Russian Language Journal

Volume 71 | Issue 3 Article 9

2021

From Error Annotation to Quantitative Analysis: Patterns in Russian Language Learning

Irina Kor Chahine

Ekaterina Uetova

Follow this and additional works at: https://scholarsarchive.byu.edu/rlj

 Part of the Slavic Languages and Societies Commons

Recommended Citation

Chahine, Irina Kor and Uetova, Ekaterina (2021) "From Error Annotation to Quantitative Analysis: Patterns in Russian Language Learning," *Russian Language Journal*: Vol. 71: Iss. 3, Article 9.
Available at: https://scholarsarchive.byu.edu/rlj/vol71/iss3/9

This Article is brought to you for free and open access by the Journals at BYU ScholarsArchive. It has been accepted for inclusion in Russian Language Journal by an authorized editor of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.

From Error Annotation to Quantitative Analysis: Patterns in Russian Language Learning

Irina Kor Chahine, Ekaterina Uetova

1. Introduction

Although learner corpus research has been progressively growing into an independent branch of corpus linguistics, the learner corpus cannot yet fully benefit from corpus analysis methods. This is due to several technical obstacles involving data collection, error annotation, and finally, data processing. When it comes to data collection, compared to corpus linguistics, learner corpus is biased because some of the learner corpora are still collected manually: Optical character recognition (OCR) is not yet sophisticated enough to transform a student's handwritten copy to a digitized text. This fact significantly slows the collection of learner corpora. Furthermore, typed students' texts present another problem: access to spell-checkers and other proofing tools obscures students' real language skills. Moreover, annotation of the learner corpora presents inherent difficulties: the learner corpus represents a collection of productions in the language, also called *interlanguage*, which deviates from the codified standard language on several linguistic levels (morphologically, syntactically, discursively), and these deviations are not yet taken into account by the processing software. This constitutes one of the challenges of current learner corpus research (Granger et al. 2015). Finally, unannotated learner corpora usually cannot be fully processed by quantitative analysis, as is the case with computerized corpora of standard texts, because of a number of erroneous forms, most of which cannot be yet recognized by the machine. However, it is possible to digitally analyze the annotated data, and this opens new perspectives particularly in the fields of foreign language acquisition and teaching.

This study presents an analysis of the Russian learner corpus, from annotation taxonomy to data processing and interpretation. The purpose of this study is to classify and quantify the data from the Russian Learner Corpus (RLC),[1] as well as to reflect on the associated difficulties and discuss the results of primary data processing.

[1] Open access: http://www.web-corpora.net/RLC/

The study is based on the annotated segment of the French subcorpus of the RLC, collected in 2015-2018. The main objectives of the study are: 1) to identify general trends in the acquisition of Russian linguistic categories in the French-speaking environment and 2) to identify the linguistically "problematic areas" for two groups of learners (FLs and HLs).

The paper is organized as follows: Section 2 presents a general overview of learner corpora processing problems and specifics of error annotation taxonomies. Section 3 is devoted to Russian Learner Corpora website and its error annotation taxonomy. Section 4 describes aspects of the working corpus and collection methods used. Section 5, classifies and analyzes learner errors through five linguistic categories, i.e. spelling, morphology, syntax, lexis and discourse. Section 6 presents general observations and suggests additional lines of research.

2. Learner corpora processing and error annotation taxonomies

The automatic processing of learner corpora is still at the beginning of its development. Although automatic error annotation can be used with learner corpora (Hana et al. 2010, Rosen et al. 2014, Rakhilina et al. 2016), it is possible exclusively for regular forms and labeling parts of speech. Many erroneous items, which are difficult to label automatically, do not allow a faithful reflection of part-of-speech usage. Currently, this represents a challenge for learner corpus research (Rosen et al. 2014, Kutuzov and Kuzmenko 2015). Therefore, the only way to effectively annotate a learner corpus is to do it manually. However, this method raises other problems: many scholars have already pointed out that, in addition to the problem of objectivity of this method, manual annotation is a labor-intensive and time-consuming task (Rosen et al. 2014, Rakhilina et al. 2016, Kisselev and Furniss 2020) that requires additional skills in identification and labeling of erroneous forms. Moreover, consistency is usually lower with manual annotations.

There is a large body of literature devoted to error annotations, and this issue has been discussed in academic papers since the very beginning of learner corpus research (Granger 2003, López 2009, Hana et al. 2010, Rosen et al. 2014, Brunni et al. 2015, Rakhilina et al. 2016, Rozovskaya and Roth 2019). What emerges from the discussions is that error annotations are highly biased by specific research purposes. Furthermore, it is often

difficult to apply the tools designed for a given language to another language, as language structures are different. Nevertheless, it is still possible (and useful) to apply these tools to linguistically close languages (Brunni et al. 2015). Moreover, to be efficient, corpus annotation needs to avoid any theoretical influence and to be more general in tag labeling (Leech 1993, Mathet and Widlöcher 2019).

There are several annotation models used in learner corpus research (see Lüdeling et al. 2005). One of the most currently applied is a multi-layer standoff model, which offers multiple choice of hypotheses for one error and gives the possibility of storing the annotation apart from the text. This design was adopted by recent learner corpora, such as the FALCO corpus of German (Lüdeling et al. 2005), the Czech learner corpus CzeSL (Rozen et al. 2013), Russian Learner Corpus, RLC (Rakhilina et al. 2016), the COPLES2 of Portuguese (del Rio and Mendes 2018).

Additionally, the adopted tag annotation taxonomy varies depending on the corpus research purposes. Some of them have a restricted annotation schema, like the COPLES2 corpus of Portuguese (del Rio and Mendes 2018) with only three linguistic categories: spelling, grammar and lexis. Others are more expansive, like the NOSE corpus of Spanish with its six linguistics categories: spelling, punctuation, word grammar, clause grammar, phrase grammar, and lexis as well as four additional layers comprising an entire tagset of 612 tags (Díaz-Negrillo 2012). Small tagset taxonomies are easy to manage but they don't allow categorisation and description of errors. By contrast, fine-grained tag annotation taxonomies are difficult to structure, and they may contain errors in annotation.

On this point, Rozovskaya and Roth's paper (2019) is particularly interesting for our study. Like our corpus, it is based on RLC tagset, and covers Russian learner corpus of American English-speaking students (RULEC-GEC). It presents an elaborate tagset of 23 items covering "syntactic and morphosyntactic errors, spelling and lexis," but presented tags include more specific tags covering not only general linguistic categories (such as punctuation or spelling) and specific phenomena (like verb:number/gender) but also mechanisms (like replace) involved in errors. As an example of the most frequently occurring errors it presents the following: spelling, noun:case, lexical choice, punctuation, missing word, replace, extra word, adj:case, preposition, word form, noun:number, verb:aspect, etc. Such taxonomy allows to calculate error rate and to

identify some frequent errors. The paper is not devoted to error analysis but there are some questions that arise about error annotation taxonomy. In particular, without more detailed information, some authors' choices remain unclear. For example, do errors in case in nouns, which are most frequent grammatical errors, depend on government (прогул*к*а по город**е** < город*у*), or occur in independent (нет автобус**у** < автобуса) or nominal construction ((в) маске *льв***у** < *льв*а), or whether they are of morphological origin in the choice of correct paradigm (читает журнал**и** < журнал*ы*)? These questions are particularly important if one wants to use the data for language instruction. This kind of tagset taxonomy is not helpful for such purposes.

The main purpose of our taxonomy was direct application of data in the teaching process, and this point of view determined our approach to the tagset design.

