Л.М. Пивоварова, Е.В. Ягунова (СПбГУ) Информационная структура научного текста. Текст в контексте коллекции // Труды международной конференции «Корпусная лингвистика–2011». – СПб.: С.-Петербургский гос. университет, Филологический факультет, 2011

Л.М. Пивоварова, Е.В. Ягунова (СПбГУ)

ИНФОРМАЦИОННАЯ СТРУКТУРА НАУЧНОГО ТЕКСТА. ТЕКСТ В КОНТЕКСТЕ КОЛЛЕКЦИИ

1. Введение

Тема доклада определяется желанием понять особенности структуры научного текста (в отличие от новостного или художественного текстов). Наша задача состояла в определении информационно наиболее важных структурных составляющих: ключевых слов, характеризующих смысл текста, и терминологически нагруженных коллокаций, характеризующих тематику коллекции.

В докладе мы продолжаем изучение роли контекста. Применительно к его теме и задачам мы ограничиваемся изучением (1) текста (его структуры) как контекста для ключевого слова (или терминологической коллокации) и (2) коллекции как контекста для текста.

2. Материал и методика

Материал [1]: (1) Монотематическая коллекция материалов конференции «Корпусная лингвистика» 2004-2008 года (далее КЛ); (2) Подвыборка: 10 текстов из данной коллекции; (3) В качестве контрастивной коллекции использовались труды

---

[1] Пользуясь случаем, хотим поблагодарить кафедру «Математической лингвистики» филологического факультета СПбГУ, и лично В.П.Захарова и О.А. Митрофанову за любезно предоставленную нам для работы коллекцию текстов «Корпусная лингвистика».

конференции «Диалог» 2003-2009 в которой реализуется большее количество лингвистических тем.

Вычислительный эксперимент основывается на мере TF-iDF как традиционном коэффициенте важности слов.

Эксперимент с информантами проводился по традиционной методике «Прочитайте текст. Подумайте над его содержанием. Выпишите 10-15 слов, наиболее важных для его содержания» (более 21 информанта).

Каждое выделенное ключевое слово (КС) оценивалось в соответствии с его весом (коэффициентом значимости): (1) значение меры TF-iDF (КС1) и (2) доля информантов, выделивших слово как ключевое (КС2).

Группа информантов состояла из студентов кафедр «Математическая лингвистика»[1] и «Информационные системы в искусстве гуманитарных науках», знакомых с предметной областью корпусной лингвистики.

Для определения терминологически нагруженных биграмм использовались меры MI и t-score. Мера MI часто используется для выделения неоднословных терминов на основании сочетаемостных ограничений. Мера t-score (опирающаяся на частоту встречаемости с поправочным коэффициентом) позволяет выделить терминологические сочетания, характерные для коллекции в целом (всех текстов или большинства из них)[2].

2. Результаты. Выводы

В работе были получены разнообразные данные, верифицирующие положение о роли контекста в анализе текстов и полнотекстовых коллекций. Предложенная методика позволила

---

[1] Пользуясь случаем, хотим поблагодарить О.А. Митрофанову за участие в проведении эксперимента среди студентов-матлингвистов.

[2] Подробнее см. Пивоварова Л.М., Ягунова Е.В. Извлечение и классификация терминологических коллокаций на материале лингвистических научных текстов (предварительные наблюдения) // Терминология и знание. Материалы II Международного симпозиума (Москва, май 2010 г.). М., 2010

изучить основные особенности информационной структуры научного текста (рассматриваемой предметной области). Оценивалась степень компактности информационной структуры текстов и коллекций и место информационной структуры текста в структуре коллекции.

Разделение текстов на тематически центральные и периферийные осуществлялась нами на основании сопоставления характеристик текста и коллекции. Характеристики коллекции: общая тематика коллекции (по термнологически нагруженным t-score-коллокациям) и основные неоднословные термины (по MI-коллокациям). Характеристики информационной структуры текста анализировались через наборы КС1 и КС2.

Для иллюстрации приведем некоторые данные. Общая характеристика тематики анализируемой коллекции (топ-список по мере t-score, в порядке убывания значения меры, после применения частеречного фильтра): *русского языка, корпуса текстов, части речи, английского языка, машинного перевода, предметной области, лексических единиц* и т.д.

Основные неоднословные термины коллекции из списка MI-коллокаций, выделяющихся и для лексем, и для словоформ (топ-список в порядке убывания значения меры): *речевой деятельности, художественной литературы, корпусная лингвистика, имена собственные, математической лингвистики, словарной статьи, предметной области, машинного перевода, семантических состояний, разрешения неоднозначности, английский язык, Национальный корпус, грамматических категорий, устная речь, база данных, лексических единиц, русский язык, корпусные данные, частей речи, морфологической разметки* и т.д.[1]

Анализ терминологических коллокаций как характеристик тематической области коллекции позволяет сделать вывод о

[1] См. подробнее Пивоварова Л.М., Ягунова Е.В. Извлечение и классификация терминологических коллокаций … М., 2010

монотематичности [1] и о компактности информационной структуры коллекции. MI- и t-score-коллокации позволяют определить такие подобласти структуры как типичные цели и задачи корпусных исследований, материал и методы, корпусные ресурсы. Эти подобласти информационной структуры коллекции соотносятся с соответствующими классами ключевых слов (КС1 и/или КС2 [2]. А наборы ключевых слов характеризуют информационную структуру единичного текста в контексте коллекции (особенно для КС1). Напомним, что КС1 – результат вычислительного эксперимента, а КС2 – эксперимента с информантами.

