UDC 81.33 DOI: 10.33910/2687-0215-2019-1-2-337-364

The project of a deeply tagged parallel corpus of Middle Russian translations from Latin

E. G. Sokolov✉[1]

[1] Institute for Linguistic Studies, Russian Academy of Sciences, 9 Tuchkov Ln., Saint Petersburg 199053, Russia

Abstract. Tagged parallel corpora are powerful tools for the analysis of natural language. Moreover, for historical linguistics, whose most peculiar shortcoming is lack of living native speakers, corpora — as paper or electronic collections of written texts — are the main source of linguistic information. Old and Middle Russian are well-documented languages, and a host of manuscripts in both idioms — including those containing numerous translations — are available for investigation. Nevertheless, up to now there is no parallel translational corpus of Middle Russian. Thus, a number of important written sources containing information valuable for linguists, literary scholars and historians cannot be studied properly. This article provides a preliminary account of the project of a deeply tagged parallel corpus of Middle Russian translations from Latin. Such corpus may prove useful in the formal description of the translation techniques of the time, which may help with dividing the anonymous texts of the time into several groups based on their language features. Such grouping may help with authorship attribution and, consequently, with incorporating each translation into a proper cultural landscape.

From the linguistic point of view, such corpus could provide researchers with crucial information on the vocabulary, morphology and syntax of Middle Russian with an emphasis on the argument structure of the verbs, usage of borrowed lexical items and set expressions and professional skills of the ancient translators. The article gives an outline of the crucial features of the prospective Middle Russian translational corpus, its possible primary contents, text standardization and annotation principles, as well as the reasons for not using a theory-neutral syntactic apparatus, characteristic of the existing historical corpora of ancient Indo-European languages, such as TOROT or PROIEL. An explanation of how the potential users of this corpus could benefit from our non-standard tagging principles is given.

Keywords: Middle Russian, Church Slavonic, Latin, translation, electronic corpora, syntactic alignment.

Introduction

Short description of the project

A parallel corpus includes a number of texts in one language with their translation into another; the corresponding fragments of the original and the translation are aligned. This usually provides tools that assist in finding not only lexical information, but other linguistic data, e. g., morphological or syntactic. Although there exist a number of Old and Middle Russian corpora (Mitrenina 2014), no parallel corpus of Old or Middle Russian translations was ever developed.

This article presents the project of deeply tagged parallel corpus of Middle Russian translations from Latin, i. e. Latin — Russian[1] translational corpus with parallel syntactic alignment (hereinafter LRC), containing the Russian translations from Latin made between the end of the 15th century and the beginning of the 17th century — that is, roughly, within period of a hundred years. It will provide a wide range of researchers with a modern instrument for an in-depth analysis of a significant layer of premodern Russian culture, namely the pretheoretic translational activity in the pre-Petrine Russia. The tasks of the project include:

[1] The notion "(Middle) Russian" here and further refers rather to the origin of the translations than to their language, comprising both Middle Russian and Church Slavonic text translated from Latin.

(1) formulating the general principles for syntactic, semantic, morphological and lexical annotation of the texts respectively in Latin and Russian parts of LRC;
(2) formulating the alignment principles for text pairs in LRC, especially the principles of syntactic alignment within such pairs;
(4) formulating the main principles of a bilingual glossary built using the LRC alignment;
(5) developing and adjusting the electronic tools for implementation of the objectives (1–4);
(7) formulating the guidelines for further annotation of LRC and other similar corpora in future.

The fulfillment of these tasks will let us launch the first deep annotated parallel historical corpus of Russian language, which will substantially enhance the research abilities of the scholars concerned with history of Slavic languages and cultures.

Some general reasons for creating LRC

Translations (mostly from Greek and Latin) constitute a substantial part of the written sources of Old and Middle Russian origin, and are undoubtedly of considerable importance for both linguists and literary scholars. These translations can provide researchers with important data concerning historical grammar and history of Russian language, including such topics as lexical, idiomatical and syntactic borrowings and their impact on written language, as well as translation principles, their variation and gradual changes. Russian pre-Petrine translations could be also of some interest for the researchers concerned with contrastive grammar description or language typology, because the process of aligning bilingual corpora provides valuable information about both the source and target languages and their grammatical properties (Grishman 1999, 225). They are also a great source of cultural and historical information, and thus must be the matter of interest for historians and literary scholars, who may also require some additional linguistic information in order not to mistake in ascribing the text to a wrong author.

For historical linguists studying the premodern languages whose only sources are sets of written texts, such sets — that is in fact, corpora (though not obligatorily in electronic format) — are the only reliable sources of linguistic data. To make such data verifiable, reliable and as exhaustive as possible (which is necessary for a serious study of any subject) they have to be easily collectable and extractable. This can only be done by means of an electronic corpus. So there exists an urgent need for a translational corpus of Old and/or Middle Russian.

To make the extraction of the information mentioned above possible, the corpus must be able to give a precise account of the structural relationships between the translation and its original. This means that the translational corpus must contain not only translations themselves, but also their originals paired to them, and that such pairs (which we will hereinafter call *parallel texts* or *bitexts*) have to bear the metalinguistic information mapping the units of the original into the units of its translation, i. e., to be aligned. The simplest way of aligning parallel texts is word alignment. But the alignment at the word level poses serious difficulties due to the complexity of the correspondences across the languages (Grishman 1999, 226). There is in fact next to no word-to-word correspondence even in a very literal translation. That is why it seems much more reasonable to compare the syntactic units of the parallel texts, not their lexical units (Grishman 1999, 226), proceeding top-down, that is, from the higher grammar units to the lower ones.

There is another reason for providing the translational corpus with parallel syntactic alignment, which is especially important for a corpus consisting of texts preserved in manuscript tradition. In the process of multiple copying a handwritten text undergoes multiple changes at nearly all levels, and syntax is the only feature of the handwritten text which remains relatively stable for a long period of time (Tomelleri 2011, 219), so it is desirable that the parallel texts in a historical corpus have a syntactic alignment.

Lack of translational corpora for pre-Petrine Russian texts

As of now there exist a substantial number of electronic resources containing historical corpora of Slavic languages, for example, the historical and Church Slavonic subcorpora of the Russian National Corpus, the MANUSCRIPT project, the SKAT project, the OCS subcorpus of the PROIEL project, the TOROT project, and some others. As a rule, such corpora are provided with lexical and morphological annotation, but most of them lack any tools for syntactic analysis of the texts. The only exceptions are PROIEL, TOROT and SKAT. The first two projects were initially designed to be syntactic treebanks (Haug et al. 2009; Eckhoff, Berdičevskis 2016, 63), i. e. sets of texts with syntactic trees associated with their units, while the syntactic module of the last project is still being developed (Alekseeva 2014).

So, in fact there are only two corpora containing the syntactic representation of OCS, Old and Middle Russian data, and both of them cannot meet the requirements imposed by our goal. The PROIEL corpus contains only three OCS codices (*Marianus, Suprasliensis* and *Zographensis*) and three Old Russian texts (*Codex Laurentianus*, *The Taking of Pskov* and *The Tale of Luka Kolocskij*) and thus is of nearly no interest to us; the TOROT project offers a lot of Old and Middle Russian sources, but doesn't aim at translation studies, containing mostly original texts.

That is why we believe that there is an urgent need for a linguistic project like LRC, focusing on the syntactic and semantic representation of aligned bitexts.

Prospective features of the LRC project

As we have already mentioned, the LRC project must be designed to meet the requirements of historical translation studies. First of all, to study a translation is to describe and explain its technique, i. e. reveal the transfer rules underlying the process of rendering a text in the source language into a text in the target language. From this point of view alignment is pairing of a subset of the nodes in the source syntactic tree with a subset of the nodes in the target syntactic tree (Grishman 1999, 228). But each set of nodes is, in fact, a particular manifestation of an abstract grammar construction; for instance, the pair *stante illa domo ~ стоѧщу домꙋ* (Fedorova 1999b, 98) is a manifestation of the *ablativus absolutus* in its Latin part and of the *dativus absolutus* in its Church Slavonic part, thus a manifestation of the *ablativus absolutus ~ dativus absolutus* correspondence between the source and goal languages. Each text, Latin or Russian, must be viewed as a set of units corresponding to a certain set of grammatical constructions. Hence, the syntactic alignment of the LRC bitexts must let the researcher establish the correspondence between the set of grammatical constructions in a source language and the set of grammatical constructions in the respective target language.

Any text can be represented by a finite sequence, or *string*, of word forms, which can be divided into several *substrings* (Partee, ter Meulen, Wall 1990, 433–435).

Furthermore, any text can be represented by a finite ordered set of separate *syntactic trees*. Syntactic tree is a tree in graph theoretic sense, over which a certain *relation* is defined. If a tree depicts the *part — whole relation* between the substrings of a certain string, it is called a *phrase structure tree* or a *phrase marker* (hereinafter PST); if a tree represents the *dependency relation* which holds between its single nodes (Gaifman 1965, 306), it is called a *dependency tree* (hereinafter DT). There is the third relation, *dominance*, defined as follows: a node α is said to *dominate* a node β, if a connected sequence of branches can be extended from α to β (Partee, ter Meulen, Wall 1990, 433–435). It is obvious that dominance holds both between the nodes of PST and DT.

Alignment can be defined as pairing of a subset of the nodes in the source syntactic tree with a subset of the nodes in the target syntactic tree (Grishman 1999, 228), see (fig. 1).

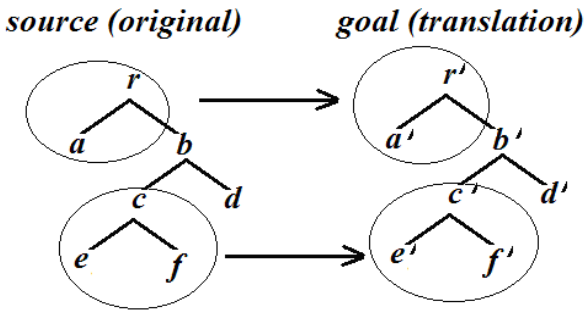

Fig. 1. Aligning the corresponding trees in the source and the goal texts

Syntactic construction. As any syntactic tree represents certain structural configuration of its non-terminal nodes, and this structural configuration may be preserved even if the word forms in the terminal nodes are replaced by other forms of the same distributional class, we may call such stable structural relation *a syntactic construction*. For instance, Latin utterances (a) and (b)[2] both contain subordinate clauses whose subject noun phrase bears accusative case and verb takes the infinitive form. They are examples of a specific construction called *accusativus cum infinitivo*, which has the following features: (1) it has the form of a clause; (2) it is the dependent of a verb or participle (here *tradidit* and *compertum*); (3) its subject bears accusative case; (4) its predicate is represented by an infinitive form of a verb.

(a) [*tradidit Herodotus*] ... *cynamomum in auium nidis reperiri* '
[Herodotus tells ...] that cinnamon can be found inside bird's nests'
(b) [*compertum*] ... *cynamomum longissime ab omni Aethiopia gigni*
'[is revealed] ... that cinnamon grows as far as possible from any land associated with Ethiopia'

Let us assume that each particular source text is generated by means of the specific set of syntactic constructions $C := \{c_1, c_2, c_3 \ldots c_n\}$, and that there is the set of grammar constructions $K := \{k_1, k_2, k_3 \ldots k_n\}$ which corresponds to it in the target language and enables the generation of the target text. Now it is possible to introduce the notion of *translation technique*.

Translation technique. Let A be a source text and A' be its translation. Then the translation technique *M* for the pair <A, A'> is the set of all the pairs of *syntactic constructions* or *syntactic subtrees* <δ, δ'> ordered by a binary relation R defined as follows: δRδ' := δ *is translated via* δ'.

$M := \{< \delta, \delta' > \mid (\delta \in A) \,\&\, (\delta' \in A') \,\&\, (R< \delta, \delta' > = 1)\}$

Hence the syntactic alignment of the LRC bitexts must let the researcher establish the correspondence between the set of grammatical constructions in a source language and the set of grammatical constructions in the respective target language in order to define the translation technique for a certain bitext, like that given below (fig. 2) for "The Letter on the Moluccas" written by *Maximilianus Transsylvanus*.