3. Error annotation in the RLC corpus

While the RLC website presents an elaborate tagset (Rakhilina et al. 2016), the error annotation process is not sufficiently systematized. With the exception of raw texts that do not have a linguistic annotation of errors, most texts contain what we can call *non-systematic annotations*. By "non-systematic annotations" we mean labeling in a non-systematic manner, when tags are not given in an orderly way. For example, the tag "Verb – Ortho – Inflexion – Morph – Miss – Lex" for the same erroneous lexeme, would be placing errors from different linguistic categories and at multiple linguistic levels in the same tag window. This makes automatic processing of such data problematic. Nevertheless, such tag labeling makes it possible to look for a certain type of errors such as an erroneous verbal form or morphological errors.

In our work we adopted position-based tags already used in other corpora (see del Rio and Mendes 2018) which we believe to be more convenient for cross-sectional studies. As far as we know, cross-sectional studies on learner errors in relation to this corpus have not been performed. The entire error annotation process comprises three steps: first, manual labeling of errors in position-based order; second, automatic processing of manual annotations and generating of Excel tables of classified errors; and, third, checking the tables and adding more detailed error labels for the fine-grained description of errors.

The first step of labeling consisted in manual annotation of errors. In our annotation system, the top-level is represented by a linguistic category. The RLC website already subdivided all tags by general categories to which we added the "discourse" label. So, the top-level tagset included five linguistic categories, namely, spelling, morphology, syntax, lexis and discourse.[2] Each linguistic category label could then be followed by additional tag(s), relevant for each category. Second-level, and, possibly, third-level labeling comprised more specific annotation, such as linguistic mechanisms for spelling (substitution, insertion, etc.). These additional annotation levels allowed more detailed classification within each category and facilitated automatic processing of errors. However, since second-level labeling design is still in progress for all categories, we will not discuss it in this paper and focus only on the top-level, since it already yields interesting results.

For automatic processing purposes, it was important that each type of error was labeled in a specific tag window: the spelling errors appeared separately from morphological or lexical errors, and so on. If a lexeme or an erroneous segment had more than one error, it was labeled by several tags. Like most recent learner corpora (Lüdeling et al. 2005, Rozen et al. 2013, del Rio and Mendes 2018), the RLC system offers multiple choices of categories for one error and contains a simple function to add tags by compounding them.[3] For example, in "фонтан безконечно работает B2, FL", *безконечно* was tagged by three tags "Ortho – Subst / Morph – Altern / Lex – Subst": i) "Orpho – Subst" was used for possible substitution з/с which are not clearly differentiated in pronunciation, ii) "Morph – Altern" for possible ignorance of morphological alternation of voiced/voiceless consonants in a word derivation, and iii) "Lex – Subst" for erroneous lexical choice because the adverb *непрерывно* is preferable in this context. In this case, the identification of the linguistic category was somewhat ambiguous (is it a spelling or a morphological error?), and the double error labeling (Ortho and Morph) was counted twice.

After the first step of manual annotation by the linguistic category and additional classification of errors if necessary, the second

[2] Punctuation is one of the problematic areas of language learning, but punctuation errors were not included in our study, nor are they included in the RLC tagset. Usually, there is no special punctuation course in Russian programs in France.

[3] About the marking process in the RLC website, see Rakhilina et al. (2016).

step consisted of automatic processing of data:[4] the tags were generated automatically into Excel tables. All tables used in this paper are available on Google Drive.[5]

Then, at the third step of data processing the task consisted in checking the tables and adding more detailed error labeling (second- and third-level) to complete error classification. This classification was made following the guidelines which have been developed (and are still under review) for each top-level category on the basis of erroneous linguistic phenomena. When structured error annotation design could not rely on previous research on Russian data, this step was executed manually. We intend to achieve a second-level taxonomy and to edit the final guidelines in the upcoming works.

4. Data and methods

This study is based on written works produced by university students from Nice, Lyon and Sorbonne University between 2015 and 2018. The working corpus includes 191 students (142 foreign learners and 49 heritage learners,[6] see Table 1 below) aged 17 to 26.

Table 1 shows that the analyzed corpus is unequally distributed with predominant levels of A2 (38.22%) and B1 (17.80%), which represent more than half of the corpus (56.02%) and make up the bulk of students studying Russian in France. In addition, data for certain levels are relatively scarce. We are aware that the number of B1 informants in the heritage language, 1.57% (3 informants), is too low for significance testing, and they do not represent a robust sample. However, using a descriptive statistical approach, the data are intended to be purely informative and allow for the facts to be observed and described. Moreover, as the percentage of errors for each level is determined by

[4] The RLC website automatically subdivides all texts into sentences which facilitates annotation and checking since each sentence is followed by two corrected versions: the first one avoids spelling mistakes and the second one shows a modified version according to annotators' suggestions. All texts with annotations can be downloaded into Excel tables including key information, such as text number, original sentence containing errors, tagset for the error reflecting its nature (spelling, morphological and so on), erroneous and corrected items, additional comments, and additional data (proficiency level, experiment group, informant's name, etc.).

[5] www.shorturl.at/cpAS7

[6] The term heritage learners refers to speakers who are fluent in two languages at the same time, with one being reserved for the family environment and the other being used in a linguistic environment outside the family (study, work, social life).

the number of errors relative to the number of words, the error ratio remains the same regardless of the size of the group.[7]

Table 1. Number of participants according to their level and group

	A1	A2	B1	B2	C1	C2	Total students
Foreign Learners	17 *8.90%*	73 *38.22%*	34 *17.80%*	11 *5.76%*	7 *3.66%*	—	142 *74.35%*
Heritage Learners	—	—	3 *1.57%*	6 *3.14%*	17 *8.90%*	23 *12.04%*	49 *25.65%*
Total	17 *8.90%*	73 *38.22%*	37 *19.37%*	17 *8.90%*	24 *12.57%*	23 *12.04%*	**191** *100%*

In addition, our corpus includes metadata containing background information (age, gender, L1, language(s) spoken at home, time spent living in France or in a Russian-speaking country) and L2 acquisition details (university of study, course, second and foreign languages, self-rated proficiency). Once again, the collected metadata show an unbalanced distribution, particularly by gender. Due to the demographics of university-level language studies in France, our corpus contains data from three times as many female students as male students (76.41% women versus 23.56% men for the analyzed corpus), as shown below.

Moreover, the French corpus is a Multi-L1 corpus reflecting the demographics of the French society, which is especially obvious at the University of Nice. Thus, our participants included native French learners but also students from various Slavic countries (Russia, Ukraine, Belarus, Bulgaria, Poland, Croatia), students from the Romance language areas (Italy, Romania), and others (native Chechens, Armenians, Hungarians). These nonnative French speakers were, however, mostly raised in France or spent several years in a French-speaking environment; thus, French was their dominant language. The most common foreign languages already

[7] It is understandable that such a small sample cannot be generalized with the same confidence as a large or diverse sample can, and these results should be checked on a larger sample.

spoken by study participants were English, Spanish and Italian, which are also the most studied languages in the French educational system.

Table 2. Level groups by gender

	Foreign learners (% of the FLs data)				**Heritage learners (% of HLs data)**			
Gender / Level	female		male		female		male	
A1	11	(5.76%)	6	(3.14%)	—		—	
A2	61	(31.94%)	12	(6.28%)	—		—	
B1	25	(13.09%)	9	(4.71%)	2	(1.05%)	1	(0.52%)
B2	8	(4.19%)	3	(1.57%)	4	(2.09%)	2	(1.05%)
C1	7	(3.66%)	0		9	(4.71%)	8	(4.19%)
C2	—		—		19	(9.95%)	4	(2.09%)
Total students by gender (% of the total)	112 (58.61%)		30 (15.71%)		34 (17.80%)		15 (7.85%)	

Another aspect of the study concerns language testing. The student's language proficiency level was determined by the students' language instructors and in accordance with the participants' self-assessment. Most of the participants were identified by their first name with their permission (or, rarely, by a nickname). This identification method was advantageous, since the knowledge of the learners dominant language or L1 would help the annotator who is familiar with them and is able to guess the students' intentions. However, we are aware that this could also be seen as a flaw in the annotation process, since objectivity and privacy are lost.