Примером текста, тематически центрального в коллекции, и текста, оказывающегося на периферии, могут служить «текст1»[3] и «текст 2»[4]. Отнесение текстов к этим группам следует на основании (1) анализа каждого из двух наборов ключевых слов (КС1 и КС2), (2) сравнения двух наборов (см. табл. 1) и (3) сопоставления характеристик текста с характеристиками коллекции в целом (наборов ключевых слов и терминологических коллокаций).

---

[1] См. подробнее Пивоварова Л.М., Ягунова Е.В. Извлечение и классификация терминологических коллокаций… М., 2010

[2] См. подробнее в Ягунова Е.В. Формальные и неформальные критерии вычленения ключевых слов из научных и новостных текстов // Русский язык: исторические судьбы и современность: IV Международный конгресс исследователей русского языка (Москва, МГУ им. М.В.Ломоносова, филологический факультет 20-23 марта 2010 г.): Труды и материалы / Составители М.Л. Ремнева,, А.А Поликарпов. – М.: Изд-во Моск. ун-та. 2010, с. 533-534

[3] Сидорова Е.А. Подход к построению предметных словарей по корпусу текстов // Труды международной конференции «Корпусная лингвистика–2008». СПб.: 2008

[4] Смирнов А.О., Смирнова О.Ю. Программное обеспечение в процессе изучения фонотактики иностранного языка // Труды международной конференции «Корпусная лингвистика–2008». СПб.: 2008

Текст 1. КС2 с максимальным весом (коэффициентом значимости) в большинстве соответствуют той терминологии, которая определялась через терминологические коллокации: *словарь, корпус, текст, термин,* и т.д. КС1 содержит слова, значимые для текста 1 в сопоставлении с рассматриваемой коллекцией, несмотря на это в него попадают ключевые слова из информационной структуры коллекции, т.е. эти слова в тексте встречаются достаточно часто, чтобы мера tf-idf для них принимала высокие значения.

На пересечении наборов КС1 и КС2 лежат: *термин, модуль, обучение, тематизация, конкорданс, иерархия, встречаемость, наполнение, статистический, словарь, классификация, предметный*, т.е. это пересечение составляет около 50% от каждого из списков. Это пересечение не противоречит информационной структуре коллекции и на основании формального совпадения – полного или частичного (например, с точностью до части речи), и на основании содержательной интерпретации (объединения слов и словосочетаний в подклассы с максимально близким смыслом).

Текст 2. КС2 с максимальным весом очень редко соответствуют той терминологии, которая определялась через терминологические коллокации. Еще меньше таких сопоставлений для набора КС1 (среди КС1 больше всего представлена фонетическая терминология). На пересечении наборов: *звук, фонема, комбинация, сочетание*, т.е. не более 16% (см. табл. 1).

Таблица 1. Сопоставление наборов КС1 и КС2 для текстов 1 и 2 [1]

текст 1 (центр)		текст 2 (периферия)	
КС1	КС2	КС1	КС2
термин	**словарь**	**звук**	язык
модуль	корпус	**фонема**	**фонема**
обучение	текст	**комбинация**	фонотактика
тематизация	**термин**	согласный	поиск
конкорданс	**конкорданс**	взрывной	иностранный
статистика	анализ	задний	**звук**
иерархия	**статистический**	преграда	***транскрипция***
тема	автоматический	сонант	программа
неразмеченный	**модуль**	***транскрипционный***	словарь
просматривать	**обучение**	передний	**комбинация**
обучать	**тематизация**	позиция	анализ
сообщение	разметка	редукция	оболочка
встречаемость	**предметный**	помочь	***звуковой***
наполнение	обработка	**сочетание**	система
статистический	структура	британский	текст
выборка	**иерархия**	гласная	статистический
словарь	информация	иноязычный	***сочетаемость***
дообучение	лексический	иностранный	поисковая
словокомплекс	**наполнение**	безусловно	электронный
терминообразовать	**классификация**	английский	**сочетание**
классификация	технология	альвеолярный	перекодировка

[1] Ключевые слова в таблице расположены в порядке убывания коэффициента значимости (веса). Слова, относящиеся к пересечению наборов КС1 и КС2 выделены п/ж шрифтом; слова, совпадающие в разных списках с точностью до части речи, дополнительно выделены курсивом.

наследование	**встречаемость**	альвеолярный-палатальный	модель
механизм	выборка	англичанин	изучение
документооборот	частота	апикальный	слово
словарный	создание	аффриката	проблема
предметный	**словарь**	боковой	
		велярный	

LM Pivovarova, EV Yagunova (SPbSU)

INFORMATION STRUCTURE OF THE SCIENTIFIC TEXT. TEXT IN THE CORPUS

We are aimed to determine the information structure of the scientific text and the scientific corpus ("Corpus Linguistics" Proceedings), as well as the nature of the interaction and cooperation between these two structure types. We continue studying the role of context. We use terminological collocation (MI and/or t-score) to define the main features of the information structure of the monothematic scientific corpus. The information structure of the text was defined according two types of keywords: (1) weighted by TF-iDF and (2) found during experiments with informants. Results are presented through specifying the position of the particular texts in the text corpus (according to their information structure).