The translation technique can seriously differ for different bitexts, which means that for two different translation techniques T_1 and T_2 there must be two correspondingly different sets of relations between the source and target constructions, say, R_1 and R_2. The LRC has to be able

[2] Here and further Latin and Russian examples are taken from The Letter on the Moluccas originally written in Latin by Maximilianus Transylvanus and translated into Russian in the mid-1520s. We have been studying this translation and its Latin source since 2013 and thus will often give examples from this text in the following sections of our article.

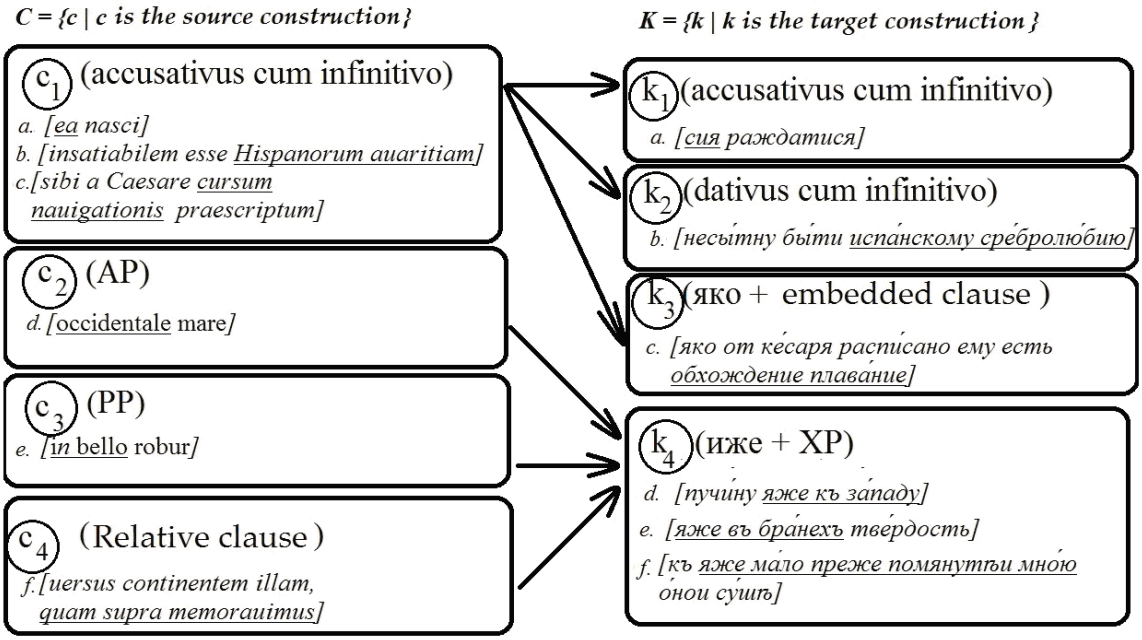

Fig. 2. An informally represented fragment of translation technique for "The Letter on the Moluccas"

to give the researcher a convenient opportunity for rendering the manually aligned pairs of syntactic nodes of a bitext into pairs of grammatical constructions (whereby the frequency of each construction must be taken into consideration as well), resulting in the set of transfer rules for each bitext. Having such sets of transfer rules, a scholar will be able to compare the translational techniques of different texts and to draw verifiable conclusions about the degree of propinquity of these texts.

For example, let us denote the bitext of "The Letter on the Moluccas" as A and imagine that there is a certain bitext B in which $R := \{(c_1, k_1), (c_1, k_3), (c_3, k_4)\}$, that is *accusativus cum infinitivo* is always translated either with *accusativus cum infinitivo* or with an embedded clause introduced by a complementizer *яко*, and the Church Slavonic non-finite construction with *иже, еже, яже* is the translation only for a prepositional phrase. In this case the machine will easily draw a Venn diagram (fig. 3) and let us see that $R(B) \subseteq R(A)$:

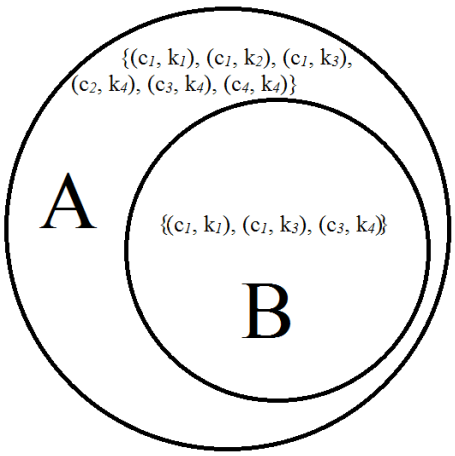

Fig. 3. An example of the Venn diagram

If all the bitexts in LRC have their own sets R, then for each n bitexts it will be possible to draw a Venn diagram consisting of n parts, each corresponding to a particular set R. If each pair in each set R is provided with its frequency rate, it will be possible to compare the frequency of a certain construction throughout all the bitexts available in the corpus. Moreover, it will enable the automatic assessment of propinquity degree for each n bitexts in corpus, preventing various scholars (especially non-linguists) from attribution faults caused by arbitrary and inconsistent reasoning.

Let us consider an example of such doubtful attribution. The first editors of the Russian version of "The Letter on the Moluccas", N. A. Kazakova and L. G. Katuškina, ascribe this translation to Dimitri Gerasimov (Kazakova, Katuškina 1968, 237–238). Their reasoning runs as follows. Dimitri was one of the most prominent translators of the time; besides, he was well-known for his diplomatic activities and took part in the famous ambassade to Rome, where he could have bought a Latin exemplar of the book mentioned above; finally, Mikhail Medovartsev, the scribe who prepared the only known copy of the text, was familiar to Dimitri. Basing entirely on these extralinguistic assumptions, the editors proceed from historical facts to linguistic conclusions about the structure of the text itself, noting that the translation preserves the literal manner typical for Gerasimov and follows the syntax of the Latin source (Kazakova, Katuškina 1968, 234). These conclusions were later repeated in Kazakova's monograph (Kazakova 1980) and D. O. Tsypkin's article (Tsypkin 1990). Nevertheless, in 1990 there appeared an article written by a German scholar Elke Wimmer (Wimmer 1990), which convincingly proved that Dimitri had no hand in this translation, giving strong linguistic reasoning by the expedient of comparison of the extracts from "The Letter on the Moluccas" with another translation surely made by Gerasimov, which demonstrated that the translation technique of The Letter differed considerably from what one could expect from a Gerasimov's translation.[3] The same conclusions were drawn in (Sokolov 2014), which is also based on some linguistic observations, despite the fact that the author was not familiar with Wimmer's article at that time. This example proves the necessity of linguistic studies for correct attribution of ancient translated texts, and the more precise such studies are, the more reliable will the attribution be. That is why we believe that the propinquity assessment mechanism similar to the one described in the initial part of this section must be implemented within the LRC project.

Primary set of texts

This section presents the primary set of texts which could become the basis of the project. Some of them have been already published, these are the translation of William Durandus' "Rationale divinorum officiorum" (Durandus 2012), the translation of Nicolaus de Lyra's "Probatio adventus Christi" (Fedorova 1999a), the so-called "Pravila gramatichnye" (Tomelleri 1999), some parts of the so-called Bruno's Psalter (Tomelleri 2004), Maxim Grek's translation of Picolomini's "De Captione urbis Constantinopolitanae" (Kloss 1975, 55–61; the text itself: Kloss 1975, 59–61), Guido de Columna's "Historia destructionis Troiae" (Tvorogov 1972), the translation of Pomponius Mela's "Chorographiae liber" (Matasova 2014). There are also several texts which need a re-edition, because their primary edition was weak from many points of view, e. g. the translation of Transsylvanus' "De Moluccis insulis... epistola" ("The Letter on the Moluccas") (Kazakova, Katuškina 1968). Some texts have never been published but are also of great importance for us, for instance, the so-called "Book of Saint Augustine" (Kalugin 2001). All these texts must be included into LRC as its basic component.

[3] See also (Wimmer 2005, 74).

Thematically related projects and software tools they use

Several projects might shed some light on how to implement the LRC project avoiding the difficulties which have already been overcome by other scholars.

An electronic database for Russian and Church Slavonic handwritten sources developed in the Vinogradov Institute of Russian Language allows for manual and semi-automatic markup, as well as automatic formation of lexical and grammatical indices, and is provided with a GUI which makes its application more user-friendly (Arkhangelskiy, Mishina, Pichkhadz 2014, 102). More importantly, its structure, based on the YAML files, allows for marking up strings of words and syntactic phrases (Arkhangelskiy, Mishina, Pichkhadz 2014, 102–103). One of the most striking features of the project is the ability to ascribe certain characteristics not to a single token but to a unit as a whole. This system also lets the annotator establish the correspondences between units within such parallel texts where units from both sides may be of arbitrary length (Arkhangelskiy, Mishina, Pichkhadz 2014, 103).

There also exists a morphological tagger suited specifically for processing Middle Russian texts, the so-called RNC analyzer, developed at Higher School of Economics (Moscow) for annotating the Middle Russian subcorpus of the Russian National Corpus (Berdičevskis, Eckhoff, Gavrilova 2016, section 1). The RNC analyzer is a rule-based system in the UniParser format (see below), which is able to give a grammatical annotation to any text for whose language there exists a properly formed grammar description. It could be also useful as an example of a successful automatic tagger for Middle Russian (even though we do not think there is an urgent need for fully automatic annotation in a relatively small corpus like LRC). More on the principles of the UniParser-based RNC analyzer one may find in (Gavrilova, Shalganova, Liashevskaia 2016).

Basic technical details

LRC must provide online access to the electronic editions of Latin and Middle Russian texts, as well as a set of search tools for processing the data contained in these texts. The texts would be stored as XML documents for the purposes of their accessibility and easy processing. As to the search and annotation instruments, some considerations on that topic will be given in the following parts of the current article.

Preprocessing and annotation levels

In the next parts of the article we will give a brief outline of the steps for processing and annotating the texts for LRC. First of all, let us shortly list these steps in the following scheme (fig. 4).

This is the tentative order in which the required steps will be applied to the texts of the corpus. Undoubtedly it could be much more fine-grained and detailed, but this is only a preliminary outline, so we kindly ask the reader to forgive us for such a short description of the annotation process. We hope that some details of the process will be further clarified below.

Preprocessing and normalizing the text

Handwritten texts may considerably differ as to their graphics, orthography and punctuation systems. If one wants to establish a searchable and uniform corpus of such texts, one cannot let the chaotic richness of particular writing systems belonging to various scribes or scribal traditions survive the digitalization of the text. Thus, every text chosen for LRC has to be previously normalized in respect to its graphics, orthography and punctuation systems.

In the following subsection we will try to give a brief outline of the normalization system which we tend to develop for the Corpus.

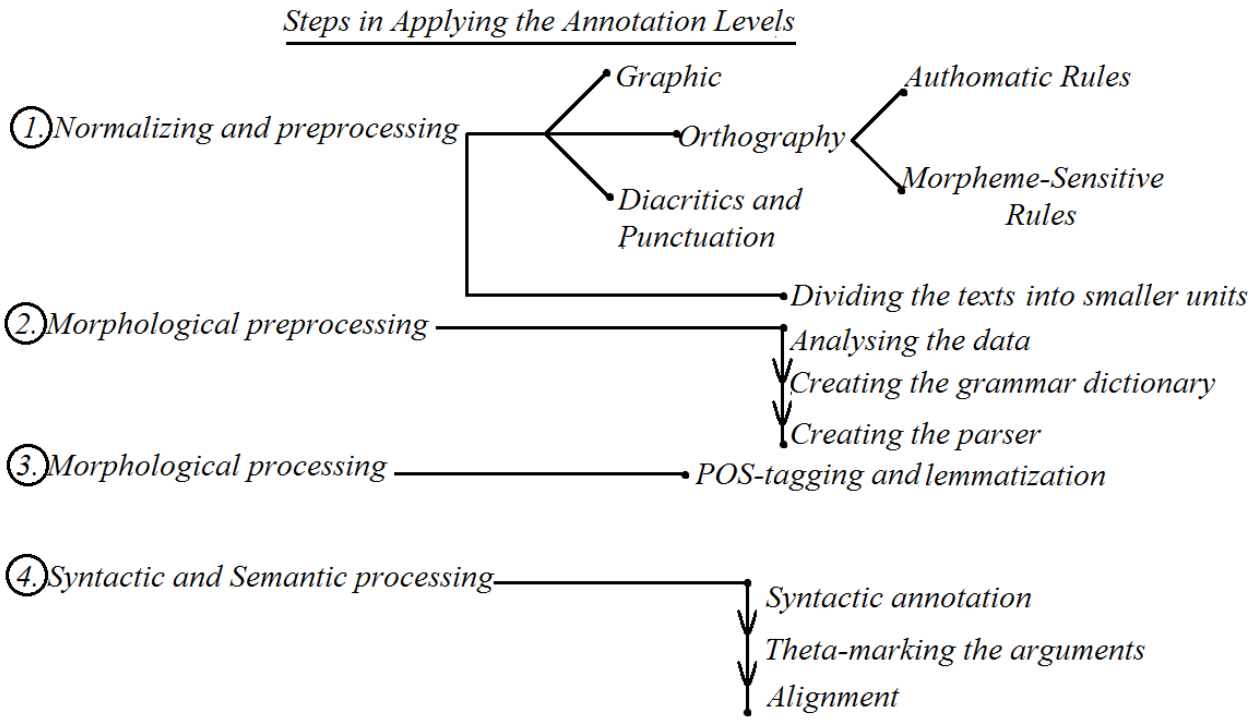

Fig. 4. Steps in applying the annotation levels in LRC

Graphics

In order to give a unified representation of the handwritten sources in LRC, the graphics system used in LRC must fit some general criteria, which are given below.

