АВТОМАТИЧЕСКОЕ ФОРМИРОВАНИЕ БАЗЫ СОЧЕТАЕМОСТИ СЛОВ НА ОСНОВЕ ОЧЕНЬ БОЛЬШОГО КОРПУСА ТЕКСТОВ

Клышинский Э.С., Кочеткова Н.А., Литвинов М.И., Максимов В.Ю.

Гипотезы в основе метода

- Синтаксическое подчинение некоторых групп слов может быть выявлено без проведения синтаксического анализа. Это справедливо для морфологически однозначных слов (однозначность части речи).
- В тексте большого объема однозначные группы слов будут встречаться достаточно часто для получения статистически значимых результатов.

Анализируемые группы

1. Следующая за единственным глаголом группа существительного синтаксически подчиняется данному глаголу.
2. Единственная группа существительного, расположенная в начале предложения перед единственным глаголом, синтаксически подчиняется данному глаголу.
3. Прилагательные, расположенные перед первым в предложении существительным или между глаголом и существительным, синтаксически подчиняются данному существительному.
4. Эти же положения могут быть применены к деепричастиям и причастиям.

Объем обработанных источников

Источник	**Объем, млн словоупотреблений**
Библиотека Мошкова	680
РИА Новости	156
Доп. корпус прозы	120
Независимая газета	89
Лента.ру	33
Российская газета	29
PCWeek	28
РБК	21
Компьюлента	9
Итого	**1165**

Результаты (по количеству вхождений)

Числитель – общее количество обнаруженных вхождений, знаменатель – количество уникальных сочетаний.

Пара	Всего вхождений, млн	>1 повторения, млн	>2 повторений, млн
Глагол+сущ.	65 / 8,3	60,3 / 3,5	57,7 / 2,3
Деепр.+сущ.	3,5 / 0,88	2,8 / 0,31	2,6 / 0,18
Сущ.+прил.	9,9 / 1,3	9,2 / 0,56	8,8 / 0,36

Статистика употреблений по частям речи

Часть речи	Приняло участие	Всего в морфологии
Глагол	21500	26400
Сущ.	53300	83000
Прил.	23700	45300

Наиболее часто встречающиеся в новостных текстах сочетания вида:

1) Глагол + существительное

Сочетание	Встречаемость	Сочетание	Встречаемость
СООБЩИТЬ РИА	624691	ПРИНЯТЬ РЕШЕНИЕ	140090
ПЕРЕДАВАТЬ КОРРЕСПОНДЕНТ	327903	ГОВОРИТЬСЯ В СООБЩЕНИЕ	132385
ПОКАЧАТЬ ГОЛОВА	304597	СООБЩАТЬ АГЕНСТВО	118615
ПРИНЯТЬ УЧАСТИЕ	271250	СКАЗАТЬ СОБЕСЕДНИК	115959
ИМЕТЬ В ВИД	201167	ИДТИ РЕЧЬ	108306

Наиболее часто встречающиеся в новостных текстах сочетания вида:

1) Существительное + прилагательное

Сочетание	Встречаемость	Сочетание	Встречаемость
БЛИЖАЙШИЙ ВРЕМЯ	23664	БОЛЬШОЙ КОЛИЧЕСТВО	17737
ПРАВЫЙ РУКА	19809	ОФИЦИАЛЬНЫЙ САЙТ	17385
ОФИЦИАЛЬНЫЙ ПРЕДСТАВИТЕЛЬ	19489	ЛЕВЫЙ РУКА	16121
ПОСЛЕДНИЙ ВРЕМЯ	18555	ИНФОРМАЦИОННЫЙ АГЕНСТВО	15503
ВОЕННЫЙ СЛУЖБА	17933	МОЛОДОЙ ЧЕЛОВЕК	14699

Причины возникновения ошибок

- Часть из ошибок объясняется не совсем корректной обработкой некоторых видов конструкций.

  Так в предложении «Хочу от лица коллектива поздравить юбиляра» конструкция «от лица» ошибочно относилась к глаголу «хотеть».

- Ассоциации, гиперболы и другие выразительные средства литературного языка. Будучи оторванными от контекста, подобные конструкции удивляют, хотя их выделение с точки зрения приведенных выше шаблонов проводится вполне корректно.

  Месяц гладит камыши

  Сквозь сирени шалаши...

- Ошибки авторов

Результаты (процент ошибок)

Количество ошибок не превышает 1%.

В области наиболее частотных сочетаний ошибки метода составляют порядка 0,1%, тогда как сочетания, встретившиеся только один раз, выделяются с примерно 1-2% ошибок.

Выводы

- Несмотря на то, что для построения баз было использовано около 1,5% всех словоупотреблений, большой объем корпуса позволил получить представительный результат.
- Проведенные эксперименты показали, что выдвинутые гипотезы вполне корректны, хотя и носят вероятностный характер.
- Точность получаемых результатов составляет порядка 99%.