"Russian Pear Chats and Stories": referential annotation guide

(Version 14.12.2018)

Evgeniya V. Budennaya[1]

1. Referential expressions marked in the corpus

The referential annotation is built on all verbal expressions with a definite reference. Among them, two basic types are annotated: deictic and anaphoric.

Deictic expressions, annotated within the frames of our corpus, include verbal expressions which coincide with speech acts. In speech, they are marked by first and second person pronouns (underlined in Examples 1-3 below):

(1) (/Is it ok if I talk first^{h}?

(2) Then you'll co= || /mment, ==)

(3) (and) we can see a rural landscape/scene".

In some cases, they may not be expressed explicitly:

(4) \ Ø don't remember this at all!

However, such zero expressions are also included in the annotation.

Anaphoric expressions include linguistic expressions which cannot be interpreted unambiguously unless there is a reference to some previous context. In the majority of cases they can be expressed by:

- third person pronouns:

(5) (h) (q) (a) he XX [is] picking /-pears,

(6) I didn't really look at him

(7) they are having some [sort of] of exchange;

- demonstrative pronouns:

(8) Well because that that [one} was a bit \dumbfounded;

- definite pronouns:

(9) /all of them have quite long hair,

- indefinite pronouns:

[1] To cite this version:
Budennaya, E.V. "Russian Pear Chats and Stories": Referential annotation guide. Version 14.12.2018. http://multidiscourse.ru

(10) well, w=|| some [are] taller,

(11) some [are] \ shorter^{w}.

- and zero pronouns:

(12) and then $\underline{Ø_{pro}}$ is going down the stairs,

(13) /What did $\underline{Ø_{pro}}$ lose)?

In our annotation scheme, all these expressions are expressed under a wider category of **reduced reference**. This category is opposed to **full reference**, i.e. constituents with a noun or a numeral head (underlined in Examples 13-14 below) which serve as antecedents for future anaphoric expressions:

(14) (m) the bike is too big for the boy
he is not going [to be] sitting in the saddle

(15) there were three baskets,
two were full,
one was empty.

These elements are also annotated within anaphoric expressions, with a 'Full'-tag.

As with other channels, referential annotation is carried out with the help of **ELAN** software. Within the annotation, anaphoric and deictic expressions form two main independent tiers (refAnaphora/refDeixis) which both depend on the Words2 tier (stereotype Included In). Each of them is a parent for several tiers where particular parameters of the referential choice are annotated (see Section 2). For the third person unexpressed (zero) pronoun, the following explicitly expressed word in the *Words* tier is annotated as a parent item.

Sometimes referential expressions form 'embedded' constituents: several syntactically related (either with coordination or subordination) anaphoric expressions form a compound expression with a definite reference:

(16) {sw} /I think there are **[[/two boys]$_{ref1}$ and [a girl]$_{ref2}$]$_{ref3}$**

(17) / Then in the shot **[[(h) (a) \boy [on a bike]$_{ref1}$]$_{ref2}$** appears

(18) #nah# I think **[one [of the three /boys]$_{ref1}$]$_{ref2}$ was a girl after all,**

Since ELAN does not allow creating a tier with different levels of annotation, all referential expressions that do not contain embedded constituents depend on the *refAnSpread* tier (stereotype Included In). Each of the annotations of the *refAnSpread* tier represents a group of one or several syntactically related expressions *refAnaphora* by default. At present, based on the recordings #4, #22, and #23 of the "Russian Pear Chats and Stories" corpus, the maximum number of embedded constituents was equal to three items. For this reason the *refAnSpread* tier is a parent to the three child tiers *refAnaphora1 refAnaphora2 and refAnaphora3*. When no embedded referential

2 On the intermediate tier *refAnSpread* for anaphoric expressions, which is located between *Words* and *refAnaphoraN* tiers in the hierarchy, see below.

constituents are found, only the first *refAnaphora1* tier is annotated, and tiers refAnafora2 and refAnaphora3 remain empty.

2. Relevant parameters of the referential choice for deictic and anaphoric expressions

The choice of a particular linguistic expression for an element actualized in discourse (referential choice) depends on numerous parameters, the list of which is still open. Generally, referential choice for anaphoric expressions refers to the choice between a full NP with a noun/numeral head (full reference), a third person pronoun, a demonstrative pronoun, and a zero pronoun (the last three types belong to a more general type of reduced subject reference). For deictic expressions in Russian, the choice is made between explicitly expressed (from 2/3 to 3/4 of occurrences) and unexpressed expressions (from 1/4 to 1/3 of occurrences).

In the given project, the following categories involved in the referential choice are chosen as relevant parameters. Each of these categories is annotated within a separate tier that depends either on the *refDeixis* or on the *refAnaphoraN* tier, according to the type of a particular referential expression (deixis VS anaphora):

1) For deictic expressions:

    a. Explicitness (*refDeiExpr* tier): explicitly expressed / not expressed (Overt/Zero). For a first or second zero pronoun, the following explicitly expressed word in the *Words* tier serves as a parent element.
    b. Person (*refDeiPerson* tier): first / second person (1 / 2)
    c. Number (*refDeiNumber* tier): singular / plural (SG / PL)
    d. Grammatical role (*refDeiSynt* tier): subject / direct object / indirect object / other (Subj / DirObj / IndirObj / Other)

