**Семинар 26 апреля 2017**

**Е.А.Смирнова**

**«Оценка репрезентативности текстовых корпусов»**

сл. 2 Корпус определяется как … subset – разновидность

Если с понятием корпуса все более ли менее ясно, то вопрос о его размере до сих пор остается открытым. Как видно из определений репрезентативность является ключевым понятием для корпуса, но что это и как ее достичь, работая с корпусом текстов, нам предстоит выяснить.

сл. 3 Репрезентативность, по определению Байбера, это степень включения в выборку изменчивости / вариативности, наблюдаемой в множестве / совокупности.

сл. 4 В корпусной лингвистике выделяется ситуационная изменчивость, т.е. разнообразие типов текстов в языке; и лингвистическая изменчивость – разнообразие лингвистических явлений, представленных в языке. Эти признаки могут быть положены в основу оценки качества корпуса.

Вторым важным для репрезентативности понятием является изучаемая совокупность. Она рассматривается в двух аспектах: границы совокупности и ее иерархическая организация. Байбер отмечает, что при составлении корпусов этим вопросам часто не уделяется должного внимания, в результате оценивание репрезентативности корпуса становится невозможным.

Сл. 5 В литературе выделяется два вида репрезентативности – качественная и количественная. Остановимся на них подробнее. Количественная репрезентативность. Проблемы с этим видом репрезентативности могут возникнуть, если тексты для корпуса низкого качества и взяты из ненадежных источников. При составлении корпуса исследователь должен четко определиться с критериями отбора материала для корпуса, а именно… Как только критерии определены, можно переходить к сбору данных, который проходит в несколько этапов…

сл.6 Качественная репрезентативность. По мнению Левида, размер корпуса является определяющим фактором его репрезентативности. Однако в корпусной лингвистике продолжаются дебаты по данному вопросу.

сл. 7 для вычисления оптимального размера корпуса предлагается использовать следующую формулу

S – оценка стандартного отклонения для всей совокупности

te – допустимая погрешность (1/2 от выбранного доверительного интервала)

t – значение t для желаемого уровня вероятности

сл 8

Но для применения данной формулы в корпусной лингвистике, необходимо решить ряд возникающих проблем, а именно:

сл. 9 Другой способ определения качественной репрезентативности корпуса основан на понятии «лексической плотности». Для того, чтобы говорить о ней, необходимо ввести ряд понятий:

сл 10

На основе закона Зипфа испанскими учеными была разработана прогрaмма N-Cor Оценивание корпуса апостериори.

сл 12

Проверяется два раза – когда файлы расположены по алфавиту и в случайном порядке. В точке, где красная и синяя линия стабилизируются можно утверждать, что корпус является репрезентативным. На графике А – видно, что корпус стал репрезентативным после включения в него 25 документов – точка, после которой кривые не варьируются, которая ближе всего к нулю. График Б показывает минимальное количество слов (токенов) необходимое для того, чтобы считался репрезентативным. В данном случае – около 300 000 слов.

сл. 13 На данный момент мы работаем с двумя корпусами. The reference corpus represents articles from leading journals in Management – Academy of Management Journal (AMJ) – 17 articles; Journal of Management (JM) – 25 articles; Journal of Management Studies (JMS) – 20 articles. All of them are well-established international academic journals with current impact factors of 6.233; 6.051 and 4.131 (2016) respectively. The journals are highly influential and their readership comprises the international business and management studies discourse community. All publications refer to the period from 2013 to 2016.

Хорошо бы добавить в papercat функцию оценки их репрезентативности.