Коллокации и конструкции в исследовании структуры текста

Лидия Пивоварова
lidia.pivovarova@gmail.com
Елена Ягунова
iagounova.elena@gmail.com

Коллокации и конструкции

Что такое для нас коллокации и конструкции?

- **Это сочетания двух и более лексических единиц, которые выделяются нами из текста на основании статистических критериев и/или экспериментов с информантами.**

Вопросы:

- Как они соотносятся со словарем и/или грамматикой?
- С номинативностью и/или предикативностью?

Положения-гипотезы:

- Нечеткость границ между **коллокациями и конструкциями**
- Типовые или ядерные **коллокации** и **конструкции** противопоставлены как парадигматические vs. синтагматические единицы
  - единицы, принадлежащие лексикону vs. синтаксису

Структурные составляющие текста

И коллокации и конструкции – **структурные составляющие текста (**и шире – **коллекции текстов)**

Опора на следующие виды **контекста**:

- минимальный контекст
  - в нем реализуются лексические и морфолого-синтаксические явления;
- текст (текстовый контекст)
  - включает в себя фрагменты текста вплоть до текста целиком;
- коллекции текстов
  - предполагает учет текстов определенного типа, т.е. формирование коллекций и подколлекций

Методика экспериментов

Вычислительный эксперимент, его результат – наборы связанных сочетаний

- связанные сочетания для коллекции (в виде частотного словаря)
- связанные сочетания для каждого текста отдельно (собственно структурные составляющие текста)

Эксперимент с информантами, его процедура – определение степени связанности путем шкалирования,

- результат связанные сочетания для отдельных текстов

Идея сопоставления вычислительного эксперимента и эксперимента с носителями языка

Степень связанности неоднословной единицы зависит от вероятностной модели, описывающей ее появление в ходе процедур обработки текста

- статистические характеристики должны описывать данные в зависимости от типа контекста: минимальный контекст – текст – коллекция текстов

Носитель языка имеет интуитивные представления о связанности (степени неслучайности) сочетаний слов в тексте

- может «подключать» текстовые базы по текстам разных функциональных стилей
- воспринимает каждый конкретный текст с точки зрения соответствия некоторой текстовой базе адресата.

Данные (процедура анализа)

Нами оценивались следующие данные:

- данные, полученные в ходе вычислительных экспериментов:
  - список наиболее связанных n-грамм по коллекции;
  - список наиболее связанных n-грамм по подколлекции (подколлекция является тематически более однородной, чем исходная коллекция);
  - отдельные тексты, представленные в виде последовательности связанных сочетаний, («сегментов» в терминологии автора программы);
- отдельные тексты, представленные в виде последовательности связанных сочетаний, полученных в ходе эксперимента с информантами

Гипотезы

- с увеличением степени однородности увеличивается объем n-грамм (увеличивается n)
  - коллекция→ однородная коллекция→текст
- с увеличением степени однородности
  - увеличивается число конструкций (в соотношении конструкция vs. типовая коллокация),
  - увеличивается число предикативных сочетаний;

  для отдельно анализируемых текстов
- сходство наборов связанных сочетаний по результатам вычислительного эксперимента и эксперимента с информантами,
- большее число предикативных сочетаний по результатам экспериментов с информантами
  - меньшее по результатам вычислительного эксперимента

- Мы обсудим с Вами полученные результаты
- Мы предложим интерпретацию результатов с разных точек зрения: конструкционной, лексико-грамматической и информационной

- а еще… покажем сходство и различие между двумя интереснейшими типами контекста: единичным текстом и кластером (или сюжетом, т.е. максимально однородной тематической подколлекцией)

Продолжение следует...

Подходите к нашему стенду!