Thus, the working corpus includes work at all language proficiency levels, from beginners (A1 level) to the highest Russian proficiency level in the CEFR, a near-native C1 and a native C2. In addition, the corpus includes written productions of two groups of students: *foreign learners* (FLs) and *heritage French-Russian learners* (HLs). The analyzed annotated

data comprised more than 42 000 words. Details concerning the corpus are reported in Table 3.

Table 3. Annotated corpus (token counts) in RLC website according to French students' level and group

Language background	Language level	Ratio	Number of words	Number of texts	Average number of words per text	Standard deviation in number of words per text
Foreign Learners	A1	5.74%	2416	22	109.82	87.16
	A2	27.74%	11673	103	113.33	77.55
	B1	16.24%	6836	45	151.91	65.19
	B2	6.00%	2527	16	157.94	54.45
	C1	4.13%	1740	9	193.33	87.92
FL Total		59.86%	25192	195	129.19	78.53
Heritage Learners	B1	0.44%	187	3	62.33	31.48
	B2	2.30%	966	6	161	58.21
	C1	12.84%	5402	34	158.88	78.75
	C2	24.56%	10336	43	240.37	166.24
HL Total		40.14%	16891	86	196.41	137.03
TOTAL		100.00%	**42083**	281	149.76	104.81

The data was collected by manual typing from handwritten sources submitted by students during Russian L2 training, from 2015 to 2018. The written works included students' essays, biographies, summaries, and occasional translations from French; some of them were written during timed exam sessions, while others were written at home.

Once the data were collected and ordered by language proficiency level, the second step was to annotate them. The text annotations were carried out by at least two annotators. However, double annotation in the RLC website was problematic: it could not allow simultaneous labeling by different annotators. The second annotator could see the edited labels and was able to make changes to the labeling, by erasing previous labels. Therefore, the annotation process was organized as follows: the markup assistant was responsible for detecting and marking errors in a raw document, the second (if there was one) made its own annotations, and then the referring annotator (authors of the paper) checked and corrected the annotations if necessary.

5. Linguistic categories and quantitative analysis

In this section, we describe linguistic phenomena found in five linguistic categories, i.e., spelling, morphology, syntax, lexis and discourse. Before discussing the results of error analysis, this general overview (see Table 4, p. 50) presents error distribution by students' group and level for each linguistic category.

5.1. Spelling errors

This is the only linguistic category that is automatically detected by the program, since the part-of-speech annotation with spelling entries is applied in the RLC. The nonnormative items are already highlighted in the raw corpus. However, not all nonnormative items should be considered spelling errors. Some errors are obviously morphological (like Арабые A2 -> Араб**ск**ие with a missing suffix in derivation), and others, involving word usage, are lexical (like пиано A2 -> пиа**нин**о as a case of direct transfer from French *piano*). Thus, the category of spelling errors is limited to errors that do not fit into any other category of linguistic development and follow four main patterns (see below).

Inspired by the RLC tagset, spelling errors are classed by four mechanisms: substitution (е**шё** B1 FL > ещё), insertion of extra letters (дол**ь**го A2 FL), omission of letters (станц(**и**)ю A1 FL), and transposition of letters (с**тра**ше C2 HL > с**тар**ше). We also mark the abusive use of Latin graphemes (**В**илет B1 FL) as a subgroup of substitution: they reveal cognitive mechanisms in acquisition. Typographical errors involving hyphenation (когда(**-**)то, кого(**-**)то C1 HL) or word or nonword spacing

(На_конец A2 FL > Наконец) were of lesser interest to us, as they did not disturb word meaning; they also represented a very small ratio of the overall errors.

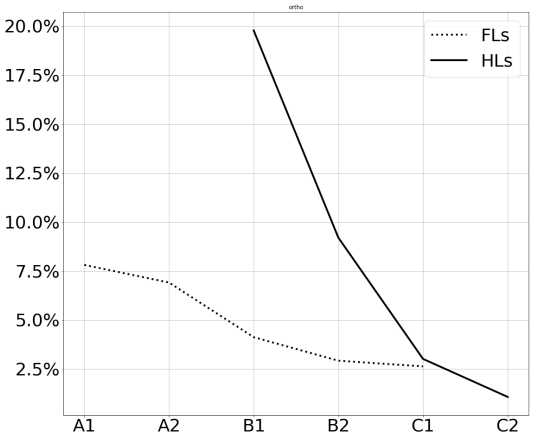

Figure 1. Progression of Spelling Errors in FLs (bold line) and HLs (dotted line)

Spelling errors characterize any written work. Figure 1 shows the progression of spelling errors made by our groups (FLs and HLs) according to the CEFR proficiency levels (from the beginners A1 and FLs to the Russian native speakers C2). In the Figure 1 here and in the Figures below, the x-axis indicates the language level of the students (from A1 to C2), and the y-axis shows a ratio between the number of errors and the number of words at each level.

The two descending curves in Figure 1 represent the gradual decrease in error number proportional to the students' increase in language proficiency, and this tendency is observed for both FLs and HLs.[8] However, the two categories of learners master spelling in different ways. For FLs, a relatively large number of errors remains stable at the two initial levels (A) and then slowly decreases until they are nearly equal at the last two levels (B2 and C1). For HLs, the problem of spelling is the greatest difficulty at the initial B1 level: nearly 2.5 times as many errors

[8] This is likely to be true for native speakers as well. However, we are not aware of any study of this kind. The learner corpus of Russian L1 CoRST (http://web-corpora.net/learner_corpus/) with its annotated corpus of 1.06 mln tokens, could be used for such a comparison.

as the FLs at the beginning levels and 4 times as many errors as at the FLs B1 level. At subsequent levels, the number of errors produced by HLs dropped sharply, and much more rapidly (especially between the first and the second level of proficiency) than for FLs (see Figure 1). Their stronger linguistic background in Russian oral skills can probably explain this fact. Therefore, our foreign students needed three levels to master Russian spelling, while our heritage learners required only two: the number of errors for both groups becomes approximately equal at the C1 level.

Table 4: Error distribution by students' group and level for each linguistic category

Language background	Language level	Students' errors							
		Spelling	Syntax	Morphology	Lexis	Discourse	Total number of errors	Complex errors	Erroneous items[9]
Foreign Learners	A1	189	329	112	102	55	787	128	659
	A2	808	1109	403	294	159	2773	477	2296
	B1	282	576	122	212	86	1278	129	1149
	B2	74	144	27	37	24	306	29	277
	C1	46	76	16	25	20	183	14	169
FL Total		1399	2234	680	670	344	5327	778	4549
Heritage Learners	B1	37	13	4	3	2	59	10	49
	B2	89	27	14	14	8	152	26	126
	C1	163	189	39	62	41	494	60	434
	C2	112	142	22	100	69	445	25	420
HL Total		401	371	79	179	120	1150	122	1028
TOTAL		1800	2605	759	849	464	**6477**	900	5577

[9] Total number of errors without complex errors.

A deeper analysis of spelling errors revealed that the most frequent spelling error involved the substitution of letters (47.5% of all spelling errors), and 70% of substitutions involved vowels (mostly between *а* and *о*). The substitution of Cyrillic letters by Latin graphemes is also quite frequent (4.3%), as it outnumbers the errors in transposition (3.5%), which are a subgroup apart. Observed only during the first three levels of language learning, the substitution of Latin graphemes is influenced by "the writing memory" of the already mastered writing system, of French in our case (**I**ван! A1 FL, **В**илет B1 FL, **uo**кавото B2 HL > **у** кого-то). The FLs also made an important number of mistakes in missing letters (фил(**ь**)м B1 FL, прие(**з**)жает B1 FL, рус(**с**)кого B2 FL).

