

«Рассказы и разговоры о грушах»: референциальная аннотация (декабрь 2018)

1. Референциальные единицы, размечаемые в рамках корпуса.

Основу референциальной аннотации составляет разметка всех языковых выражений с конкретной референцией. Среди них выделяются два основных типа единиц – дейктические и анафорические.

К дейктическим единицам, размечаемым в рамках нашего корпуса, относятся языковые выражения, совпадающие с непосредственными участниками речевого акта. В речи они обозначаются личными местоимениями 1-го и 2-го лица (подчеркнуты в примерах 1-3):

- (1) (/Можно я расскажу^h?
- (2) А потом ты ко= || /откомментируешь, ==)
- (3) (а) мы видим сельский /–пейзаж,

В ряде случаев они могут не выражаться эксплицитно –

- (4) \Вообще Ø не помню этого!

однако в разметку подобные нулевые выражения также включаются.

К анафорическим единицам относятся языковые выражения, однозначная интерпретация которых невозможна без отсылки к некоторому предыдущему контексту.

Основной массив анафорических единиц составляют:

- личные местоимения 3-го лица:

- (5) (h) (q) (a) он ХХ собирает /–груши,
- (6) Я ведь на него не /смотрела как-то,
- (7) у них происходит такой \обмен;

- указательные местоимения:

- (8) Ну потому что /тот немножко был в \шоке ;

- Определительные местоимения:

- (9) /волосы у всех достаточно длинные,

- Неопределенные местоимения:

- (10) ну-у нь= || кто-то /выше,

кто-то \ниже^w.

- нули:

(11) а потом Ø_{pro} спускается по /лестнице,

(12) /Что Ø_{pro} потерял?

В рамках разметки эти единицы объединяются в более общий тип **редуцированной референции**. Ему противопоставлена **полная референция** – именные группам с вершиной-существительным или числительным (подчеркнуты в примерах 13-14), которые выступают в роли антецедентов для анафорических единиц:

(13) (a) (m) велосипед мальчику /велик
он едет не в /седле

(14) /корзин там было три,
две /полные,
/одна \пустая.

Данные элементы также размечаются в рамках анафорических единиц, с пометой Full.

Как и другие каналы, референциальная аннотация в рамках проекта осуществляется в программе **ELAN**. В рамках этой разметки анафорические и дейктические единицы составляют два основных независимых слоя (refAnafora / refDeixis), зависящих от слоя Words¹ (стереотип Included In) и при этом, в свою очередь, включающих в себя зависимые слои с частными конкретными параметрами референциального выбора (см. раздел 2). Родительским элементом по отношению к эксплицитно невыраженному местоимению 3-го лица в программе ELAN выступает непосредственно следующее эксплицитно выраженное слово в слое *Words*.

В ряде случаев референциальные единицы являются «вложенными» друг в друга: несколько синтаксически связанных (как с помощью сочинительной, так и с помощью подчинительной связи) анафорических единиц образуют составную единицу с конкретной референцией:

(15) {sw} /по-моему это было [[/два мальчика]_{ref1} и [девочка]_{ref2}]_{ref3}

(16) /Потом в кадре появляется [[(h) (a) \мальчик [на велосипеде]_{ref1}]]_{ref2}

(17) #неа# по-моему [один [из трёх /мальчиков]_{ref1}]]_{ref2} был всё-таки девочкой,

¹ О промежуточном слое *refAnSpread* для анафорических единиц, расположенном в иерархии между слоями *Words* и *refAnaphora*, см. ниже.

В связи с тем, что в программе ELAN невозможно создать слой с аннотациями разного уровня, все референциальные единицы, не содержащие вложенных составляющих (=дискретные аннотации слоя *refAnafora*), подчинены слою *refAnSpread* (стереотип *Included In*). Каждая из аннотаций слоя *refAnSpread* по умолчанию представляет собой группу из одной или более синтаксически связанных единиц *refAnafora*. На данный момент, по данным записей 4, 22 и 23 корпуса «Рассказы и разговоры о грушах», максимальное число вложенных составляющих составляло 3, в связи с чем слоя *refAnSpread* по умолчанию является родительским для 3-х подчиненных слоев *refAnafora1*, *refAnafora2* и *refAnafora3*. В случае отсутствия вложенных референциальных составляющих дальнейшая разметка производится только в слое *refAnafora1*, а слои *refAnafora2* и *refAnafora3* остаются пустыми.

2. Релевантные параметры референциального выбора для дейктических и анафорических единиц.

Выбор конкретного языкового выражения для актуализованного в дискурсе элемента (референциальный выбор) зависит от многих параметров², точный список которых на данный момент остается открытым. В общем и целом референциальный выбор для анафорических единиц заключается в выборе между полной именной группой с вершиной-существительным/числительным (полная референция), субъектном местоимении 3-го лица, указательном местоимении и нуле (последние три типа относятся к более общему типу редуцированной субъектной референции). Для дейктических единиц в русском языке выбор осуществляется между эксплицитно выраженными (от 2/3 до 3/4 высказываний³) и невыраженными единицами (от 1/4 до 1/3 высказываний).