Preserving the Necessary Graphemes. The first criterion demands that the texts have to preserve as many graphic units from the original manuscript copies as necessary, no less and no more. In general, if there is a distinction between the graphic unit G_1 and the graphic unit G_2, this distinction must be preserved, if it regularly conveys some necessary grammatical information and is typical for the text (or texts). For example, there are two graphic units **е** and **ѣ**. We can see that the second one, **ѣ**, is regularly used in a certain set of morphemes, where **е** is never found, e. g. in certain roots (***крьи́-: крьп-каа, у-крьп-лѧ́ше***) or in the *ā-declension and *ŏ-declension allomorphs of the locative singular morpheme: ***по ѡ́ноу пучи́н-ѣ*** and ***ѡ’мѣст-ѣ***. Correspondingly, if there is no morphemic or other grammatical distinction inside a pair of graphic units (that is, if they are both possible in the same morpheme under the same circumstances), they are in fact free graphic variants of the same grapheme, and one of such graphic units must be eliminated and substituted by the other one everywhere, like in the following examples:

(a) ***я’ : ѧ’***

There are some instances where the use of ***я’*** or ***ѧ’*** is fully arbitrary, as in the following pair of contexts, where both units occur in the same position and in the same word:

вели́кую пучи́ну, и’<u>я́зы́ки</u> синейскыѧ : стра́ну ѡ̀ну и’<u>ѧ́зы́къ</u>
sinum magnum & Sinarum <u>populos</u> : regionem et <u>gentem</u>

It is obvious that in the examples given above (opposite to the situation in modern Church Slavonic) there is no real distinction between the first and the second grapheme.

(b) ***у : ѵ : оу : ов***

The examples given below demonstrate that the use of these four graphic units doesn't exhibit any grammatic or semantic differences, or complementary distribution between them:

у҃ра(з)ѵмѣвъ : оудо́бнѣ : у҃бо : ѿту́доу : не ов҃ : ов҃́бо : тре́бовю(т)

That is why the four variants have to be replaced by only one of them, namely the variant ***у***.

Finally, some pairs of graphic units are superfluous from the point of view of our knowledge about the phonological system of the language of the time. For instance, even though the graphic unit ***s*** in the pair ***s : з*** has a specific distribution, being used in a certain set of roots, e. g. ***svѣр-, sим-, snл-***, in the 16th century it neither reflects any special phoneme distinct from /z/ nor is used to distinguish homophonous morphemes. That is why it must also be replaced by its more common counterpart, **з**.

Avoiding the allographic variation. Once the set of necessary graphic units, i. e. graphemes, is defined, and all alternations in the same position are eliminated, we must also eliminate the sets of variants whose members differ only with respect to their mutual complementary distribution. If there is any complementary distribution (i. e. allographic variation) within a grapheme, the number of allographs must be reduced to one. Below we will consider the most widespread example, the pair ***ї : и***.

From the very time of the so-called Second South Slavic Influence (Grot 1894, 59), the pair of graphic units ***ї : и***, whose distribution had been rather free before, became allographs of the grapheme ***и***, where ***и*** served as a primary allograph, and the second allograph, ***ї***, was placed before the vowels (this rule stayed nearly untouched[4] until the 1918 orthographic reform).

Undoubtedly, there is no need to preserve such allographic alternations in LRC, so all such cases are subject to reduction. The discussed case, for example, has to be reduced to a single graphic variant of the ***ї : и*** grapheme, namely ***и***. For instance, in both forms ***sі́мныѧ*** and ***помоли́шѧ*** the pair ***ї : и*** should receive the same representation as ***и***: ***зи́мныя*** and ***помоли́ша***.

Reducing unnecessary diacritics. The Second South Slavic Influence reintroduced a number of long-forgotten and unnecessary diacritic signs borrowed from the Greek script. Most of them, like aspiration signs, are *completely useless*. Let us call them the signs of the *first type*. The other ones (which we will call *the signs of the second type*), usually stress signs, *may be of some use* for us, primarily because they indicate the place of the word stress. The signs of the first type must be completely eliminated from the text. Among the signs of the second type, all signs that have similar function must be reduced to one. After such procedures the text will preserve the stress signs, but lose all the unnecessary South Slavic ornamental elements, like in the following example:

но̂ sі́мныѧ snѣлности . и҃ стра́ны о҃́ноѧ лю́тости у҃боя́вшесѧ , помоли́шѧ своего кора́бленача̂́лника маггела́на → но́ зи́мныя зѣлности и стра́ны о́ноя лю́тости убоя́вшеся, помоли́ша своего кора́бленача́лника Маггела́на.

In this way all the linguistic information is preserved, while the text makes one more step towards normalization.

[4] Except the cases of morphemic borders like шести-аршинный, пяти-этажный, ни-откуда see (Grot 1894, 60).

Orthography

The next necessary step in the normalization of a historical text is its orthographical normalization. That means that a uniform set of orthographic rules is applied to each text in order to render it homogenous and searchable. Below we will consider some of the issues connected with this task.

Establishing the set of automatic rules. By *automatic orthographic rule* we mean a rule which is applied to any string of symbols independently from its morphemic status. In modern Russian language the examples of such rules are writing the strings ***чу, щу*** with letter ***у*** instead of ***ю***, and the strings ***жи, ши*** with letter ***и*** instead of ***ы***.

If there are any alternations of graphemes in the same context, one of them must be chosen to be used in this context permanently. In the modern examples given above only the strings ***чу, щу, жи*** and ***ши*** are approved, while the homophonous strings ***чю, щю, жы*** and ***шы*** are considered to be clumsy errors. The same principle can be applied to strings in any texts, including those in LRC. Given such pairs *as* ***'разсудѝшя'~ 'разсудѝша', 'бы́шя'~'бы́ша', 'ви́дѣшя' ~ 'ви́дѣша', 'приплу́шя' ~ 'приплуша'*** etc., one can easily deduce that the differences within such pairs (where each member is the 3rd person plural active aorist form of the verbs ***разсудити, быти, видѣти*** and ***приплути*** correspondingly) are merely orthographic, and can be unified by applying the rule ***[š'a] → ша*** (or conversely ***[š'a] → шя***). Then such forms can be unified by means of a simple method like the following one (written in Python 3.6): *text = text.replace('шя','ша')*, where *text* is the name for the variable containing the text subject to normalization. A similar example is provided by the pair ***щи ~ щы (су́щимъ ~ су́щымъ, преимѣ́ющыхъ ~ преиму́щихъ)***, and it can be dealt with after the same manner. Of course, there are a great deal of similar cases, and thus it is crucial to create a sufficient set of automatically applied rules, which could rule out the orthographical contradictions where it is possible to do so without human supervision.

Establishing the set of morpheme-based rules. Unfortunately, we have come to the point where the normalization issue overlaps with the annotation one. It is quite evident that a corpus without annotation is nearly useless for most specialists. A corpus must be at least POS-tagged and morphologically annotated. It is undoubtedly desirable that a corpus also contain a morpheme annotation, i. e., that words in the corpus are represented as lists of morphemes of which they consist. Then such lists could be transformed into a set of all morphemes found in the corpus. Many morphemes have more than one allomorph, and it is strongly desirable that the orthographic representation of the allomorphs have some basic principles. Thus, to represent the allomorphs correctly one has to establish the set of rules considering their morphological representation. Allomorphs can be defined either phonologically, or morphologically. For example, given a root */(slad ~ slat) ~ slažd/* we can say that the first two allomorphs are phonologically defined (<u>slad</u> comes before a voiced consonant or a vowel, *slat* comes before an unvoiced consonant), and the last is morphologically defined (Nida 1949, 44–45), because it appears only before a special subset of suffixes. We suppose that the phonologically defined allomorphs do not have to be orthographically distinguished at all, while the morphologically defined ones do. Thus, the two forms of the same root like in *рѣд-ко* and *рѣт-костию* have to be unified, being its phonological allomorphs: *рѣд-ко* and *рѣд-костию*. We also consider it reasonable to apply the same unification principle to suffixal and prefixal elements. Hence, sets (1a), (2a) and (3a) given below have to get the same orthographical representation (1b), (2b), (3b) of their phonologically defined allomorphs of the suffix *–id– 'belonging to a certain geographical region'* (Table 1).

Table 1. Actual and normalized forms of the lexemes featuring the suffix –id–

a	ACTUAL FORM	b	NORMALIZED FORM
1)	перс-ид-скою	(1)	перс-ид-ск-ою
2)	еспер-и́т-скыхъ еспер-и́т-цкимъ еспер-и́ц҃-кихъ	(2)	еспер-и́д-ск-ихъ еспер-и́д-ск-имъ еспер-и́д-ск-ихъ
3)	молук-и́т-цкимъ молук-ит-цкимъ молук-ит-цкихъ молук-ит-цкия молук-ит-цкымъ молук-ит-цкыхъ молук-ит-цкыя	(3)	молук-и́д-скимъ молук-ид-скимъ молук-ид-скихъ молук-ид-ския молук-ид-скымъ молук-ид-скыхъ молук-ид-скыя

So, the general rule for a morpheme-based orthography must be the following:

(a) If allomorphs are phonologically defined, they preserve the default orthographic form.

(b) If allomorphs are morphologically defined, they orthographically represent the actual phonemic form.

Proper names. As to the proper names, it seems to us that it is desirable to preserve their actual form with some minor exceptions. The basic principles are the following: (a) the uniformness of the root morpheme; (b) the regularity of affixes. The uniformity of the root morpheme means that there must be as few graphic variants of this morpheme as possible. If there are two or more competing graphic variants with the same or nearly the same pronunciation, we have to choose one of them to be the only one throughout the whole text. The most important factors here are (1) the frequency of such graphic variant in the text, (2) its conformity with the orthographic rules accepted for the text, (3) its conformity with the corresponding form in the Latin source of the text.

For example, given a set of tokens with the root ***субу́ѳ-/субу́сѳ-: субу́сѣъ, субу́ѳскому, субу́ѳускаго, субуѳскии, субуѳскому, субуѳстяномъ, субуѣъ-***, we choose the variant ***субу́ѳ-***, because it is the most frequent one and corresponds better to the Latin root *Subuth-* in the words like *Subuth, Subuthicus* etc.

The affixes of proper names have to be treated as the affixes in any other word (see the examples with the stems ***перс-ид-ск-, еспер-и́д-ск-, молук-и́д-ск-*** in the previous section).

Punctuation

Defining the set of punctuation marks. The sets of punctuation marks for different manuscripts of the 16th century can vary in an unpredictable manner. The rules according to which these marks were used seem very vague, if not completely arbitrary. At least, for nearly each scribe there was a unique system of using them. That is why it seems reasonable to limit the use of punctuation marks, preserving only those which are used in the modern Russian punctuation system.

