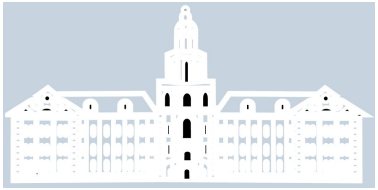


Лекция 2

Поиск в корпусе

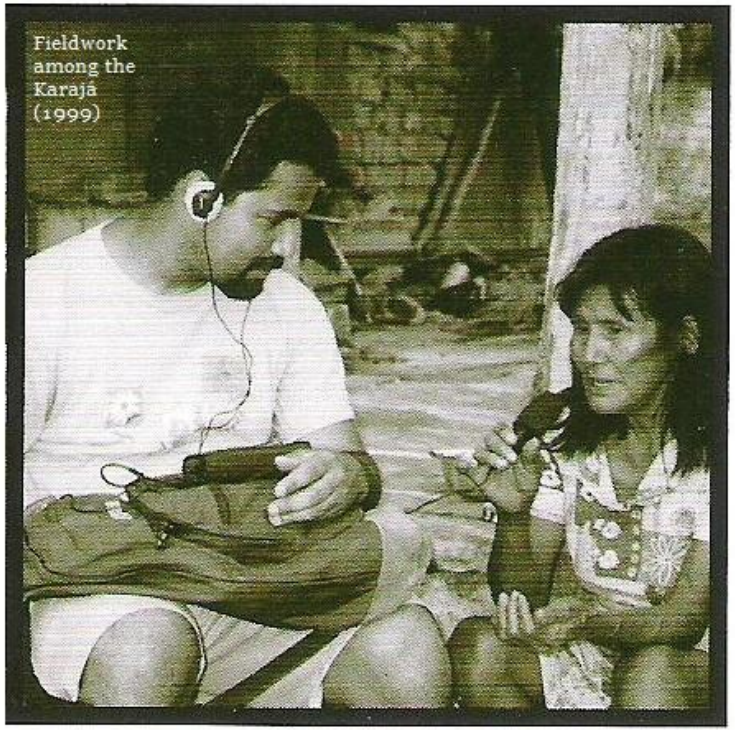
Ольга Ляшевская ** olesar@yandex.ru

Курс “Лингвистические данные”, 1 курс ФиКЛ ВШЭ

В прошлой лекции

Основные типы данных

- грамматика
- словарь
- корпус
- наборы стимулов для экспериментов
- наборы данных для машинной обработки
- лингвистические базы данных
- тренажеры и другие CALL
- GIS (геоинформация)



Информация к ближайшим семинарам

- ELAN
- данные для вашего индивидуального проекта
- обучающие материалы

Корпус: основные принципы составления

- Представительность (репрезентативность)
- Сбалансированность (J. Biber 1993; J. Leech 2007;)
- Ориентированность на пользователя
- Учет условий для разработки

    - ручной отбор данных VS автоматический отбор данных
    - ручная разметка VS автоматическая разметка

Корпус: основные принципы составления

- Представительность (репрезентативность) (J. Biber 1993; J. Leech 2007;
- Сбалансированность дискуссия в Adel 2020)
- Размер
- Каковы критерии отбора? >>

- Ориентированность на пользователя
- Учет условий для разработки

  - ручной отбор данных VS автоматический отбор данных
  - ручная разметка VS автоматическая разметка

Domain	%	Date	%	Medium	%
Imaginative	21.91	1960-74	2.26	Book	58.58
Arts	8.08	1975-93	89.23	Periodical	31.08
Belief and thought	3.40	Unclassified	8.49	Misc. published	4.38
Commerce/Finance	7.93			Misc. unpublished	4.00
Leisure	11.13			To-be-spoken	1.52
Natural/pure science	4.18			Unclassified	0.40
Applied science	8.21				
Social science	14.80				
World affairs	18.39				
Unclassified	1.93				

Баланс текстов разного типа в BNC (письменные тексты - 90% корпуса)

Корпус: слои разметки

- Основные уровни языка
  - лексическая информация
  - морфологическая информация
  - синтаксическая информация
  - семантическая информация
  - фонетика
  - интонация
  - жесты
- Переводы, глоссы
- Специальная разметка (стиховедческая, в параллельных корпусах и т.д.)
- Метаданные
- Надкорпусные базы данных

Иллюстрации

- НКРЯ
  - Инструкция
  - Состав корпуса: статистика
  - Задать подкорпус: снятая и неснятая омонимия в разборах
  - Разборы слова (по клику в выдаче)
  - Распределение по годам (*вертывать vs *ворачивать; без сахару vs без сахара)
  - N-граммы (не без …)
  - Дополнительные признаки
- Стиховедческая разметка в Поэтическом корпусе
- Жестовая и орфоэпическая разметка в Мультимедийном корпусе
- Ударение в Акцентологическом корпусе

Надкорпусные базы данных

- Для разметки корпуса
  - например, база устойчивых оборотов НКРЯ и мн. др.
- Сводные данные на основе корпуса
  - CoCoCo - сочетаемость слов (на основе НКРЯ, Taiga, i-RU)
  - Google Ngrams - частотная информация о словах и сочетаниях
  - rusVectores - близость по смыслу на основе сочетаемости
- Сводные данные + исследование
  - база данных союзов и частиц (коннекторов) - классификация примеров употребления, научный комментарий
  - база данных межъязыковых эквивалентов

Литература

- Ädel, Annelie. Corpus Compilation. In: A Practical Handbook of Corpus Linguistics (ed. by M. Paquot, S. Gries). New York: Springer, 2020. Pp. 3-24.
- Biber, Douglas. Representativeness in Corpus Design. Literary and Linguistic Computing, Volume 8, Issue 4, 1993. Pp. 243–257.
- Leech, Jeoffrey. New resources, or just better old ones? The Holy Grail of representativeness. In: Corpus Linguistics and the Web (ed. by M. Hundt,N. Nesselhauf, C. Biewer). New York: Rodopi, 2007. Pp. 133-149.
- McEnery, T., Wilson, A. Corpus Linguistics: An Introduction. Edinburgh, Edinburgh University Press, 1996.
- Копотев, М. В. Введение в корпусную лингвистику: учебное пособие для студентов филологических и лингвистических специальностей университетов. Praha : Animedia Company, 2014.
- Плунгян В. А. Корпсная лингвистика. ПостНаука, 2013.