The factors that influence spelling errors may be of two types: contextual and noncontextual. The contextual (or syntagmatic) factors mainly concern transposition errors, where two inverted letters are situated nearby (пр**ыби**ли A2 FL > прибыли, в**тс**ретила B1 FL > встретила, пр**ова**славный B2 HL > православный). The noncontextual (or paradigmatic) factors have various origins, i.e., cognitive, intra- and interlinguistic, and extralinguistic. They mostly lead to omission errors. Thus, in cases like воз(**в**)ращаюсь A2 FL, Трина(**д**)цать B1 FL, Чю(**в**)ствовать C1 HL, the omission is motivated by peculiarities of pronunciation (assimilation or devoicing) and therefore by the sound perception of learners; the missing consonants are less audible for a non-Russian speaker (also лес(**т**)нице A1 FL, быстр**а** A2 FL). Cases of the substitution of *а* by *о* may be the result of a lack of attention or of "hypercorrection", i.e., a self-correction of the frequent item (here a letter) in a wrong position (пасс**о**жиры B1 FL, сначал**о** C1 HL). However, contextual and noncontextual factors are complementary, and this is often the case in most errors of substitution and insertion.

5.2. Morphological errors

Errors that involve the morphemic structure of an item or its inherent morphological features were considered morphological errors. During the annotation process, two main morphological aspects were identified as most problematic, namely morphological mechanisms (alternation and derivation) and morphological features (gender, number).

Morphological mechanisms represent almost 90% of all morphological errors, according to our data. We considered alternation

as a contextual phenomenon where the choice of a correct form depends on the left-hand (nature of the ending phoneme) or the right-hand (nature of the initial phoneme) context. This group was primarily divided into *strictly alternation errors* and *errors in inflectional endings*.

In the **strictly alternation subgroup** (26% of the category), we deal with the alternations occurring in roots, which vary in nature. The alternation of this type occurs mostly in verbal roots (50% of errors). They concern palatalization patterns, such as *т* / *ч*, *с* / *ш*, *ск* / *щ* (попи**с**ут A1 FL > подпи**ш**ут, подпи**ш**али A2 FL > подпи**с**али, хо**т**ет B1 FL > хо**ч**ет) but include other cases of alternations occurring in verbal roots (м**ы**ут A1 FL > м**о**ют, б**ра**ет B2 FL > б**ер**ёт, зак**о**нчивают B2 FL > зак**а**нчивают). Even if the previous cases may occur in nouns as well, errors of this kind are rare in nouns. Most cases in nouns affect epenthetic vowels (рын**о**ке A2 FL > рынке, за**я**йца B1 FL > зайца). Other cases with irregular nouns, such as дру**га**ми A2 FL > дру**зья**ми, дере**ва**ми B2 FL > дере**вья**ми, must be mentioned.

Other alternations involve affixes. In addition to the cases of inflectional affixes shown below, we mostly find verbal suffix alternations, such as *-ова-/-у-*, *-ну-/ –*, etc. (рис**ова**ет A2 FL > рис**у**ет, достиг**ни**ли A1 FL > достигли) or postfixes with *-ся* / *-сь* alternation (встречала**ся** A1 FL > встречала**сь**, одеваю**ся** A2 FL > одеваю**сь**) but also find errors in prefixes (под**о**бежала A2 FL > подбежала, без**к**онечно B2 FL > бе**с**конечно).

Finally, there are some errors related to the sandhi phenomenon. Errors of this kind usually occur with prepositions (25% of errors) involving a misuse of epenthetic vowels with *в, с, к*, etc. (в**о** парке A1 FL > в парке, в(**о**) Францию A2 FL) or with third-person pronouns where an epenthetic *н* is missing or wrongly inserted (у (**н**)их есть A2 FL, старше **н**ее B1 FL > старше её).

As Russian endings appeared to be the main source of difficulties for non-Russian learners, appearing in approximately one-third of all morphological errors, we chose to classify them into a particular subgroup: **errors in inflectional endings**. The quantitative analysis shows the following error distribution by part of speech: inflection errors occur mostly in nouns (56%), the ratio of verbal and adjective errors is 25% and 17%, respectively, and the remaining 2% involve pronouns and numerals. The alternation errors are found in nominal and adjective

inflections, where the inappropriate form of the flexional alternation was chosen: украинц**и** C1 FL > украинц**ы**, так много волос**ов** A1 FL > так много волос, *люд***ов** B2 FL > *люд***ей**; родительск**его** B1 FL > родительск**ого**, *младш***ом** B2 FL > *младш***им**. Thus, according to Russian phonological and spelling norms, the ending *и* after *ц* (instead of *ы*) is due to the frequent confusion of *ц* with a hushing consonant (which implies such a choice), and a flexional *е* does not appear after a velar consonant (*родительского vs. среднего*) nor does an unstressed *о* appear after a hushing consonant (*cf. старшего*).

It is important to emphasize here that the errors of this kind concern only "obviously correct" forms on the syntagmatic level, such as the choice of regular plural genitive endings for nouns (between *-ов/-ей/ zero* flexion for the errors above) or the 1st and 2nd plural inflections of the verb *(-ем/-им; -ете/-ите*): хоч**ем** B1 FL > хот**им**, увид**ете** C1 HL > увид**ите**. When the choice is wrong on the paradigmatic level (i.e., a genitive inflection morpheme instead of a dative morpheme), the problems are not morphological but syntactic. However, if the error cannot be explained by alternation mechanisms (on the syntagmatic or paradigmatic level), we are dealing with derivational instances.

Another set of errors concerns **derivational mechanisms** includes various phenomena. It may appear in the cases of "paradigmatic intruders" when a morpheme combination does not belong to the word paradigm, while the morphemes are correct independently. This is the group of word form creations in which the inflectional (in most cases) morpheme is chosen from another paradigm. Thus, in forms such as письм**ы** A1 FL and человек**и** A2 FL, linguistic features of the items are not respected: if the nominal inflection *-и* is used to mark plural, it is used here with nonrespect to the morphological gender of a noun (the plural neuter implies *-а*, *письма*) and to its suppletive plural form (*люди*). The same problem occurs in the following examples: ног**ими** A1 FL (noun with adjectival flection), по выходн**а**м A2 FL (adjective with noun flexion), плак**и**ть A1 FL (confusion of verb derivational suffix), Толсто**йа** B1 FL > Толстого (adjective declension confused with a nominal declension for Russian last names), *л*учшее A1 FL > *л*учше (inappropriate comparative suffix for this suppletive form), езд**и**ет B2 HL > ездит (fusion of two paradigms: infinitive basis of *<езди>ть (ездит)* and the third-person inflection of *ехать (ед<ет>)*).

In addition to this case of regular morphemic items, another group is represented by word creations that disturb the morphemic entity of the word. We consider here examples such as отношен**ов** A2 FL > отношений, французс**к**ов A2 FL > французов, климат**н**ой C1 HL > климатической, which contain inappropriate or missing morphemes. Word creation of this kind usually indicates a lack of mastering derivation mechanisms: in отношен**ов**, the ending *-ий* of *отношений* is wrongly interpreted as an inflection (not as a part of the root), in французс**к**ов, the plural genitive noun is derived from an adjective (<*французск*>*ий*), and климат**н**ой does not use the appropriate derivational adjective suffix. However, all such errors cannot be so easily explained, and word creations like остат**не**ваться A2 FL > оставаться, пасу**ют** A2 FL > пасутся are usually not clear without a large context. Most of them are apparently conditioned by a cognitive ability of individual memory (phonetic or written memory of words). Nevertheless, some particular but rare creations are remarkably good and worth mentioning as well: лыжит A1 FL > едет/катается на лыжах, добрость A1 FL > доброта. Finally, errors that imply categorical change (like the use of an adjective *русский* instead of an adverb *по-русски*) are considered lexical errors (see below).