Partitioning the text

Latin and Russian texts of the 16th century are usually divided into parts which we call sentences, i. e. strings of words from a full stop to a full stop. It is obvious that such division doesn't have any linguistic base. To allow for a correct in-depth syntactic analysis of the Corpus, one has

to choose the maximal unit of division based on syntactic criteria, not an arbitrary string of words beginning with a capital letter and ending with a full stop. For this purpose, we propose a special unit called *block*. In terms of phrase structure, *block* is the phrase which is not dominated by (or, in other words, not included in) any other phrase. In terms of dependency structure, *block* is the dependency tree whose head is not dependent from any other head. In general, *block* is the tree which is not a subtree of any other tree.

Let us consider the beginning of the Latin text of "The Letter on the Moluccas" (Table 2).

Table 2. Some examples of the so-called blocks

LATIN SOURCE TEXT	RUSSIAN TRANSLATION
1. De Moluccis insulis, itemq*ue* aliis pluribus mira*n*dis, quae nouissima Castellanorum nauigatio, Serenis*simi* Imperatoris Caroli V auspicio suscepta, nuper inuenit, Maximiliani Transyluani ad Reuerendis*simum* Cardinalem Saltzburgensem epistola lectu perquam iucunda.	1. О Молуки*дскихъ* о́стровѣхъ и и́ныхъ мно́г*ихъ* ди́вныхъ, *ихже* новѣишее плава́ние кастелла́новъ, ре́кше испа́нскихъ, по*мщ*а́ниемъ кро*мч*а́ишаго самоде́ржьца Каро́ла пя́таго събра́но, е*ж*е но́во обрѣ́те, Ма*ѯ*имилиа́на Тра́нъсилва́на къ ч*е*стнѣи́шему кардына́лю салтызву́ръенскому еписто́лиа краснѣи́ш*а* чте́ние*м*ъ.
2. REVERENDISSIME ac Illustris*sime* Domine, domine mi unice, humilli*me* commendo.	2. Ч*е*стнѣи́шии мнѣ́ и пресвѣтлѣишии вл*а*ды*к*о, вл*а*ды*к*о мо́и въж*д*елѣннѣи́шии, <...>
3. Rediit his diebus una ex quinq*ue* illis nauibus, quas Caesar sup*er*ioribus annis, du*m* Caesareae Augustae esset, in alienu*m* et tot iam s*a*eculis incognitu*m* orbem miserat ad inquirendu*m* insulas, in quibus aromata p*ro*ueniu*n*t.	3. Възврати́лся е́сть въ днехъ си́хъ единъ отъ пятихъ кора́бль онѣхъ, и́хже ке́сарь въ пре*ж*нихъ лѣтехъ, въ ни*х*же кеса́рское управля́ше нача́лство, посла́лъ е*сть* въ стра́нныи и толи́кими уже вѣки незна*е*мыи ми́ръ къ разсмотрѣ́нию острововъ, въ ни*х*же ражаю*т*ся арама́ти.

Here the numbers 1, 2 and 3 mark the corresponding maximal syntactic units of the texts. Number 1 is the title. Number 2 is the address. Number 3 is a compound sentence. All three belong to different phrase categories, but all three are *blocks*, because for each of them there is no higher phrase which includes them (or no higher head from which their heads depend, in terms of dependency grammar).

Morphological processing and annotation

Inflectional morphology

The grammar annotation for a corpus must include lemmatization and grammar analysis. For a language featuring relatively high degree of variability in morphology, the inflectional model cannot be determined by an apriori set of rules taken from a kind of textbook: *it must be drawn from the corpus data*. Regarding the mixed Russian — Church Slavonic character of the language of Middle Russian translations from Latin, this is the only possible solution for the LRC texts, because it is not likely that any existing grammar description of Middle Russian or Church Slavonic could adequately account for the degree of variation found in real texts.

To extract the grammar data from a set of texts, one must assume a particular descriptive model. We tend to assume two closely related models for inflectional morphology description, namely that proposed by A. E. Polyakov for the Church Slavonic subcorpus of the Russian National Corpus (Polyakov 2014, 251) and that created by T. A. Arkhangelskiy and known as UniParser. Polyakov's model consists of two basic components:

(1) a grammar dictionary;

(2) a table of inflectional types (paradigms).

The grammar dictionary is a list of lexemes with information on their inflectional features. Each lexeme in the dictionary must contain the following information (Polyakov 2014, 251):

— lemma and its variants;

— POS tags;

— inflectional type or paradigm (represented by a certain paradigm code), and irregular inflectional forms.

Polakov claims that a dictionary entry may also contain some explanatory remarks on the meaning of rare words, but we regard it to be superfluous, mostly due to the fact that it doesn't have any connection with grammar. As mentioned before, Middle Russian paradigm types must be extracted by means of analyzing the set of texts, not by any existing language description. Hence there are some major steps constituting the process of creation of the inflectional morphology model for the Middle Russian part of LRC (Polakov 2014, 252–253):

1) First of all, it is necessary to create the list of inflectional forms found in the texts of the corpus. This is easily done, for example, by means of the following Python 3 program (where *text* is the variable for the set of texts in the corpus):

```
slovar = text.split(' ')
slovar_1 = sorted(set(slovar))
for i in slovar_1 :
  print(i)
```

Having a list of inflectional forms, we can sort it by desinence using the following program (where *text* is an alphabetic string of inflectional forms created by the previous program):

```
text_reversed = text[::-1]
text_reversed_2 = text_reversed.split(' ')
text_reversed_3 = sorted(set(text_reversed_2))
text_reversed_4 = str(text_reversed_3)
slovar = text_reversed_4[::-1]
slovar_2 = slovar.split(' ')
for i in slovar_2 :
 print(i)
```

2) Secondly, the most frequent words must be manually POS-tagged and lemmatized. We consider it reasonable to adopt the list of POS tags used in RNC Middle Russian corpus (Berdičevskis, Eckhoff, Gavrilova 2016, section 3.3.1.), which is with some slight modifications given in the POS tags chart (Table 3).

3) The next step is forming the inflectional classes for the lemmatized words. Though these classes have to be drawn from the corpus analysis, it is possible to borrow some basic principles of their representation (as well as some coinciding paradigms) from Polyakov's Church Slavonic grammar dictionary.

4) Finally, the annotator must apply the created paradigms to the rest of the inflectional forms, checking the results of the tentative automatic or semi-automatic analysis and improving them by manually correcting the wrong guesses of the analyzer.

Another way of producing a formalized description of a language accounting for the facts extracted from the corpus is the UniParser formalized grammar description format developed by T. A. Arkhangelskiy (Arkhangelskiy 2012), which is freely available here: http://languedoc.philol.msu.ru:8082/fieldling/uniparser/. UniParser allows for describing the grammar of a certain natural language as a set of UTF–8 plain text files. Its universal character (it is not designed for any particular language) lets us assume that it could be applied to our material as well.

As to the Latin part of LRC, it is unlikely that its morphology could feature any major deviations from the standard variant of this language, and that is why we consider it possible to use the existing information on the Latin inflectional types instead of thoroughly analyzing the inflectional morphology of the Latin texts in the corpus. The Uni-Parser format, being a universal tool for formalized language description, may be applied to Latin material too. In addition, there is also the so-called Classical Language Toolkit (CLTK) for Python 3.6 whose aim is to provide users with automatic processing tools for Greek and Latin. It may be used to lemmatize the Latin part of LRC as well.

Derivational morphology

At the moment we do not plan to annotate the derivational structure of lexemes in the corpus, but it may be a prospective task, especially regarding the fact that some of the words in Russian translations from Latin may feature one-to-one morphemic correspondences to their Latin prototypes, as in the following examples given by E. S. Fedorova (Fedorova 1999b, 90): ***ab-surdum → о-глушено, ob-iectio → о-пирание / воз-спорь, con-venit → съ-идется*** etc.

Syntax

Syntactic model

Before developing the fine properties of syntactic annotation for a corpus, one must choose a certain syntactic model whose principles would underlie the annotation structure. In general, there are two most widespread models of syntactic representation for natural language, namely dependency and constituency (Osborne 2014, 604). Both PROIEL and TOROT projects use a variant of dependency grammar (Haug et al. 2009, 27; Eckhoff, Berdičevskis 2016, 63). The developers argue that their choice was predominantly determined by the fact that the languages in both corpora had a rather free word order, so it would be convenient to use the formalism where word order information were kept out of the syntactic model (Haug et al. 2009, 27). We are not inclined to regard this property of dependency grammar (hereinafter DG) as an advantage; moreover, it seems to us that DG has a number of other disadvantages, especially

Table 3. POS tags used in RNC Middle Russian corpus

POS tag	Its meaning
A	Adjective
A-PRO	Adjective pronoun
ADV	Adverb
ADV-PRO	Pronominal/interrogative adverb
CONJ	Conjunction
INTJ	Interjection
N	Noun
N-PRO	Nominal pronoun
Q	Quantifier word/cardinal numeral
P	Preposition
V	Verb
D	Determiner

in comparison with some variants of phrase structure grammar (hereinafter PSG), based on constituency principle. Thus, we will continue this section as a gradual comparison of DG in the variant adopted for PROIEL and TOROT, and PSG in the variant of X-bar theory (Chomsky 2015, 45 ff.; Carnie 2008, 112–132).

Basic notions of the X-bar theory

There are some concepts which distinguish the X-bar theory from the basic variants of PSG. The first one is the notion of a head, which X-bar theory shares with dependency grammars. Each phrase is regarded as having a single head, i. e., the element which determines the syntactic properties of the whole phrase (Chomsky 2015, 47; Melchuk 2014, 13; Zwicky 1985). The head projects higher layers of structure, adding one element (each containing one dependent phrase) at a time (Carnie 2008, 120). The head of a phrase is an item of the lexicon; if a head item has substantive content, it is called *lexical* head; a head item without substantive content is called *functional* head (Chomsky 2015, 47–48). The following table (table 4) exhibits the most common categories of *lexical* heads and the corresponding phrases headed by these categories:

Table 4. The most common lexical heads

CATEGORY		PROJECTED PHRASE	
NAME	**ABBREVIATION**	**NAME**	**ABBREVIATION**
Noun	N	Noun phrase	NP
Verb	V	Verb phrase	VP
Adjective	A	Adjective phrase	AP
Adverb	Adv	Adverbial phrase	AdvP
Preposition	P	Prepositional phrase	PP
Complementizer	C	Complementizer phrase	CP

The phrases projected by different heads are believed to have the same inner structure. Let us substitute X, Y, Z or W for any head category. Then the inner structure of XP, i. e. the phrase headed by X, and the Phrase Structure Rules for its formation will be as in the following figure (fig. 5).

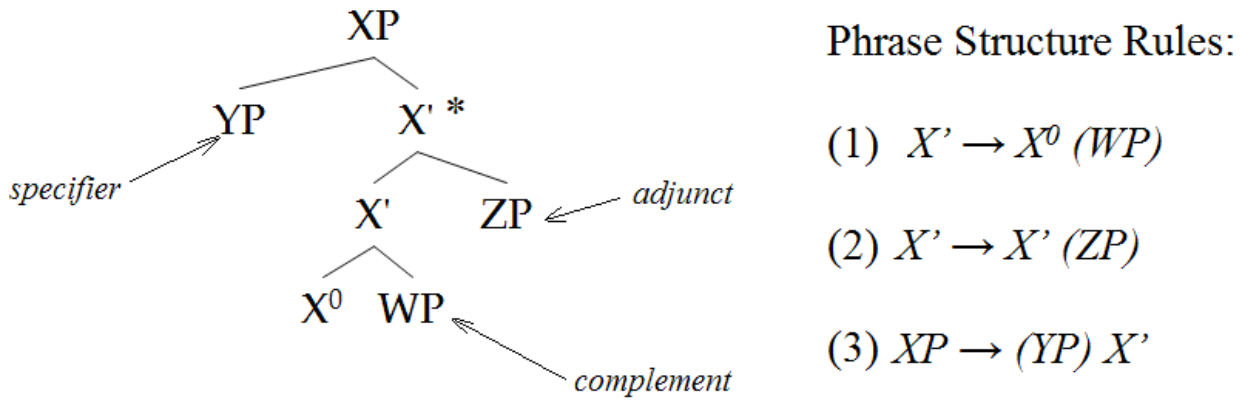

Fig. 5. The structure of XP and the corresponding phrase structure rules

The XP consists of several levels (Chomsky 2015, 48–49). The first level is formed by the head marked as X^0 and its sister phrase marked as WP, see rule (1). They form the first X', i. e. the first X-bar level. The first X' level can add a ZP and form a new X' level, see rule (2). This rule is recursive; every X' level (called intermediate projection) is able to add a dependent phrase, thus forming a new X' level. The iterative character of the X' level is marked by an asterisk in the tree given above. Finally, there must be such an element YP, the addition of which finishes the construction of the phrase headed by X^0, marking the whole construct as maximal projection XP, see rule (3). The layered character of a phrase allows for establishing the relations between its parts regarding their position. There are three basic relations, namely *complement-of*, *adjunct-of* and *specifier-of* (Chomsky 2015, 47–48), which we have marked by arrows in the above illustration. Below we give their definitions, taken from (Carnie 2008, 122).