В рамках данного проекта в качестве релевантных параметров, задействованных в референциальном выборе, размечаются следующие категории, каждая из которых представляет собой отдельный слой, подчиненный слоям *refDeixis* или же *refAnaphora1-3*, в зависимости от типа референциальной единицы (дейксис VS анафора):

1) Для дейктических единиц (*refDeixis*):

- а. Употребление в речи (слой *refDeiExpr*): единица эксплицитно выражена / не выражена (*Overt / Zero*). Родительским элементом по отношению к

² См., например, Кибрик А. А., Добров Г. Б., Залманов Д. А., Линник А. С., Лукашевич Н. В. Референциальный выбор как многофакторный вероятностный процесс. // Кибрик А. Е. и др. (ред.). Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог». М.: РГГУ, 2010.

³ Относительно количественных данных см., в частности, Grenoble L. Conceptual reference points, pronouns, and conversational structure in Russian. *Glossos 1*, 2001. Онлайн на https://slaviccenters.duke.edu/sites/slaviccenters.duke.edu/files/media_items_files/grenoble.original.pdf
<http://multidiscourse.ru>

эксплицитно невыраженному местоимению 1-го и 2-го лица в программе ELAN выступает непосредственно следующее эксплицитно выраженное слово в слое *Words*.

- b. Лицо референта (слой *refDeiPerson*): 1-е / 2-е (1 / 2)
- c. Число референта (слой *refDeiNumber*): единственное / множественное (SG / PL)
- d. Синтаксическое выражение (слой *refDeiSynt*): подлежащее / прямое дополнение / косвенное дополнение / прочее (Subj / DirObj / IndirObj / Other)

2) Для анафорических единиц (*refAnaphora1-3*):

- a. Тип референции (слой *refAnType*): полная / редуцированная (Full / Reduced)
- b. Конкретный тип референциальной единицы (слои *refAnFull* and *refAnReduced*, зависят от слоя *refAnType* и имеет стереотип Symbolic Association): именная группа с вершиной существительным/ числительным (NomP/ NumP) в случае полной референции VS личное местоимение 3-го лица / указательное местоимение / определительное местоимение / неопределенное местоимение / нуль в случае редуцированной референции. (Pron3 / Dem / Def/ Indef/ Zero). Кроме того, при наличии предлога выбирается соответствующая единица с пометой Prep.
- c. Род референта (слой *refAnGender*): мужской / женский / средний / смешанный/ прочее (M / F / N/ Mixed/Other). Значение Mixed выбирается в случае составной единицы, относящейся к нескольким референтам разного рода (как в примере 13). Значение Other выбирается в случае словарной невозможности определения рода для парных элементов типа *штаны*, *волосы* и т. д.
- d. Число референта (слой *refAnNumber*): единственное / множественное (SG / PL)
- e. Синтаксическое выражение (слой *refAnSynt*): подлежащее / прямое дополнение / косвенное дополнение / прочее (Subj / DirObj / IndirObj / Other).

Кроме того, к каждому слою *refAnaphora1-3* и *refDeixis* привязан слой с комментариями, где может указываться тип связи для составных единиц (Coordination / Subordination), наличие предшествующего определительного (With Def) или неопределенного (With_UNDEF) местоимения, зависимого относительного придаточного

(With REL), а также случаи эмфатического (Emphasis) или контрастивного употребления (Contrast).

Пример референциальной аннотации можно видеть ниже на рис. 1, где представлен скриншот разметки для примера (9):

The screenshot displays the ELAN software interface for referential annotation. At the top, a timeline shows time intervals from 00:16:21.500 to 00:16:23.500. The main workspace shows the text 'волосы у всех достаточно длинные' with a blue bar for 'C-205' and a red bar for 'Spr 150'. The left sidebar lists various annotation types, including 'EDU', 'Words', 'refAnSpread', 'refAnaphora1', 'refAnType', 'refAnFull', 'refAnRedu', 'refAnGender', 'refAnNumbe', 'refAnSynt', 'refAnComm', 'refAnaphora-2', 'refAnType-2', 'refAnFull-2', 'refAnRedu', 'refAnGender', 'refAnNumbe', 'refAnSynt-2', 'refAnComm', and 'refAnaphora-3'. The main workspace shows a large blue bar for 'C-205' and a red bar for 'Spr 150'. The 'refAnType' bar is labeled 'Full' and 'NounP'. The 'refAnType-2' bar is labeled 'Reduced'. The 'refAnFull-2' bar is labeled 'PrepDef'. The 'refAnRedu' bar is labeled 'Mixed'. The 'refAnGender' bar is labeled 'PL'. The 'refAnNumbe' bar is labeled 'PL'. The 'refAnSynt-2' bar is labeled 'Other'.

Рис. 1. Вид референциальной аннотации в программе ELAN на примере (9).