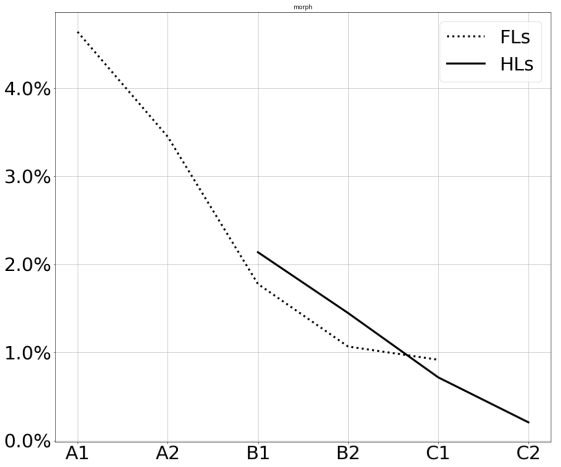

Figure 2. Progression of Morphological Errors of FLs (bold line) and HLs (dotted line)

Another group of learner errors illustrates a problematic use of **morphological features**. The small number of errors of this kind (approximately 10% of the category) does not mean that they do not represent a source of difficulties for our participants. Rather, the small number is explained by our methodological choices: only inherent morphological features of nouns are taken into account here, and the verbal aspect, due to its borderline nature putting it between morphology and syntax, is considered among syntactic problems. Thus, erroneous interpretation of gender in nouns and erroneous use of number in *singularia* or *pluralia tantum* remain relatively rare. The errors in gender are not be considered inflectional errors and their identification is mostly possible through a larger context (particularly, thanks to adjective agreement): for instance, Россия <многонациональный> государств(**о**). C1 HL. Errors in gender are obviously influenced by cognitive and/or interlinguistic factors; however, their interpretation is debatable. Thus, in К ним подбежает собак(**а**) Шарик. A1, FL, the noun *собака* is missing the flection, like most masculine nouns, visibly by association with the semantic genre of the noun that refers to a male dog Sharik, but it can also be explained by the influence of its French masculine counterpart (*le chien*). On the other side, words like температур(**а**) A2 FL, гитар(**а**) A2 FL with the same missing ending were not interpreted as morphological errors in gender (their French counterparts are feminine: *la temperature, une guitare* and the context does not suggest any information about their gender);[10] so, errors like this have to be considered as lexical errors by loan translation from French.

As for errors in number, they are usually influenced by a semantic factor: the inappropriate plural forms are prevalent and correspond to collective nouns: (*люблю* есть) рыб**ы** A1 FL > рыбу (*(j'aime) le poisson*, singular in French). Once again, the interlinguistic influence may be strong: in French, the plural is often required in such a case: покупать одежд**и** A2 FL, картофел**ы** A2 FL (e.g., *acheter les vêtements, les pommes de terre*). In addition to these cases, errors in number can also be of a syntactic (agreement in number for adjectives: красив**ый** <фотки>. A1 FL) and discursive nature (the choice of plural for generalization, for instance)[11] and will be treated in the corresponding sections.

[10] (…) температур(**а**) тепло и не обычно облачно. (Marine A.-C., A2, FL)

[11] Я *люблю* смотреть экстремальные виды спорта, новосты и теле сериал(**ы**). (Petya, B1 FL)

Therefore, the use of morphological features is directly linked to extralinguistic factors such as cognitive ability of individual's memory and interlinguistic influence. These morphological issues require special attention at all stages of learning, both for foreign and heritage students.

Morphological errors may also be viewed from the perspective of language acquisition. Morphology is the first grammatical domain in Russian language training in the university educational system. Syntax, although present at the introductory level, is reduced to some basic constructions. Thus, it is predictable to find a higher number of morphological errors at the introductory levels. Indeed, quantitative data based on the error ratio in our corpus at each proficiency level of Russian shows a high ratio of errors at the beginning A1 level and its steady decrease from level to level indicating gradual mastery of morphology with language training (see Figure 2). This tendency is characteristic for both groups. Thus, while the number of morphological errors made by FLs is very high at the initial A levels, their number declines gradually through B2 level. The difference in the ratio of morphological errors made by FLs and by HLs becomes small at the high intermediate level. Overall, in regard to both morphology and spelling, our two groups (FLs and HLs) are situated equally at the B2 level. As the data show, the first four levels of training, which usually correspond to the number of years of training, are fundamental for mastering Russian morphology; with the assimilation of morphological forms, the number of errors is gradually reduced. However, it is remarkable that at the B2 and C1 levels, the situation becomes stable. It turns out that FLs, who are fluent in Russian, continue to make certain morphological mistakes and do not reach, according to our data, the level of morphological mastery characteristic of native speakers C2, who themselves still make a certain number of errors.

The factors influencing morphological errors have yet to be fully determined, but this preliminary view shows that both intralinguistic (determined by other Russian forms) and interlinguistic (i.e., linguistic transfer, motivated by French, English or other L1) factors are strong influences.

5.3. Syntactic errors

Syntactic errors violate the rules of word combination. Since syntactic errors are varied in nature, they will have to be analyzed in greater

detail in a future study. We propose here only an overview of syntactic points included in this category and discuss specific problems related to the annotation choices. For efficient data processing, we divided all syntactic errors into three groups, grouping them by their proximity. Thus, the first group contains all agreement errors, where case, gender, number or person agreement is not respected. The second group involves errors occurring in syntactic constructions. Finally, the third section includes other syntactic errors covering mainly the argument structure of particular items and parts-of-speech syntax. **Agreement errors** may affect any variable part of speech: nouns, adjectives, verbs, pronouns, and numerals. According to their nature, they may vary on four morphological parameters: on case (Ex. 1), gender (Ex. 2), number (Ex. 3) and person (Ex. 4):[12]

1) Французский народ и его культуру протеста. (Fr2, F, A2, FL) > культур**а** (A / N)
2) <Конференция> состоя**л**ся на прошлой неделе в Париже. (Emilie C., A2, FL) > состоя**ла**сь (m / f)
3) Арабск**ий** <страны> готовы инвестировать 12 миллион долларов к проекту. (Ed, A2, FL) > арабск**ие** (m sg / pl)
4) <Моя подруга> очень люб**лю** читать. (Amandine M., B1, FL) > люб**ит** (1sg / 3sg)

There are no specific difficulties in labeling errors of this kind except for the subgroup of case agreement. Indeed, the RLC tagset presents three tags that are very similar and particularly difficult to distinguish at first sight. They are a "Case agreement", "Government" and "Constructions", which are not exclusive, particularly, "Constructions". What is the appropriate syntactic tag to be used for a sentence like "Город Москва нахоится на европейск**ий** част**ь** росси**й**. (Marielle, A2, FL)"? One can say that there is an error in a verbal construction since *европейская часть* does not respect the case implied by the prepositional government after the verb. From a semantic point of view, it might be so, but we decided not to include cases like this in the construction phenomena group and reserve the term "construction" for specific patterns (see below). Therefore, in the example above, the error had to be classified as an agreement or a government phenomenon. We classified *част**ь*** as a government error

[12] Due to lack of space, we will not detail the parameters of each morphological class and refer the reader to any Russian grammar.

implied by a verbal prepositional government (*находиться на* + L) with a nonrespect of implied case (nominative instead of locative).[13] However, the adjective *европейск**ий*** is an example of an agreement error, as it does not agree with the associated noun in gender (but it is in agreement in case, which is assumed to be inanimate accusative by the student) and has to be treated separately. As for *России (росс**ий**)*, we believe that for such a case, the error probably comes from spelling, under effect of the left-hand context (otherwise, it is a government error since the genitive is required by its function as a nominal complement). Therefore, a government error is an error that always implies hierarchical dependency between the main word and a subordinate word(s), while an agreement error appears in an equal relationship of word compatibility. That is why errors in case agreement are specifically errors of adjectival or participial case agreement (Ex. 5) and of a subject marking (Ex. 6):

5) Много студентов и ректоров сожалели об этом, потому что это перидаёт плохую картину об итальянск**ие** университетах (…) (Chiara, B2, FL) > итальянск**их** (N=A / L)
6) У него есть жену. (Chloé, A1, FL) > жена (A / N)

The subject postposition (Ex. 6) presents an additional difficulty in case marking, as the postposition is usually associated with an object position (most likely, by transfer effect): many erroneous examples that use accusative instead of the nominative case testify to this fact. The errors of this kind occur in the HLs' productions as well (На столе стоит черн**ую** лампу. Ruslan, B1, HL).