A phrase that is a sister to a head is its *complement*. Phrases that are sisters to X' levels and are daughters of other X' levels are *adjuncts*. A phrase that is a sister to the X' level and a daughter of a maximal category (i. e. of XP) is a *specifier*. As we will see later, these relations can prove extremely useful in defining some crucial linguistic notions sometimes regarded as primitives.

Also note that phrase structure tree may as well be represented as a bracketed string of characters, see (fig. 6).

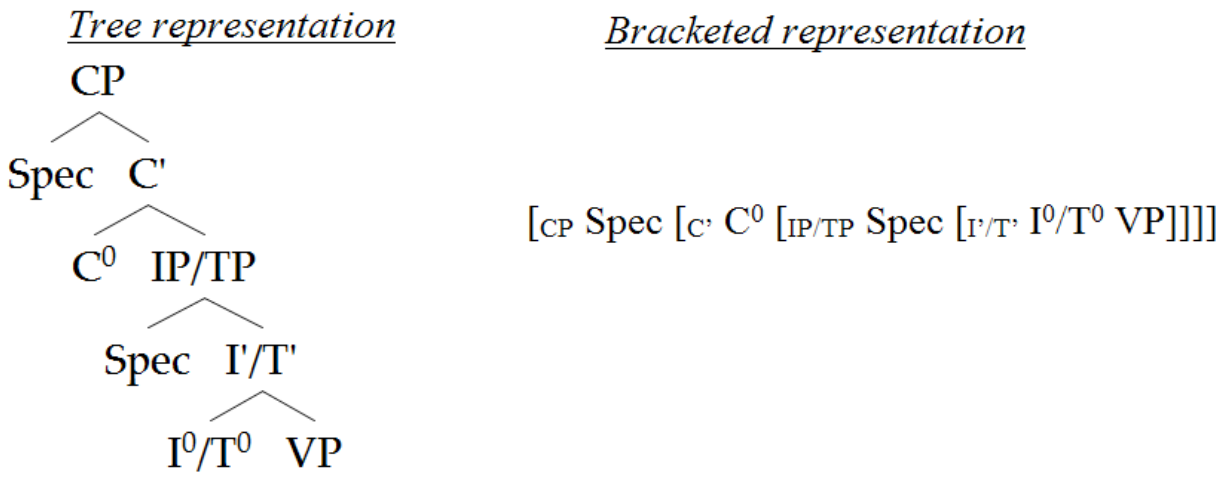

Fig. 6. Tree representation and bracketed representation

The other property of X-bar theory which is crucial for us is the idea of clause structure. Here we will give a brief outline of this structure, following (Chomsky 2015, 49). The full clause is assumed to be headed by a complementizer C hence being a CP. The complement of C is a propositional phrase headed by a functional category I (inflection) or T (tense), thus being IP or TP. I^0 or T^0 has the obligatory complement VP, which contains the predicate and its arguments:

The last crucial notion is the idea of *movement*, which means that the elements are able to change their initial position and move to a new one having left a *trace* (or a silent copy) of themselves in the position which they obtained initially. This simple mechanism has many advantages for the explanation of the linear order of elements and other relations in the tree. The displaced element and its traces are coindexed by a subscript letter (typically *i, j, k*) and linked by an arrow line. Of course, there is a number of other substantial notions which are characteristic of X-bar theory and other modules of generative grammar, but it would be superfluous to consider them in this proposal. Now, let us proceed to the comparison between the opportunities given by DG and the X-bar theory.

Subject, Object and Other Notions: A Comparison Between DG and PSG

The PROIEL DG annotation system allows for specifying each relation between two elements by ascribing it a particular type, e. g. *subject-of* (SUBJ), *object-of* (OBJ), *attribute-of* (ATR), *adverbial-of* (ADV) and so on, as demonstrated in an example from the Latin source of "The Letter on the Moluccas", see (fig. 7)[5]:

Magellanus his dictis mirifice irritatus corrigit socios

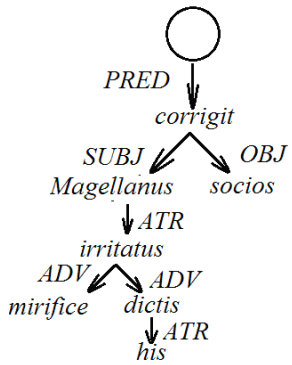

Fig. 7. Some syntactic relations in PROIEL

Subject, object and other relations are regarded here as linguistic primitives. In fact, they are not. Let us consider the following examples, where each subject and object is ascribed a semantic role (fig. 8).

All the examples demonstrate, that neither subjects nor objects can be associated with a single semantic role (also called theta-role). In addition, the examples (1) and (2) make it clear that such roles can be even opposite (*experiencer vs. stimulus, source vs. goal*). The example (3) shows that it is the construction, not the predicate itself, that defines the syntactic status of an argument: for example what would be an object in an active construction, becomes a subject in a passive one, as in (3b), preserving the same semantic role (*adducunt Serranum ~ Serranus adducitur*). Finally, the example (4) provides the evidence that subject cannot be associated with strictly one case, because it bears the nominative in a finite clause and the accusative in a non-finite one (which is traditionally called *accusativus cum infinitivo*).

[5] All the examples of PROIEL annotation presented here are composed using the PROIEL guidelines. Due to some technical considerations they are not taken from actual PROIEL corpus texts.

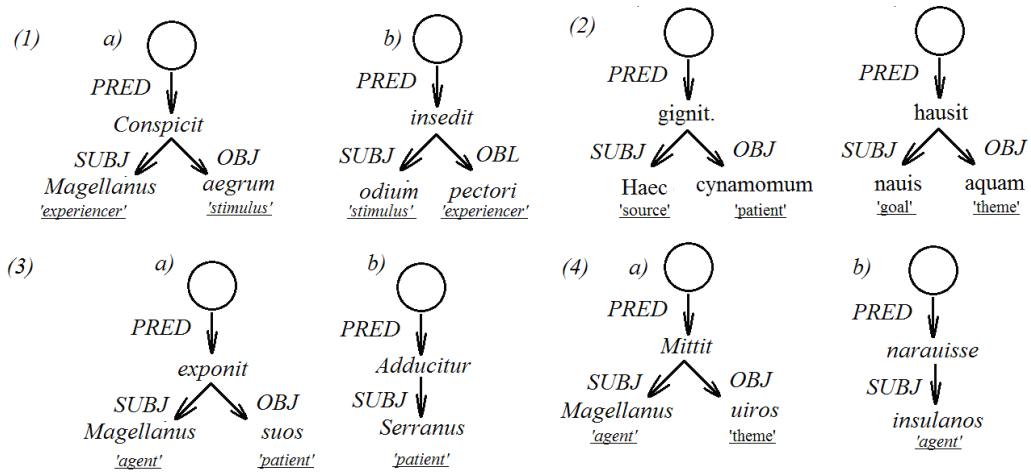

Fig. 8. The semantic roles of various subjects and objects

The DG system is not able to account for all these difficulties — as well as for some others, for example for the evident syntactic prominence of subject, see (McCloskey 1998, 197–198) — and its notions of subject, object and other syntactic relations stay vague and unclear. On the contrary, the X-bar system gives a precise and simple account of the given facts, which is based on the assumption that subject, object, attribute and so on are complex notions formed from a set of more simple relations. All we need for that are the notions of specifier, complement and adjunct, and the tripartite structure of a clause. The subject is taken to be a phrase base-generated in the specifier of VP, where the predicate ascribes it the semantic role (McCloskey 1998, 203–216). Then this phrase is raised into the position of the specifier of TP (frequently noted as [Spec, TP]) in order to get its case. If T^0 is finite, then the phrase in the specifier of TP gets nominative case; if T^0 is non-finite, it fails to ascribe the specifier of TP the nominative case, and it gets another case (accusative in Latin, dative in Slavic languages). So the subject is nothing but the specifier of TP (or IP). The object is, in its turn, the complement of VP (Chomsky 2015, 49). If the object phrase is raised to the position of the specifier of TP, as it usually happens during passivisation, it becomes a subject and gets the case associated with the subject of a particular clause (finite or non-finite). To sum up, the structural prominence of subject and the properties of object are explained by their syntactic position. That is the first fact that urges us to prefer the X-bar representation, like the one given below, to any possible DG annotation (fig. 9).

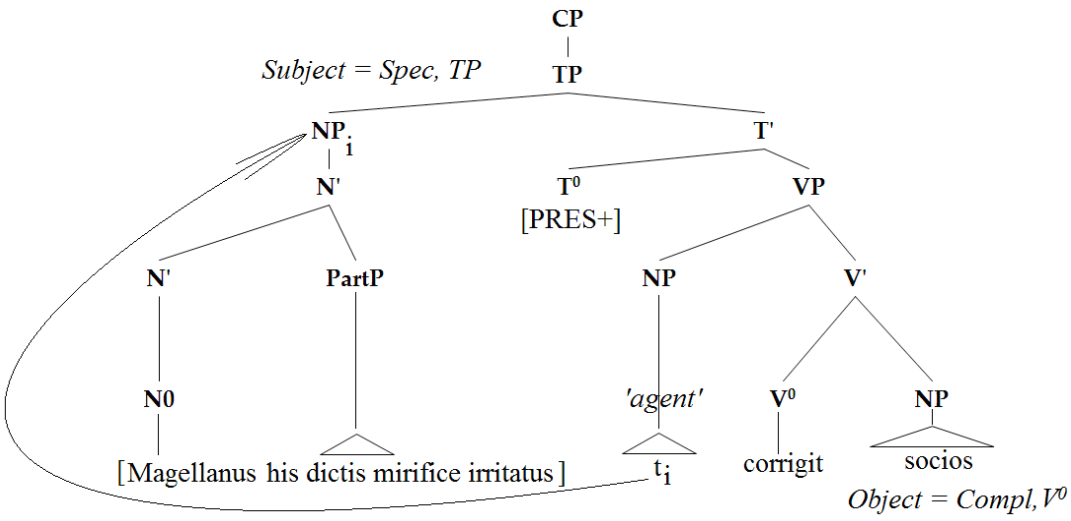

Fig. 9. Subject raising to the [Spec, TP] position in an X-bar tree

Relative Clauses and Movement

Now let us consider the following DG graph of the pair of syntactic units from the Latin-Russian parallel text of "The Letter on the Moluccas" (fig. 10).

regio, quam terram firmam uocant ~ стра́на, ю́же и зе́млю тве́рду нари́чютъ

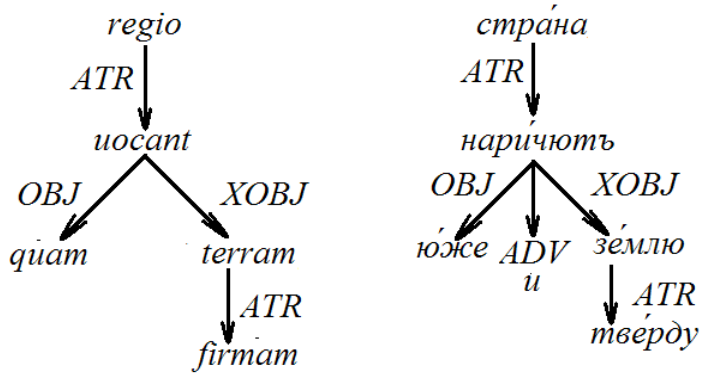

Fig. 10. A PROIEL-like DG representation of a relative clause

The annotation of this sample pair follows the guidelines given for relative clauses in PROIEL (Haug 2010, 38–43). One can see that the relative pronoun *quam* resp. *юже* is taken to be nothing but one of the dependents of the verb. It is really a dependent of the verb, but in addition it is the element which determines the properties of the whole relative clause: a relative clause lacking any relative element is a nonsense! But what determines the properties of a syntactic unit is the head of this unit. So, we come to a controversial situation: on the one hand, the relative element depends on the verb, being its argument; on the other hand, the same relative element is the head of the whole relative clause which contains the verb; to sum up, the relative pronoun holds two separate syntactic positions at the same time. As we can see, the PROIEL annotation cannot account for this problem. Some other types of DG annotation seem to be able to manage this problem a bit better, adopting some mechanisms of displacement, see (Osborne 2014, 619). Nevertheless, it seems to us that the simplest way is to adopt the well-known solution given by the generative grammar: namely, the idea that the relative pronoun is base-generated in a certain position related to predicate or the other element in the tree, and then moves to the leftmost position in the clause, becoming the specifier of a silent complementizer C:

regio, [*quam*$_i$ *C*^{o} [*terram firmam uocant* t_i] ~ *стра́на,* [*ю́же*$_i$ *C*^{o} [*и зе́млю тве́рду нари́чютъ* t_i].

