Контролируемый корпус: рассказы по картинкам

Г. Мороз

последняя версия: **https://goo.gl/qwmmXZ**

1. Введение

Контролируемый корпус — это корпус, собранный в результате обработки диалогов двух информантов, в ходе которого один информант рассказывает историю, основываясь на стимулах, предложенных исследователем. Традиционно в лингвистике используются визуальные стимулы (изображения или видео), но, естественно, аналогичным образом можно исследовать кодирование в языке вкусовых, тактильных или звуковых ощущений носителей или даже моделировать языковые поведение в тех или иных ситуациях. Метод контролируемого корпуса позволяет

- избавится от основных проблем элицитации, так как влияние исследователя на порождаемое информантом значительно снижается;
- значительно повысить вероятность появления исследуемых явлений, что было бы невозможно при сборе репрезентативного естественного корпуса языка.

Метод контролируемого корпуса не стоит воспринимать как аналог элицитации или корпусной лингвистике. При элицитации исследователь достаточно часто обращается скорее к знаниям носителя о языке, задавая вопросы "Как сказать *X*?", "Можно ли сказать *X*?", в то время как корпусные методы позволяют фиксировать случаи использования языка. Данные сущности часто достаточно сильно расходятся, что было показано в работе [Labov 1964: 300]: "*In the conscious report of their own usage, however, New York respondents are very inaccurate. <...> We shall see that when average New Yorkers report their own usage, they are basically giving us their norms of correctness.*". На знания о языке достаточно часто оказывает влияние те единицы описания языка, которые используются при обучении в школе. В результате исследователь и информант оказываются в заложниках таких понятий, как *предложение*, *слово*, *слог* и т. п. единиц письменного языка, которые редко в полной мере подходят для описания единиц звучащей речи — language in use (см., например, [Miller et al. 1998]). К тому же письменный и устный варианты языка могут значительно расходится друг с другом (см. [Lyons 1968: 40-42], [Tannen 1982] и др.), а во время элицитации часто именно письменный язык является "образцом" для информанта. Некоторые лингвисты (Блумфильд, Лайонс) однозначно высказывались о производности письменного варианта от устного, считая устный вариант в каком-то смысле важнее письменного. Современные исследования показывают, что разные варианты языка находятся в очень сложных отношениях и активно влияют друг на друга. Надо полагать, в случае адыгских идиомов разрыв между письменным и устным вариантами не столь велик, как, скажем в случае русского или китайского языков, однако в большинстве аулов, в которых нам приходилось работать, язык СМИ и школы (имеется в виду литературные адыгейский или кабардино-черкесский языки) часто отличаются от родного идиома информантов.

Идея контролируемого корпуса не нова и восходит к работе [Chafe 1980], однако контролируемые корпуса чаще всего используют в исследованиях дискурса. В данной работе эта методика будет использована для исследования фонологии и морфосинтаксиса.

1.1 Сбор данных

В каждой сессии информанты участвуют парами. Сначала каждому информанту выдается по одному набору картинок, рассказывающие две истории. Потом через какое-то время на одного из информантов надевается микрофон и он рассказывает свою историю. Второй информант его слушает, и если что-то не ясно, то спрашивает. Потом информанты меняются ролями и все повторяется. Чтобы избежать соблазна пересматривать картинки (и тем самым шуметь на аудиозаписи), картинки у рассказывающего изымаются. Как видно из описания, для данной работы необходимо подготовить минимум две истории: одну информант будет рассказывать, другую — слушать.

После записи аудиозаписи аннотируются в программе ELAN, где сказанное разбирается вместе с носителями языка: происходит сегментация текста на дискурсивные единицы, создается пословный перевод, а также перевод на русский язык. Позже полученные тексты глоссируются, паузы размечаются в рамках системы нотации, предложенной в [Кибрик et al. 2014]. Последним этапом работы с корпусом является создание необходимой аннотации для каждого исследуемого признака и извлечение целевой информации.

1.2 Исследуемые явления

- фонологические
  - скорость речи
  - длительность сегментов и их составляющих
    - длительность гласных
    - VOT стопов и аффрикат
    - длительность фрикативных
  - форманты гласных
  - спектральная характеристика фрикативных
  - чередование ɐ ∼ a
  - ? исчезновение j в интервокале
  - ...
- грамматические
  - использование релятивизации
- лексические

1.3 Диалектные особенности

При создании корпуса планируется использовать материал не только одного идиома, но хотелось бы потенциально иметь возможность собрать материал разных адыгских идиомов. В связи с этим при создании картинок, следует учесть междиалектные различия, которые позволили сделать из рассказа своего рода тест, позволяющей по набору тех или иных диалектных черт относить идиом к той или иной диалектной группе. При выборе признаков мы опирались на диалектная классификацию, представленную на рисунке 1.