2) For anaphoric expressions (*refAnaphora1-3*):

    a. Type of reference (*refAnType* tier): Full/ Reduced
    b. Type of referential expression (*refAnExpression* tier; depends on the *refAnType* tier and belongs to *Symbolic association* stereotype): NP with a noun / numeral head (NomP / NumP) for full reference VS third person pronoun/ demonstrative pronoun / definitive pronoun / indefinite pronoun / zero for reduced reference (Pron3 / Dem / Def / Indef / Zero). In the case of a preceding preposition, a corresponding expression with Prep-tag is to be chosen.
    c. Gender (*refAnGender* tier): male / female / neutral / mixed / other (M / F / N / Mixed / Other). 'Mixed' is used for a compound expression which refers to several different entities (as shown in Example 13). 'Other' is used when it is impossible to define the gender of paired elements, such as *'trousers', 'scissors'*, etc.
    d. Number of the referent (*refAnNumber* tier): singular / plural (SG / PL)
    e. Syntactic expression (*refAnSynt* tier): subject / direct object / indirect object / other (Subj / DirObj / IndirObj / Other).

Apart from this, each *refAnaphoraN* and *refDeixis* tier has an dependent tier with comments where the type of relation for compound elements (Coordination / Subordination), a preceding definite (With Def) or indefinite pronoun (With Undef), or relative clause (With REL) can be marked, along with emphasis (Emphasis) or contrastiveness (Contrast).

An example of referential annotation is presented below in Figure 1 (an ELAN annotation screenshot for Example 19):

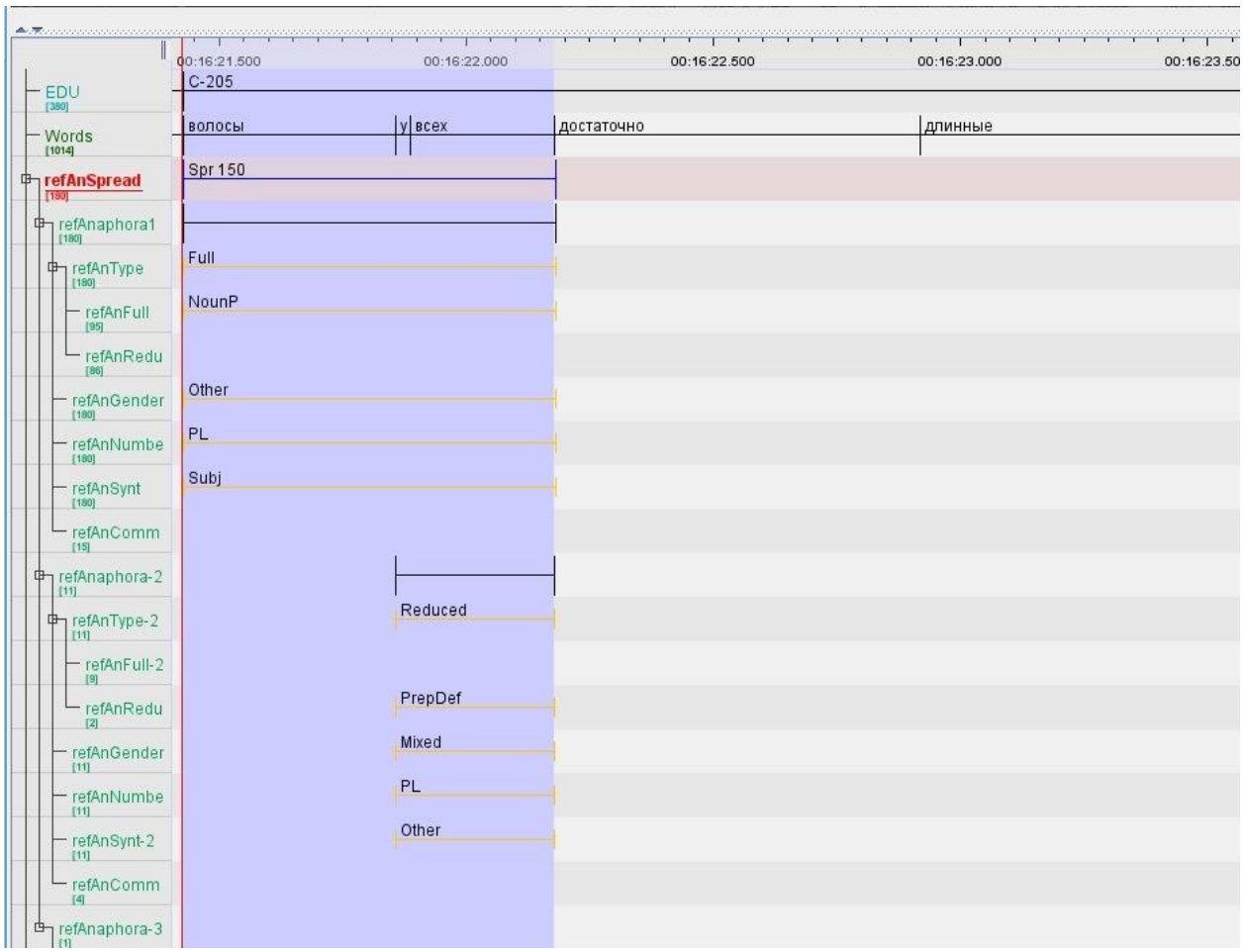

Fig.1 An example of referential annotation in ELAN (19).