Word order representation and discourse configurationality

Latin, Church Slavonic and Middle Russian are *discourse-configurational languages*, i. e. the languages whose linear word order is predominantly determined by the information structure of utterances (Kiss 1995, 6). There seem to be at least two major notions associated with the informational structure of sentence, namely those of *topic* and *focus* (Kiss 1995, 6, 7–14, 15–24; Bailyn 2012, 266–267). The elements in discourse-configurational languages can be *topicalized* or *focalized*, that is moved to the positions associated with *topic* or *focus* (Bailyn 2012, 267).

We have already mentioned that DG graphs are usually independent from any particular linear order and thus are unable to represent the informational structure of utterances. On the contrary, the X-bar theory demands that the phrase marker for a particular sentence retain the linear order of its elements. That is why X-bar phrase markers allow for marking the informational structure of the elements by hosting them in special functional projections named Topic Phrase

(TopP) and Focus Phrase (FocP). Let us consider the following DG representation of a sentence taken from the Russian variant of "The Letter on the Moluccas" (fig. 11).

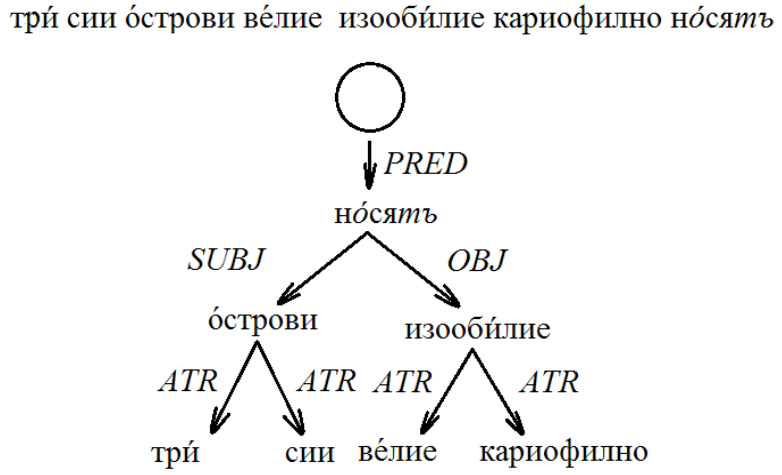

Fig. 11. A DG tree which conveys no information on linear order

The DG representation indicates the dependency relations between the elements; it also marks the type of each relation, but it still may correspond to a large number of possible linear orders:

a. *[Три сии острови] носятъ [изобилие велие кариофилно].*

b. *Носятъ [три сии острови] [изобилие велие кариофилно].*

c. *[Изобилие велие кариофилно] носятъ [три сии острови].*

d. *[Изобилие велие кариофилно] [три сии острови] носятъ.*

e. *[Изобилие кариофилно велие] [три сии острови] носятъ.*

f. *[Три сии острови] [велие изобилие кариофилно] носятъ.*

And so on, having at least 3! = 6 positions for three elements *{три, сии, острови}* inside the subtree headed by the subject, the same number for those inside the subtree headed by the object and also 3! = 6 positions for the subject, object and verb themselves. No part of the dependency tree can throw any light on the word order in the sentence.

It can be argued that some kinds of DG allow for including the word order information into the structure of tree-generating rules, like that described in (Hays 1964), but in fact such rules cannot cope with what is called *discontinuous* or *non-projective phrases*[6], and, Middle Russian and Church Slavonic being discourse-configurational languages, with these languages themselves. Now let us consider the PSG representation of the same sentence in the form of X-bar phrase marker (fig. 12).

Here not only the linear order of elements is preserved as it is, but also the middle-field topicalisation of the noun phrase *велие изообилие кариофилно*, which results in SOV linear order (Bailyn 2012, 273–274), is marked as a syntactic operation (XP-movement). Hence one can easily decide that PSG has more explanatory force as to the linear and informational organization of utterances in discourse-configurational languages than DG.

[6] *Discontinuous or non-projective phrases* are *phrases which are linearly interrupted by some material which doesn't belong to the phrase*.

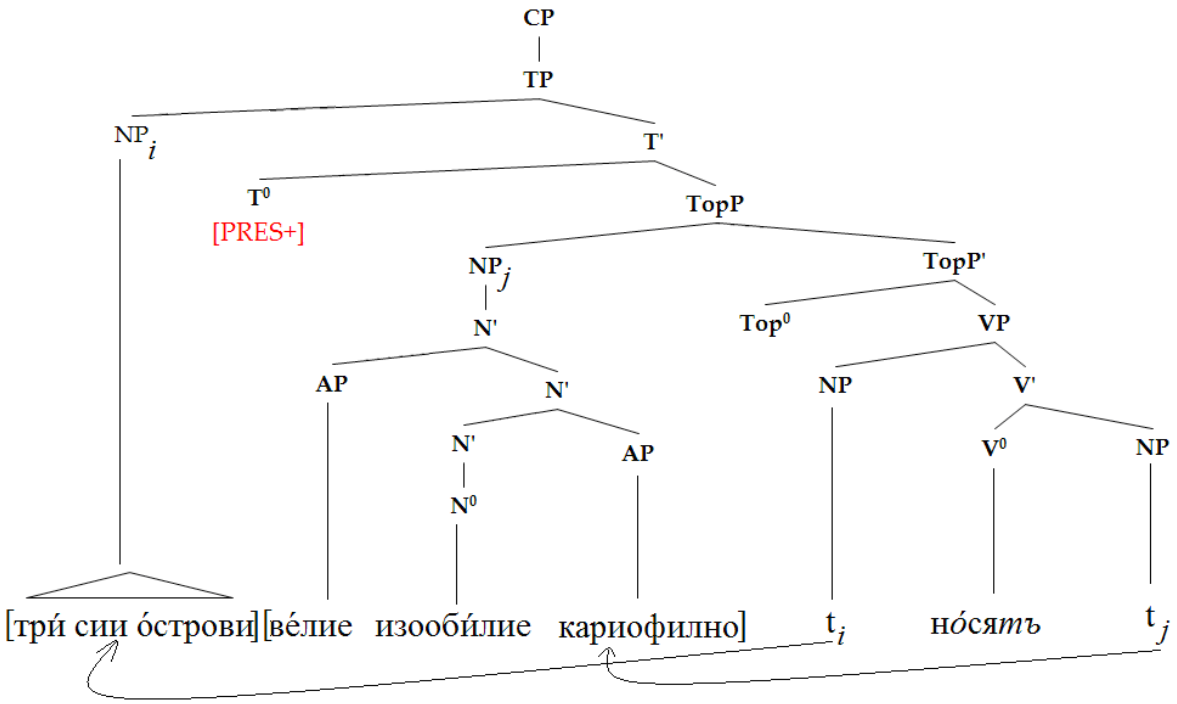

Fig. 12. A phrase marker of [Три сии острови] [велие изобилие кариофилно] носятъ.

Which structure hosts more information?

To conclude with, now it is obvious that both DG and X-bar structures host the essential information about the dependency relations between the syntactic units, but the latter also has a number of mechanisms which can explain the linear order and information structure of utterances, as well as the complex nature of some basic linguistic concepts, while the former is unable to perform the same task. This simple reason convinces us that the X-bar theory is a better basis for corpus annotation than the DG system.

Syntactic annotation

At the current time we are not aware of any syntactic annotation tools for Latin or Middle Russian that use the PSG formalism. Latin and (Old, Middle, Modern) Russian corpora either lack syntactic annotation or are annotated in terms of dependency grammar. The only exception is the SKAT project, which includes some elements of phrase structure annotation in XML format, but its syntactic module is not ready for use so far (Alekseeva 2014), and in addition it is nonetheless based on the system of syntactic relations similar to that from the Russian National Corpus, i. e. dependency relations.

At any rate, there are some opportunities for establishing a PSG annotation system for these languages. These opportunities are based on two main facts:

(1) There exist a number of works exploring the limits and abilities of PSG description for old Indo-European languages, including Latin and old Slavic languages, for instance (Oniga 2014; Danckaert 2011) for Latin, (Mitrenina 2012) for Middle Russian, partly (Isakadze 1999) for Old Russian. Some of such works consider the problem of annotating the texts in these languages, e. g. see (Dimitrova 2011) for Old Church Slavonic.

(2) There are some open source syntax tree generators which let the users render a bracketed representation of a phrase into a syntax tree and back, for example:

Linguistic Tree Constructor (LTC): http://ltc.sourceforge.net/about.html

Syntax Tree Editor: http://www.ductape.net/~eppie/tree/

Besides, as we have already mentioned before, there exists a tool which allows for aligning the strings of word forms (Arkhangelskiy, Mishina, Pichkhadze 2014, 102). In addition, the Natural Language Toolkit for Python (NLTK) (Bird, Klein, Loper 2009, 291–326) supports some patterns of syntactic parsing which may be applied to our material too. All these instruments can be used as a base for our own syntax tree construction device included into LRC, but the question of the subtleties of their application remains quite unclear so far and requires a thorough analysis of the existing parsing techniques, tools and data. Of course, marking up the corpus after the generative manner requires quite a high level of expertise and is not a simple and straightforward task; that is why we hope to make final decision on the details of this procedure only after a tentative mark up of some text fragments.

Prospective directions of research

Grammatical and lexical variation in parallel texts

The relation between the sets of lexical items of the two languages cannot be reduced to a one-to-one correspondence. Particular lexical items are represented in the text only by the word forms, which are obligatory included in larger syntactic units called phrases. Thus, the comparison must begin not at the word level, but rather at the phrasal level considering the paired phrases from the source and target texts. The phrases in both parts of a parallel text must be divided into constituents, which must be paired using the appropriate software until the correspondences between the terminal nodes, i. e. the word forms, are established.

Let us consider three stages of analysis for an imaginary annotator illustrated below:

(1) Finding the phrase structure correspondences for the source and target phrase

In our example (see Fig. 12) the correspondences have to be established between two prepositional phrases: *[ad Taprobanen quam nunc Zamataram uocant]* and *[к Тапрова́ни, уже нынѣ Зама́тару нари́чютъ]*. The corpus software must support manual syntactic annotation for phrases. The next step is the manual alignment of the corresponding nodes in the two phrase markers. Such alignment must look like that: $PP \rightarrow PP$, $P^0 \rightarrow P^0$, $NP_1 \rightarrow NP_1$ and so on, until the moment when each node of the Latin phrase marker is paired with the corresponding Russian node.

(2) Forming the list of lexical correspondences

After that the machine analyses the set of node pairs formed on the previous stage and offers to the annotator the corresponding set of paired word forms, like the one given on the second part of the illustration. If a word form corresponds to a terminal node, the annotator can relate it to a certain existing lemma or add a new one:

Taprobanen acc.sg. | *Taproban-e, N. f. 1gr.* ~ *Тапрова́ни acc.sg.* | *Тапрова́ни, N f. indecl.*

So for a pair of aligned terminal nodes of the source text and the target text one could derive a corresponding pair of lemmata:

Taproban-e, N f. 1gr. → *Тапрова́ни, N f. indecl.*

Such ordered pairs of lemmata *L(l,s) := 'the Latin lemma l corresponds to the Russian lemma s'* may serve as the basis for an automatically formed vocabulary for each parallel text.