Таким образом, для составления изображений, которые потом будут использованы в исследовании, необходимо составить некоторый список фонетических, морфологических и лексических особенностей адыгских идиомов.

2. Исследование скорость речи

2.1 Введение

Судя по ссылкам, о скорости речи говорили еще в начале XX века, но первые квантитативные исследования, видимо начались с работ [Goldman-Eisler 1954] и [Goldman-Eisler 1956].

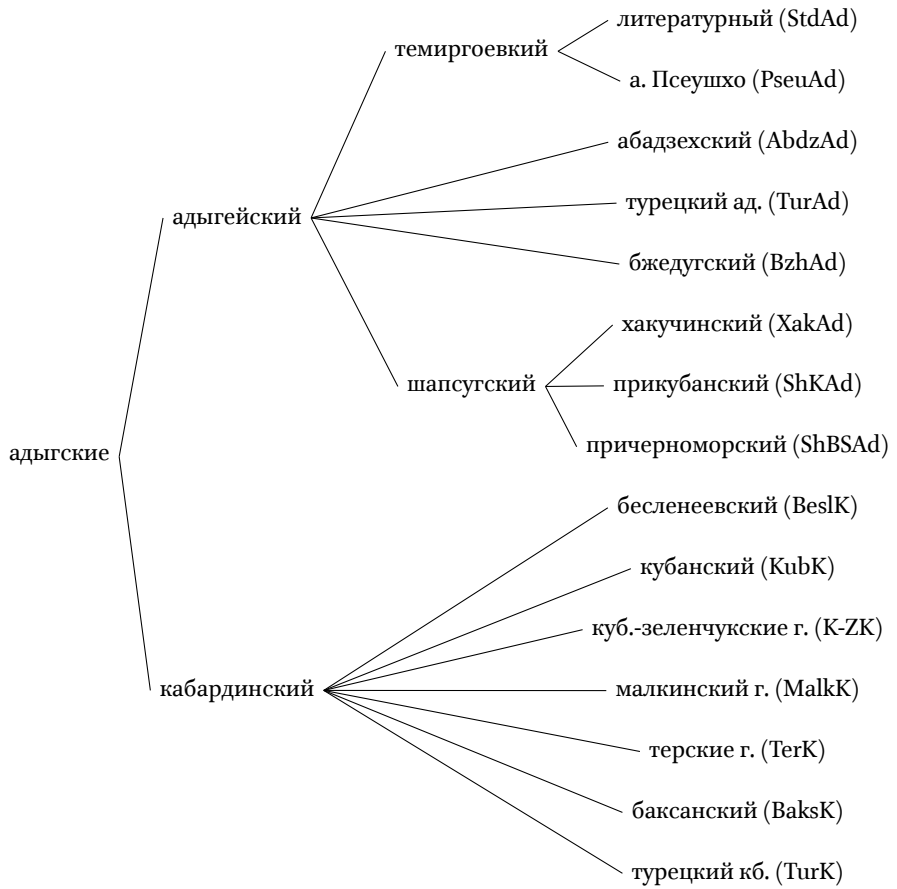

Рис. 1: Классификация адыгских идиомов

3. Исследование лексики

Chafe, W. (ed.) (1980). *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Ablex.

Goldman-Eisler, F. (1954). A study of individual differences and of interaction in the behaviour of some aspects of language in interviews. *The British Journal of Psychiatry 100*(418), 177–197.

Goldman-Eisler, F. (1956). The determinants of the rate of speech output and their mutual relations. *Journal of Psychosomatic Research 1*(2), 137–143.

Labov, W. (2006 (1964)). *The social stratification of English in New York City.* Ph. D. thesis, Columbia university.

Lyons, J. (1968). *Introduction to theoretical linguistics*. Cambridge university press.

Miller, J. E., J. Miller„ R. Weinert (1998). *Spontaneous spoken language: Syntax and discourse*. Oxford University Press, USA.

Tannen, D. (1982). Oral and literate strategies in spoken and written narratives. *Language*, 1–21.

Кибрик, А., В. Подлесская„ др. (2014). *Рассказы о сновидениях: Корпусное исследование устного русского дискурса*. Litres.