The subgroup of constructions covers various syntactic patterns, such as comparative (Ex. 7), impersonal and related to it constructions (Ex. 8, 9), negative constructions (Ex. 10) and other constructional errors (like Ex.11 with missing subject), but the latter are rare:

7) У нее брат [старше ее (**на**) девять лет]. (Mathez, B1, FL)[14]
8) [Люд**и** без квалификации есть много], им трудно найти хорошую работу на бирже труда. (Arlo, C1, HL) > люд**ей** (N / G)
9) [Вам нужн**ы** более двадцать четыре часа], чтобы приехать в Новой Зеландии. (Caroline, B1, FL)

[13] We choose not to differentiate the terms "locative" and "prepositional" and use only "locative" for both cases. From the other side, the nominative has to be considered as a mixed case of "Nominative/Accusative" since *часть* is a homonymous form for both of these grammatical cases.

[14] The borders of construction are marked here by [...].

10) *Дети* [(**не**) *знают ничего*] *о мире*. (Fr1 (M), A2, FL)
11) *Я купался в море*, [(**оно**) *было теплым*]. (Marion, A1, FL)
12) *Конференция состоялась* [**в** *прошлой неделе*] *в Париже*. (Laurie, A2, FL)

In this section, we also take into account errors in prepositional constructions viewed as independent prepositional phrases (marking time, space, purpose) (Ex. 12), which have to be distinguished from errors in government (see below). The ratio of errors in prepositional constructions is very high, since a lot of attention is given to them in a typical training course and because prepositions are among the most used items in Russian[15] and mastering prepositions in L2 is usually difficult. A preliminary analysis of data reveals that a high number of errors at A1 level are due to omission of prepositions (especially for *в-* and *на*-constructions), but at A2-B1 levels mistakes in prepositions are more frequent (particularly in the same constructions with *в* and *на*). For a more detailed presentation of prepositional constructions in the French learner corpus see Kor Chahine, Perova-Nouvelot, and Uetova 2019.

The **last subgroup** presents the remaining syntactic issues, which can be divided into two sections: argument structure problems and a parts-of-speech syntax. The preliminary results show that the verbal argument structure (verbal government) is the most problematic point for our learners (Ex. 13), along with the usage of verbal categories (aspect, tense, mood) (Ex. 14, 15):

13) *Я увлекаюсь спортом, музыку*. (Bogan, B2, FL) > музыкой (A / I)
14) *Обышно я* **опоздаю**. (Djaïa, A2, FL) > опаздываю (PF / IPF)
15) *Дядя Федор* **решить** *уйди искать клад*. (Cosme, A1, FL) > *решил* (inf / pst)

The most frequent errors in this subgroup concern government (76%). Interestingly, the number of government errors does not decrease between the A1-B2 levels. This peculiarity can be explained by the fact that the study of items and their government patterns is an arduous process, since each verb must be memorized separately due to its specificity, in contrast, for example, to mood and passive voice for verbs, the material of which is more grammatical and can be summarized in general rules.

[15] Cf. Dictionary of frequencies 2009 by Olga Ljaševskaja and Sergej Šarov; open acces: http://dict.ruslang.ru/freq.php?.

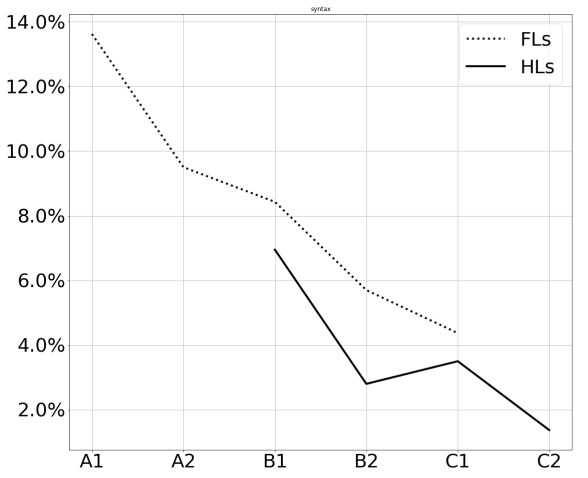

Figure 3. Progression of Syntax Errors in FLs (bold line) and HLs (dotted line)

Other syntactic errors deal with the use of auxiliary verbs, the choice of attribute case markings for nouns or the use of long / short forms for adjectives and participles; other syntactic errors can be labeled with additional tags and be associated with the aforementioned errors:

16) Ты <будешь **быть** счастливее> чем все животные! (Fr1, M, A2, FL) > будешь счастливее (Aux extra)

17) <Любовь очень опасн**ая**>. (Fr1, M, A2, FL) > опасна (LF / SF)

18) Отец <был преподават**ель**> в университете. (Camille, B1, FL) > преподавателем (N / I)

This overview of syntactic issues that cause errors among learners can also be supplemented by other syntactic points discussed in Uetova et al. (2019). As for the quantitative data including all syntactic errors, its general picture is presented in Figure 3. According to the descending curves and a high error ratio at the initial levels, A1 for FLs and B1 for HLs, a new syntactic system of Russian apparently destabilizes students' syntactic habits in some way. However, the "grammatical shock" does not last more than one level, and at the next step, the A2 and B2 levels for FLs and HLs, respectively, the error ratio becomes almost stable, before decreasing until the C1 level for FLs, where the syntactic error ratios of both groups are very similar. A slight increase in the number of errors

at the C1 level in HLs may be explained by learners' confidence in using more diversified linguistic constructions, but a qualitative analysis of syntactic errors is needed to confirm this hypothesis. Thus, as for the morphological category, at the end of their training, FLs and HLs master syntactic questions to almost the same level, but they continue making some syntactic errors that prevent them from approaching the C2 level of native speakers. This observation seems to be a general tendency for our participants who, for the most part, continued academic training and did not leave the educational system.

5.4. Lexical errors and errors in discourse

Lexical errors and errors in discourse follow similar trends and can be discussed together. They also present a slightly different picture in error progression compared to the previously mentioned purely grammatical parameters, as we will see below. However, these categories concern different linguistic questions, and they are annotated separately for both data systematization and more sufficient automatic processing.