(3) Forming the strings of categorial symbols for a particular phrase

The strings of categorial symbols must be formed automatically, being simply the regular linearization of the Latin and Russian tagged phrase markers from the stage 1 (see fig. 13).

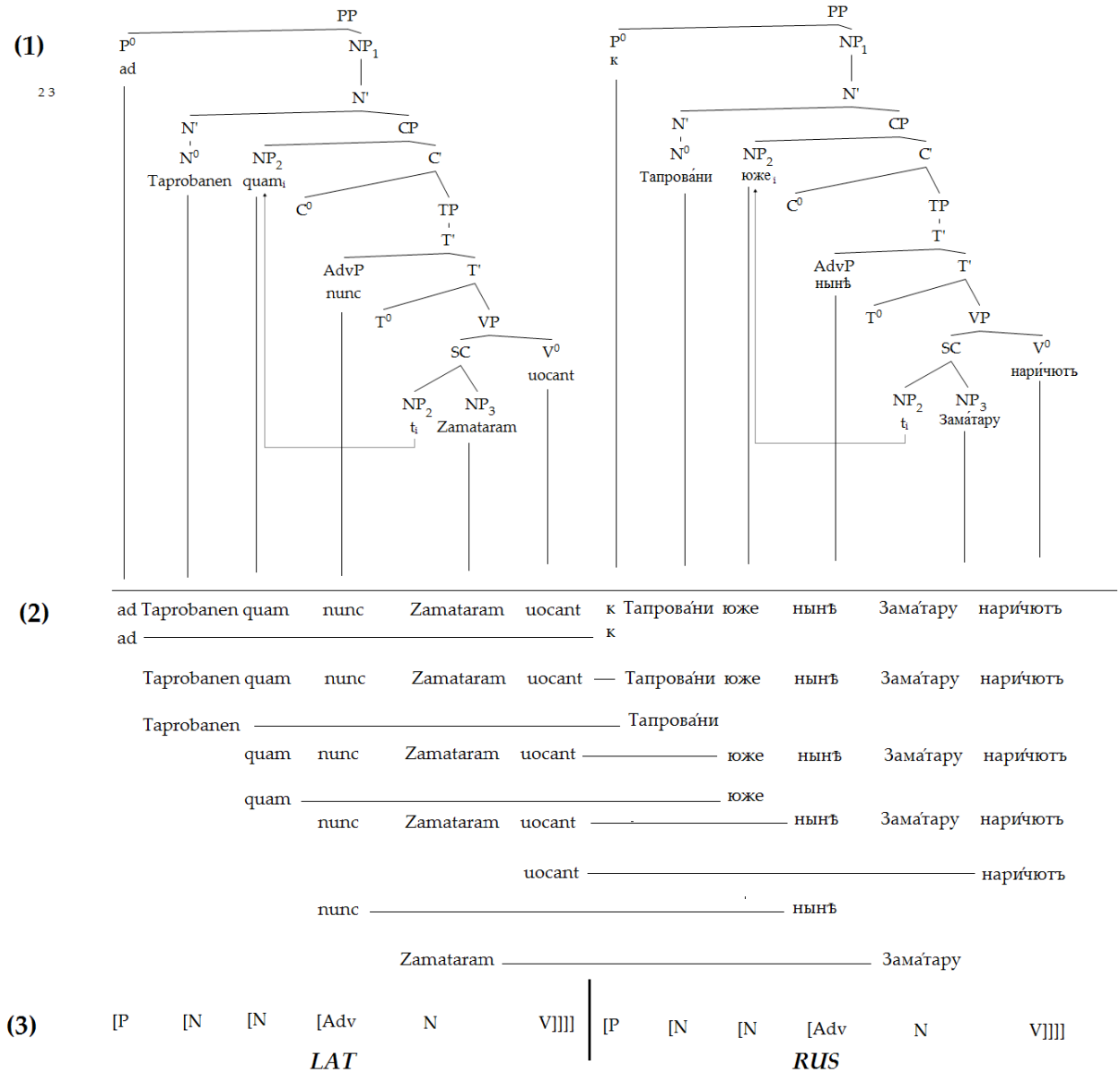

Fig. 13. An example of the lexical and phrasal correspondences in a particular bitext

Sometimes it is impossible to form a one-to-one correspondence between all terminal nodes of the source phrase and all terminal nodes of its target counterpart. For example, in a pair of correspondent noun phrases *[tropicum Capricornum] 'Tropic of Capricorn' ~ [вéсненаго солнечнаго возвращéниа] 'Tropic of Capricorn (a translator's periphrase)'* it is possible to analyse the phrases themselves, but impossible to pair their constituents. In such cases the phrasal material must be stored in the vocabulary as a whole:

[$_{NP}$ tropicus Capricornus] → [$_{NP}$ вéсненое солнечное возвращéние] In another case the Latin noun *Pigmeos acc.pl. | Pigmeus N m. 2* is translated by a substantivated adjective phrase *[$_{NP}$ Ø [$_{AP}$ лакотных возрастом]].* The vocabulary entry for this pair must look like this:

Pigmeus N m. 2 → [$_{NP}$ Ø [$_{AP}$ лакотный возрастом]].

In other words, if on a certain stage there cannot be found any further one-to-one correspondences between the constituents of the Latin and Russian phrases, then the annotator has to enter into the vocabulary the last pair of correspondences, even if this pair contains nonterminal nodes.

Selective features of lexical items

The LCA project must also provide its users with an opportunity to research the lexical compatibility and selective features of the lexical items in the corpus.

Lexical compatibility search

Having an annotated and aligned phrase marker pair for a certain parallel text, one could easily find in it all the entries featuring a certain lexical head. One could also get a list of dependents it is compatible with, as well as the corresponding material in the opposite part of the same bitext.

Thus, the syntactically annotated corpus must give an opportunity to form a compatibility list for the lexical items in it or to supplement the existing vocabularies with such information.

Forming the subcategorization frames

A syntactically annotated corpus can be supplemented with a module serving for extraction of the subcategorization frames. If annotators tag the arguments of each predicate with their semantic role (theta-role), such a module will be able to compare this information with the syntactic position of the arguments and form a subcategorization frame for a certain predicate. Consider the pair of sentences in the (fig. 14).

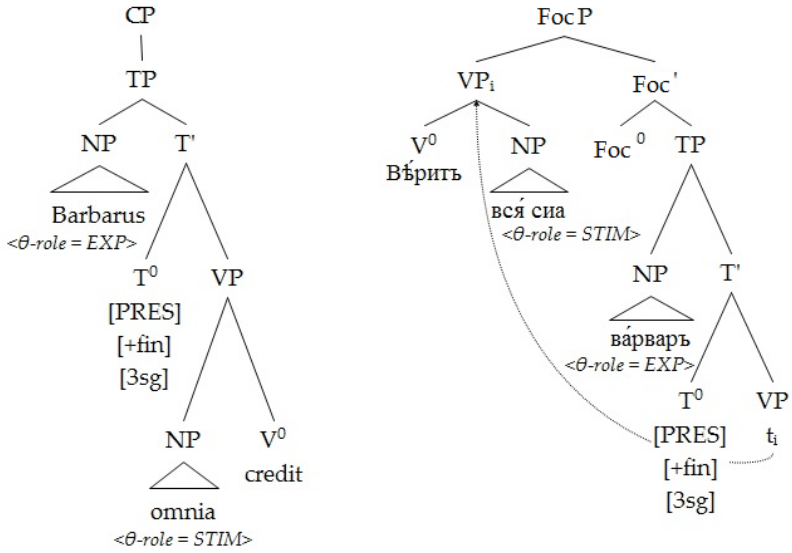

Fig. 14. A pair of sentences with the arguments of the predicate marked with their θ-roles

All the annotator has to do is to mark the subject in these sentences as experiencer (EXP) and the object as stimulus (STIM). Once the annotator does it, the module will form a certain subcategorization frame (Table 5).

Table 5. An example of corresponding subcategorization frames

LAT	RUS
Credo	вѣ́рити
EXP: Spec, CP (subject) STIM: Comp, V0 (object)	EXP: Spec, CP (subject) STIM: Comp, V0 (object)

Such subcategorization frame doesn't mark the structural cases of the arguments, because they depend on the clause type, not on the semantic role of a certain argument (for instance, as we have already mentioned, the subjects of finite clauses bear nominative case, while the subjects of infinite clauses bear accusative or dative).

Finding omissions and additions in parallel texts

It is typical of a Latin-Russian parallel text to include a number of omissions and additions. Translators (or later scribes) could remove large fragments, like in the example from the (Table 6), where the corresponding fragments are highlighted in bold type:

Table 6. An example of omission in translation

LAT	RUS
Nam ut reliqua omitta*m*, tradidit Herodotus alioqui clarissimus autor, cynamomum in auiu*m* nidis reperiri, in quos uolucres illud ex lo*n*gissimis regionibus, et praesertim Phoenix (cuius nidu*m* nescio quis unq*uam* uiderit) detulissent.	**И да́ж*д*ь про́чая оста́влю.**

The syntactic annotation must show where a nonempty set of nodes in a Latin text is conveyed by an empty set of nodes in its Russian counterpart. The search in the corpus or in a particular bitext must find all the results of that kind, letting the researcher assess the reasons and regular patterns of the omissions in a translated text.

There can be found some opposite cases, in which translators added something to the translation. In the example from the (Table 7) the addition is highlighted in bold type:

Table 7. An example of addition in translation

LAT	RUS
atq*ue* adeo Nili fontes et Troglodytas inuenerunt	И са́мыя ни́ловы исто́чники, и глаголе́мыя троглоди́ти, ре́кше подъ землею въ пеще́рахъ живу́щаа, обрѣто́шя

The syntactic annotation must show such superfluous nodes in the translation, to ascertain the absence of any correspondences in the Latin source. Such option could also help the future researchers.

Marking the translator's faults

During the annotation process it is impossible not to find some translational mistakes. This can be regarded as an additional advantage of manually annotated parallel corpora. The translator's fault must be marked with a special metatag, allowing for a grammar note, which could possibly result in lists of typical translational mistakes for various parallel texts.

Conclusion

In this article we have presented a preliminary project of a deeply tagged parallel corpus of Russian translations from Latin, including the information on its goals, purposes, applicability

and structure. We have proposed annotation models for various levels of language representation, paying special attention to the issues of orthographical normalization, morphological and syntactic tagging of the corpus. The creation of such a corpus could provide researchers with a powerful instrument for scholarly activities in the fields of historical linguistics, literary studies and history of culture.

