфавита. По критериям, характеризующим степень совпадения изображения с маской, логическая схема вырабатывает решение о том, какая буква предъявлена для распознания. Дальнейшее развитие компьютерных технологий естественно привело исследователей к более гибким методам, использующим правила языка. Ведь тексты написаны по грамматическим правилам конкретного языка, поэтому было естественным разработан алгоритм поиска распознаваемого знака прежде всего среди букв, ожидаемых по правилам грамматики.

Мы в своей работе в качестве объекта исследования взяли слог. Была сформулирована задача построения модели слога, для распозна-

вания текста. Слог – одно из основных понятий языка. Для определения слога и для разбиения слова на слоги в языкознании существуют много разных подходов. Суть нашей работы – математическое моделирование и изучение квантитативных характеристик слогов в контексте письменного текста.

Рассмотрим множество, состоящее из согласных и гласных звуков алфавита, из этих звуков будем строить слоги. Слог это совокупность гласных и согласных звуков, с единственным ограничением – слогообразующей фонемой является одна гласная фонема (т.е. слоги не могут образовываться без гласной фонемы). Учитывая данное ограничение определяем типы слогов. Например: Г, ГС, СГ, ГСС, ... и т.д.

В реальном языке при предположении существования слогов, например до k букв, можно предположить возможность образования k(k+1)/2 подмножеств слогов. Это только по двоичному признаку буквы делятся на гласные и согласные. На самом деле гласные и согласные в естественном языке в свою очередь различаются по разным признакам, поэтому реальных вариантов подмножеств слогов получается значительно больше.

Каждый язык характеризуется определенными типами слогов [1]. Количество этих типов в языке различно и зависит от числа элементов слога и существующих между ними отношений, что составляет структуру слога.

При изучении слоговых элементов текстов квантитативными методами возникает задача выявления количественных и качественных критериев отбора текстов для статистического анализа. При определении достаточного объема выборки для исследования текстов мы использовали методику, выработанную Р. Г. Пиотровским и К. Б. Бектаевым, с уточнениями, применительно к слогам [2].

По этой методике "объем выборки определяется исходя из требований к покрываемости наугад взятого текста наиболее частыми единицами составляемого списка" и вычисляется по формуле, предложенной Ципфом:

$$N = \frac{A^2 \cdot (1-f)}{g^2 \cdot f},$$

где $N$ – объем текста;

$A$ – постоянная, значение которой берется по заданному значению доверительной вероятности (надежности),

$f$ – относительная частота лексической единицы, значение которой равно значению нижней границы достоверных частот F,

$g$ – относительная (допустимая) ошибка.

Используя данную формулу и постоянные величины, которые выработаны в современной лингвостатистике (такие вычисления для случая слов проведены в работе [3]), мы определили необходимый объем выборки для статистического исследования текста на татарском языке. С этой целью был составлен частотный словарь слогов. Оказалось, что 70%-ное покрытие татарского текста дает начальная зона словаря в 24376 слогов, при этом нижним порогом этого массива является $f = 0.0011$.

Вычисления дают:

$$N = \frac{1{,}96^2 \cdot (1 - 0{,}0011)}{0{,}33^2 \cdot 0{,}0011} = 32034 .$$

Итак, для статистического исследования достаточна выборка объемом в 32034 слога, при этом средняя длина слова в татарском языке три слога. Отсюда получаем с округлением, что достаточна выборка в 10000 слов.

В результате получены следующие основные результаты:
– предложен алгоритм построения слога;
– получены формулы для подсчета количества слогов той или иной модели слога;
– выявлен статистический закон распределения различных моделей слогов на базе русского и татарских языков.

Предложенная модель построения слога может быть использована для решения задач сканирования и распознавания текстов. Полученные результаты позволяют производить квантитативный анализ текстов различных стилей на базе реальных языков.

ЛИТЕРАТУРА:

1. Сторчак Л.В. *Модели слоговых структур русского литературного языка//* Автореферат диссертации на соискание ученой степени кандидата филологических наук. – Ростов на Дону: РГУ, 1991.

2. Пиотровский Р. Г., Бектаев К. Б., Пиотровская А. А. *Математическая лингвистика.* – М.: Высшая школа, 1977.

3. Ризванова Л.М., Якушев Р.С., Хадиев Р.М. *Использование АРМ «КЭЛИМЭ» в лингвистических исследованиях татарского текста//* Модели национальных языков. Труды научного семинара «Формально-логические и компьютерные модели языков» в рамках российской конференции по искусственному интеллекту «КИИ - 96»./ Казань: Фэн, 1996. – С. 88-94.