Lexical issues are usually related to semantic questions. However, in the case of foreign language acquisition, the area involving **lexical errors** is broader and exceeds the limits of semantics. Some errors reveal a derivational mechanism of word creation, such as direct loans from French (or other languages, L1 or L2) (like температур(**а**) A2 FL, гитар(**а**) A2 FL, seen above), which do not conform to Russian linguistic norms, or cases of code switching, namely, for proper names (в городе **Annecy** в Альпах A2 FL > Аннеси), but this mechanism does not truly represent a linguistic error, as it is currently well represented in Russian media. Other errors are lexical calques, i.e., loan translations (бо*л*ьшие **окно двери** A2 FL > французские окна, from fr. *portes-fenêtres*). However, errors of this kind are rare. Most of the errors (almost three quarters) involve a substitution of words belonging to Russian vocabulary. Such errors occur in any part of speech[16] (nouns, verbs, adjectives, adverbs, conjunctions and numerals). Nonetheless, lexical errors of this kind involve not only semantics (очень **много** меняется C1 HL > быстро, from fr. *change*

[16] As prepositions are always related to singular patterns (as a prepositional phrase or depending on a verb or a noun), they are described in the syntax category. As for interjections, they are extremely rare in written texts.

beaucoup) but also grammatical features of words: in many examples the semantically close lexemes are grammatically confused (Меня зобут Жиль и я **французскый**. A2 FL > француз, from fr. *je suis Français* (Noun / Adj)). The preliminary analysis of data confirms the previous statement that lexical substitutions have extralinguistic origins involving linguistic transfer.

Thus, the lexical category comprises various errors, namely, errors in word substitution as mentioned above, errors in conjunctions and reflexive verbs, and usage of erroneous part-of-speech forms and idioms:

19) Дядя Федор читает что-то **и** кот ест колбасу с молоком. (Cosme, A1, FL) > а
20) Когда мы устаем мы идем в ретсоран и пить вино, чтобы с**ебя разогреть**. (Amandine M., B1, FL) > согреться
21) Позже, когда он стал **старший**, она всегда садилась в кресле и он приносил не подушку, а цветы. (Manon, A2, FL) > старше
22) Однако самый север страны находится за **полярным поясом**, там где лет никогда не тает. (3730, C1, HL) > полярным кругом

It is not surprising that students make most errors (half of the errors) with verbs, since, in addition to the semantics of simple verbs, Russian verbs can differ by their prefixes *(перейти-пройти, уходить-выходить)*, which determine their meaning. Lexical errors are also frequent in nouns, adjectives and adverbs. Erroneous substitution of pronouns occurs rarely. As for the misuse of prepositions, such errors should not be attributed to vocabulary, but rather to syntactic constructions (see above), since the choice of a preposition depends on the word that governs it in most cases (usually a verb or a noun).

While lexical error annotation does not present particular difficulties, errors in discourse raise some questions. The **discourse section** itself is closely related to lexis, but the word usage here depends on different contextual parameters. As a result, we find here perfectly grammatical constructions that, nevertheless, turn out to be anomalous in relation to a wider context. Even though the discourse category was not as detailed as other categories, the preliminary results reveal some characteristic trends. Thus, most errors in discourse represent examples of a misuse of referential lexis (principally, a subject or object being inserted or missing) in anaphoric or cataphoric position (Ex. 23-25) and errors in discourse word order (Ex. 26-27). The explanation of erroneous

word order usually comes from a transpositional mechanism: as learners first lexicalize their thoughts in French, they frequently transpose the word order of French constructions to Russian (notably with *есть* "il y a" in the initial position, Ex. 26). However, transpositions from English are not so rare, as the learner wants to move away from the syntax of his or her native language: this is probably the case in Example 27, where the Russian construction follows the same word order as in French (cf. *le soutien des Français* / *French support*).

23) Я сказала, что увидела много интересных местов и **я** встретила много друзей. (Vanya, B2, FL)

24) Но подожди, я буду объяснять вам почему я **думаю это**. (Fr1 (M), A2, FL) > я так думаю

25) В регионе Веллингтона, есть много вина, особенно белое **вино**. (Caroline, B1, FL)

26) Город лежит на Северном острове (**есть два острова в Новой Зеландии**). (Caroline, B1 FL) > в Новой Зеландии есть два острова

27) Оппосиция получила **французков поддержку.** (Ed, A2, FL) > поддержку французов, probably from English

In addition to the referential lexis and word order, this category also includes errors in discursive lexis (Ex. 28-30) and discursive constructions (Ex. 31), which are close to purely lexical issues. Annotating these phenomena separately enables a more detailed analysis in the future.

28) В дом моих мечтаний, будут **тоже** чердак над спальнями… (Fr1 (M), A2, FL) > также

29) (…) массовый туризм разрушаеть земля и **более конкретно** туристических объектов. (Alexandre, B1, FL) > в частности

30) **Правда говоря** я ужо задумывался об этих фактах раньше, когда я изучял Француский язык. (Rouslan, B2, HL) > по правде говоря

31) **Это каникулы**! (Alexis, A1, FL) > Наступили каникулы!

Thus, the quantitative data of lexis and discourse reveal quite similar trends in error progression. Therefore, Figures 4 and 5 (see below) are the first not to show steadily decreasing lines. Instead, they seem to point out gradual changes in error number, both in lexis and in discourse, which, however, still decreases with improved proficiency. This tendency is typical for both groups of learners (foreign and heritage students).

In addition, preliminary data also show some differences between our two groups. At the B2-C1 levels HLs make more mistakes in idiomatic expressions and conjunctions than FLs.

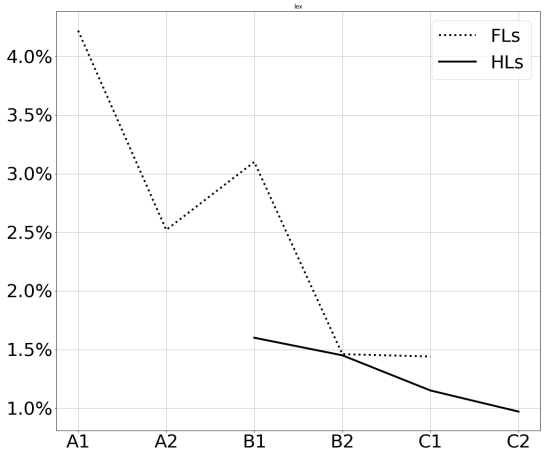

Figure 4. Progression of Lexical Errors of FLs (bold line) and HLs (dotted line)

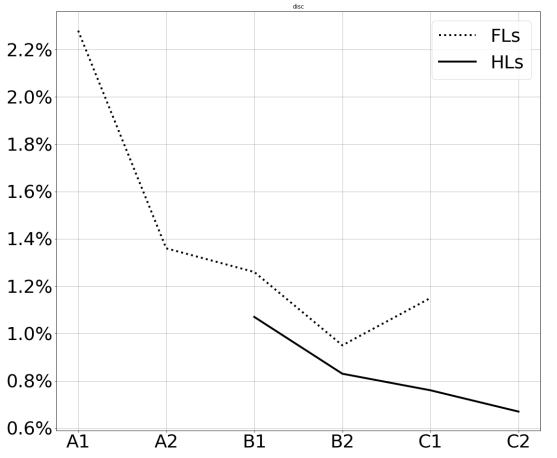

Figure 5. Progression of Discourse Errors in FLs (bold line) and HLs (dotted line)

Finally, the acquisition of discourse category also takes place gradually: while it is relatively easy to learn a linguistic form, its usage in an appropriate context is much more difficult and needs more practice. Our figures also show that the intermediate B1 level represents an important step in lexical or discursive acquisition, as the error ratio suddenly increases. A plausible explanation for this finding may be that, at this level, learners "feel more confident" in the Russian grammatical system and are expected to explore more challenging lexical topics and types of writing that go beyond simplified lexical domains and expository texts. However, factors influencing lexical and discursive mastery require further investigation.

6. General observations and further perspectives

The main purpose of this paper was to show that even "simple" primary data, without the usage of sophisticated statistical manipulations, can yield interesting results for use in learner corpus research. Thus, Figures 6 and 7 present the overall error distributions of the FLs and HLs (see next page).