References

Alekseeva, E. L. (2014) Sintaksicheskaya razmetka korpusa drevnerusskikh agiograficheskikh tekstov SKAT [Syntactic tagging of Saint-Petersburg corpus of hagiographic texts (SCAT)]. In: *Strukturnaya i prikladnaya lingvistika.* Iss. 10. Saint Petersburg: Saint Petersburg State University Publ., pp. 345–351. (In Russian)

Arkhangelskiy, T. A. (2012) *Printsipy postroeniya morfologicheskogo parsera dlya raznostrukturnykh yazykov. Extended abstract of PhD dissertation (Philology)*. Moscow, Moscow State University, 24 p. (In Russian)

Arkhangelskiy, T. A., Mishina, E. A., Pichkhadze, A. A. (2014) Sistema elektronnoj grammaticheskoj razmetki drevnerusskikh i tserkovnoslavjanskikh tekstov i ee ispol'zovanie v veb-resursakh [A system for digital morphological tagging for Old Russian and Church Slavonic texts and its use in web resources]. In: V. A. Baranov, V. Zhelyazkova, A. M. Lavrent'ev (eds.). *Pismenoto nasledstvo i informacionnite tehnologii. El'Manuscript–2014*. Sofia; Izhevsk: Bolgarskaya akademii nauk Publ., pp. 102–104. (In Russian)

Bailyn, J. F. (2012) *The Syntax of Russian.* Cambridge; New York: Cambridge University Press, XVIII, 373 p. (In English)

Berdičevskis, A., Eckhoff, H., Gavrilova, T. (2016) The beginning of a beautiful friendship: Rule-based and statistical analysis of Middle Russian. In: V. P. Selegej (ed.). *Computational linguistics and intellectual technologies: Proceedings of the International conference "Dialogue 2016"*. Vol. 15 (22). Moscow: Russian State University for the Humanities Publ., pp. 99–111. (In English)

Bird, S., Klein, E., Loper, E. (2009) *Natural language processing with Python*. Beijing: O'Reilly, XX, 479 p. (In English)

Carnie, A. (2008) *Constituent Structure*. Oxford; New York: Oxford University Press, XVIII, 292 p. (Oxford surveys in syntax and morphology. Book 5). (In English)

Chomsky, N. (2015) *The Minimalist Program: 20. Anniversary edition.* Cambridge, MA: MIT Press, XIII, 393 p. (In English)

Danckaert, L. (2011) *On the left periphery of the Latin embedded clauses. PhD dissertation (Philology)*. Ghent, Belgium, Ghent University, XVII, 387 p. (In English)

Dimitrova, Ts. (2011) *The Old Bulgarian noun phrase: Towards an annotation specification*. Saarbrücken: VDM Verlag Dr. Müller, VII, 273, 28 p. (In English)

Eckhoff, H. M., Berdičevskis, A. (2016) Automatic parsing as an efficient pre-annotation tool for historical texts. In: *Proceedings of the Workshop on language technology resources and tools for digital humanities (LT4DH)*. Stroudsburg, PA: The COLING 2016 organizing committee; Association for Computational Linguistics, pp. 62–70. (In English)

Fedorova, E. S. (1999a) *Traktat Nikolaja de Liry "Probatio adventus Christi" i ego tserkovnoslavyanskij perevod kontsa XV veka*: In 2 books. Book 1. Moscow: Prosvetitel' Publ., 287 p. (In Russian)

Fedorova, E. S. (1999b) *Traktat Nikolaja de Liry "Probatio adventus Christi" i ego tserkovnoslavyanskij perevod kontsa XV veka: In 2 books. Book 2: Prilozheniya*. Moscow: Prosvetitel' Publ., 120 p. (In Russian)

Gaifman, H. (1965) Dependency systems and phrase-structure systems. *Information and Control*, 8 (3): 304–337. DOI: 10.1016/S0019-9958(65)90232-9 (In English)

Gavrilova, T. S., Shalganova, T. A., Liashevskaia, O. N. (2016) K zadache avtomaticheskoj leksiko-grammaticheskoj razmetki starorusskogo korpusa XV–XVII vv. [Lexico-grammatical annotation of the Middle Russian corpus 1400–1700: A computational approach]. *Vestnik Pravoslavnogo Svyato-Tikhonovskogo gumanitarnogo universiteta. Seriya III: Filologiya — St. Tikhon's University Review. Series III: Philology*, 2 (47): 7–25. DOI: 10.15382/sturIII201647.7-25 (In Russian)

Grishman, R. (1999) Iterative alignment of syntactic structures for a bilingual corpus. In*:* S. Armstrong, K. Church, P. Isabelle et al. (eds.). *Natural language processing using very large corpora*. Dordrecht: Springer, pp. 225–234. (Text, Speech and Language Technology. Vol. 11.). DOI: 10.1007/978-94-017-2390-9_14 (In English)

Grot, Ja. K. (1894) *Russkoe pravopisanie; Rukovodstvo, sostavlennoe po porucheniyu 2-go Otdeleniya Imperatorskoj akademii nauk akademikom Ya. K. Grotom.* 11th ed. Saint Petersburg: Tipografiya Imperatorskoij Akademii Nauk Publ., XII, 120, XL p. (In Russian)

Haug, D T. T. (2010) *PROIEL guidelines for annotation*. [Online]. Available at: https://folk.uio.no/daghaug/syntactic_guidelines.pdf (accessed 15.08.2019). (In English)

Haug, D. T. T., Jøndal, M. L., Eckhoff, H. M. et al. (2009) Computational and linguistic issues in designing a syntactically annotated parallel corpus of Indo-European languages. *TAL (Traitement Automatique des Langues)*, 50 (2): 17–45. (In English)

Hays, D. G. (1964) *Dependency theory: A formalism and some observations*. Santa Monica, CA: RAND Corporation, VII, 39 p. (In English)

Isakadze, N. V. (1999) *Otrazhenie morfologii i referentsial'noj semantiki imennoj gruppy v formal'nom sintaksise. Extended abstract of PhD dissertation (Philology)*. Moscow, Moscow State University, 23 p. (In Russian)

Kalugin, V. V. (2001) "Kniga svyatogo Avgustina" v russkoj pis'mennosti XVI — XIX vekov. In: A. M. Moldovan, V. S. Golyshenko (ed.). *Lingvisticheskoe istochnikovedenie i istoriya russkogo yazyka*. Moscow: Drevlekhranilishche Publ., pp. 108–163. (In Russian)

Kazakova, N. A. (1980) *Zapadnaya Evropa v russkoj pis'mennosti XV–XVI vekov. Iz istorii mezhdunarodnykh kul'turnykh svyazej Rossii*. Leningrad: Nauka Publ., 278 p. (In Russian)

Kazakova, N. A., Katushkina, L. G. (1968) Russkij perevod XVI v. pervogo izvestiya o puteshestvii Magellana (Perevod pis'ma Maksimiliana Transil'vana). In: D. S. Likhachev (ed.). *Trudy otdela drevnerusskoj literatury*. Vol. 23. Leningrad: Nauka Publ., pp. 227–252. (In Russian)

Kiss, K. É. (ed.). (1995) *Discourse Configurational languages*. New York; Oxford: Oxford University Press, 402 p. (Oxford Studies in Comparative Syntax). (In English)

Kloss, B. M. (1975) Maksim Grek — perevodchik povesti Eneya Sil'viya "Vzyatie Konstantinopolya turkami" [Maxim the Greek — translator of Aeneus Silvius' narrative "Seizure of Constantinople by Turks"]. In: *Pamyatniki kul'tury. Novye otkrytiya. Pis'mennost', iskusstvo, arkheologiya*. Moscow: Nauka Publ., pp. 55–61. (In Russian)

Matasova, T. A. (2014) Pervaya kniga "Geografii" Pomponiya Mely v drevnerusskom perevode: O retseptsii antichnogo naslediya v russkoj kul'ture XV-XVI vv. [The Old-Russian translation of the first part of Pomponius Melas' "Cosmography": Perception of classical heritage in Russian culture in XV-XVI centuries]. *Aristej: vestnik klassicheskoj filologii i antichnoj istorii — Aristeas. Philologia Classica et Historia Antiqua*, IX: 310–343. (In Russian)

McCloskey, J. (1998) Subjecthood and subject positions. In: L. Haegeman (ed.). *Elements of grammar: Handbook in generative syntax*. Dordrecht: Springer, pp. 197–235. DOI: 10.1007/978-94-011-5420-8_5 (In English)

Mitrenina, O. V. (2012) Sintaksis psevdokorrelyativnykh konstruktsij s mestoimeniem *kotoryj* v starorusskom [The syntax of pseudo-correlative constructions with the pronoun *Kotoryj* ("Which") in Middle Russian]. *Slověne. International Journal of Slavic Studies*, 1 (1): 61–73. DOI: 10.31168/2305-6754.2012.1.1.4 (In Russian)

Mitrenina, O. V. (2014) The corpora of Old and Middle Russian texts as an advanced tool for exploring an extinguished language. *Scrinium. Journal of Patrology, Critical Hagiography, and Ecclesiastical History*, 10 (1): 455–461. DOI: 10.1163/18177565-90000109 (In English)

Melchuk, I. (2014) Dependency in language. In: K. Gerdes, E. Hajičová, L. Wanner (eds.). *Dependency linguistics. Recent advances in linguistic theory using dependency structures*. Amsterdam; Philadelphia: John Benjamins Publishing Company, pp. 1–32. (Linguistik Aktuell / Linguistics Today. Vol. 215). (In English)

Nida, E. A. (1949) *Morphology: The descriptive analysis of words*. Ann Arbor: University of Michigan Press, XVI, 342 p. (In English)

Oniga, R. (2014) *Latin: A linguistic introduction*. Oxford: Oxford Universty Press, XVIII, 345 p. (In English)

Osborne, T. (2014) Dependency grammar. In: A. Carnie, Y. Sato, D. Siddiqi (eds.). *The Routledge handbook of syntax*. Abingdon: Routledge, pp. 604–626. (In English)

Partee, B. H., ter Meulen, A., Wall, R. E. (1990) *Mathematical methods in linguistics*. Dordrecht; Boston; London: Kluwer Academic Publishers, XX, 663 p. (In English)

Polyakov, A. E. (2014) Korpus tserkovnoslavyanskikh tekstov: Problemy orfografii i grammatiki [Church Slavonic corpus: Spelling and grammar problems]. In: A. Kiklewicz (ed.). *Przegląd Wschodnioeuropejski [East European Review]*. Vol. V (1). Olsztyn: University of Warmia and Mazury in Olsztyn, pp. 245–254. (In Russian)

Durandus, W. (2012) *"Rationale Divinorum officiorum" Wilgelmi Durandi v russkom perevode kontsa XV veka*. Moscow; Saint Petersburg: Indrik Publ., 261 p. (In Russian)

Sokolov, E. G. (2014) "De moluccis insulis" Maksimiliana Transil'vana v russkom perevode XVI v.: Zadachi i perspektivy lingvisticheskogo issledovaniya ["De Moluccis Insulis" by Maximilianus Transylvanus in 16th century Russian translation: Tasks and prospects of the linguistic study]. *Vestnik Sankt-Peterburgskogo universiteta. Yazyk i literatura — Vestnik of Saint Petersburg University. Language and Literature*, 11 (3): 60–70. (In English)

Tomelleri, V. S. (ed.). (1999) *Die "Pravila gramatichnye", der erste syntaktische Traktat in Rußland*. München: Verlag Otto Sagner, 159 p. (In German)

Tomelleri, V. S. (2004) Il Salterio commentato di Brunone di Würzburg in area slavo-orientale: Fra traduzione e tradizione (con un'appendice di testi). München: Verlag Otto Sagner, XVII, 343 p. (Slavistiche Beiträge. Bd. 430). (In Italian)

Tomelleri, V. S. (2011) Latinskaya traditsiya u vostochnykh slavyan (nekotorye zametki). In: *Aktual'nye problemy filologii: Antichnaya kul'tura i slavyanskij mir*. Minsk: National Institute For Higher Education Publ., pp. 214–221. (In Russian)

Tvorogov, O. V. (ed.). (1972) *Troyanskije skazaniya. Srednevekovye rytsarskie romany o Troyanskoj vojne po russkim rukopisyam XVI–XVII vekov*. Leningrad: Nauka Publ., 232 p. (In Russian)

Tsypkin, D. O. (1990) Skazanije "O Molukitskykh ostrovekh" i Povest' o Loretskoj Bogomateri (Iz sbornika BAN, Arhangel'skoe sobr., D. 193, XVI v.). In: D. S. Likhachev (ed.). *Trudy otdela drevnerusskoj literatury*. Vol. 44. Moscow: Nauka Publ., pp. 378–386. (In Russian)

Wimmer, E. (1990) Die russisch-kirchenslavische Version von Maximilian Transylvans *De Moluccis insulis ... epistola* und ihr Autor. *Zeitschrift für slavische Philologie*, 50 (1): 51–66. (In German)

Wimmer, E. (2005) *Novgorod — ein Tor zum Westen? Die Übersetzungstätigkeit am Hofe des Novgoroder Erzbischofs Gennadij in ihrem historischen Kontext (um 1500)*. Hamburg: Kovac, 229 S. (Hamburger Beiträge zur Geschichte des östlichen Europa. Bd. 13). (In German)

Zwicky, A. M. (1985) Heads. *Journal of Linguistics*, 2 (1): 1–29. DOI: 10.1017/S0022226700010008 (In English)

Author:

Evgenii G. Sokolov, ORCID: 0000-0001-5782-8093, e-mail: pan_liwerij@mail.ru

For citation: Sokolov, E. G. (2019) The project of a deeply tagged parallel corpus of Middle Russian translations from Latin. *Journal of Applied Linguistics and Lexicography*, 1 (2): 337–364. DOI: 10.33910/2687-0215-2019-1-2-337-364

Received 24 August 2019; reviewed 11 September 2019; accepted 12 September 2019.