As the data show, in both groups, the error ratio gradually decreases for each linguistic category. In addition, the large error ratio shows which areas are problematic for our groups and are likely to generate errors. It is worth mentioning here that the error rate in our groups is distributed as follows (see Table 5): the FLs are at 21.15% error rate, while the HLs reach almost 11% on average. For the advanced students (B2 and C1 levels), the error rate falls between 9.48%-13.11%, which is slightly higher than 6.3% error rate in RULEC-GEC Russian data, but still remains low on average compared to other learner corpora (English, Arabic) (Rozovskaya and Roth 2019, 6).

Table 5. Error rates

	A1	**A2**	**B1**	**B2**	**C1**	**C2**	**Total**
Foreign Learners	32.57	23.76	18.70	12.11	10.52	–	21.15
Heritage Learners	–	–	31.55	15.73	9.14	4.31	10.96
Total	32.57	23.76	19.04	13.11	9.48	4.31	15.39

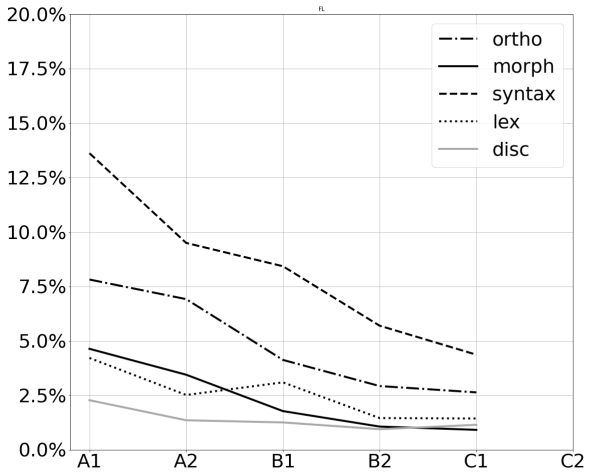

Figure 6. Dynamics of Errors in FLs by linguistic category

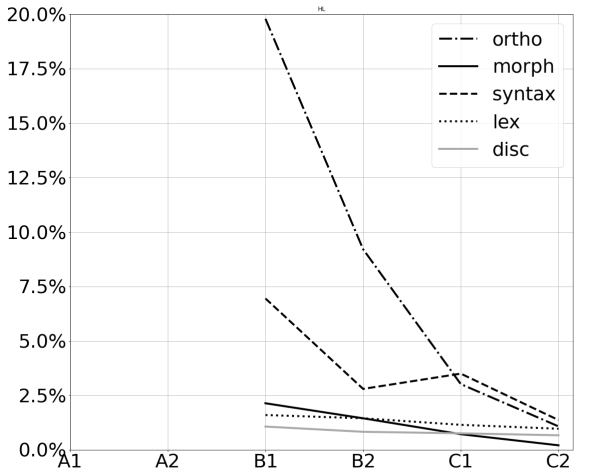

Figure 7. Dynamics of Errors in HLs by linguistic category

Due to specificity and different purposes of each learner corpus analysis, the comparison with other learner corpora data can be made only partly. Thus, Russian corpus RULEC-GIC of American students's texts places Russian spelling, vocabulary, noun case usage, punctuation and missing word at the top five error categories for FLs and HLs (Rozovskaya and Roth 2019, 6), while English corpus analysis points out English vocabulary, articles, and spelling as most problematic areas for Spanish students (López 2009, 684). As for our data, all mistakes made by FLs and HLs of Russian in a French-speaking environment are classified by linguistic category as follows:

Table 6. Ratio of errors made by FLs and HLs

	Spelling	Morphology	Syntax	Lexis	Discourse
Foreign Learners	26.26	12.77	**41.94**	12.58	6.46
Heritage Learners	**34.87**	6.87	32.26	15.57	10.43

The data show that the most problematic area for our FLs group is Russian syntax (42%), followed by spelling (26%), morphology (13%), lexis (12.5%) and, finally, discourse (6.5%). On the other hand, for our HLs group, the greatest challenge is spelling (35%), closely followed by syntax (32%), lexis (15.5%), discourse (10%), and morphology (6.9%). In summary, except for the spelling problems for HLs, syntax turns out to be the most problematic area for our groups of learners. However, this statement needs to be more nuanced: morphology and syntax in Russian are closely related. For example, inflectional morphemes in nouns are not only cumulative – they mark gender, number and case – but they also assign a syntactic role in a phrase. This also applies to verbs: morphologically inherent aspectual features imply restrictions to the verbal syntax. Thus, case and aspect choices were counted as syntactic problems in our data. Perhaps, it would be more appropriate for these particular questions to be further investigated by distinguishing a morphosyntactic category via a tag label, which would allow a more detailed picture of general error distribution to be drawn.

Besides general patterns reflecting Russian language proficiency, our study reveals important aspects for improving teaching methods: understanding typical areas of difficulty for specific learner groups allows to pay more attention to these issues during training. Identification of more problematic issues in each linguistic category of Russian grammar and for each linguistic level and group would be the next stage in error analysis. For these reasons, we believe it is necessary to set up a more effective error annotation system with a fine-grained description of each category. Moreover, quantitative analysis should gain in effectiveness when it is complemented by qualitative analysis since the same error ratio doesn't imply the same type of errors in different learner groups, and this is a topic for future research.

Acknowledgements
The authors would like to thank Natalia Partenheimer, Amanda Edmonds and the anonymous RLJ reviewers for their expertise, valuable comments and stylistic suggestions which substantially helped the authors improve the first draft of the paper.

References

Brunni, Sisko, Liisa-Maria Lehto, Jarmo H. Jantunen, and Valtteri Airaksinen. 2015. "How to Annotate Morphologically Rich Learner Language. Principles, Problems and Solutions." *Bergen Language and Linguistic Studies (BeLLS)* 6 (May): 133–52. https://doi.org/10.15845/bells.v6i0.812.

del Río, Iria, and Amália Mendes. 2018. "Error annotation in a Learner Corpus of Portuguese." *11th International Conference on Language Resources and Evaluation*. May 7-12. Miyazaki, Japan. http://hdl.handle.net/10451/36511 // https://www.aclweb.org/anthology/L18-1649.pdf. Accessed May 30, 2021.

Díaz-Negrilloa, Ana, and Valera Salvador. 2010. "A Learner Corpus-based Study on Error Associations." *Procedia Social and Behavioral Sciences* 3: 72–82. https://doi.org/10.1016/j.sbspro.2010.07.014

Granger, Sylviane. 2003. "Error-tagged Learner Corpora and CALL: A Promising Synergy." *CALICO Journal*: 465–80.

Granger, Sylviane, Gaëtanelle Gilquin, and Fanny Meunier, eds. 2015. *The Cambridge Handbook of Learner Corpus Research*. Cambridge

Handbooks in Language and Linguistics. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139649414.

Hana, Jirka, Alexandr Rosen, Svatava Škodová, and Barbora Štindlová. 2010. "Error-tagged Learner Corpus of Czech." In *Proceedings of the Fourth Linguistic Annotation Workshop*, ACL July 2010, Association for Computational Linguistics, edited by Nianwen Xue and Massimo Poesio: 11-19. Uppsala, Sweden.

Kisselev, Olesya, and Edie Furniss. 2020. Corpus Linguistics and Russian Language Pedagogy. In *The Art of Teaching Russian*, edited by Evgeny Dendub, Irina Dubinina and Jason Merrill: 307-29. Washington: Georgetown University Press.

Kor Chahine, Irina, Yulia Perova-Nouvelot, and Ekaterina Uetova. 2019. "Some Russian Prepositional Constructions through Russian Learner Corpus." In *AATSEEL Conference*, Feb 2019, New-Orleans, United States. hal-02099331

Kutuzov Andrey, and Elizaveta Kuzmenko. 2015. "Semi-automated typical error annotation for learner English essays: Integrating frameworks." *Proceedings of the 4th workshop on NLP for Computer Assisted Language Learning. NEALT Proceedings Series* 26. Linköping Electronic Conference Proceedings 114: 35–41. Vilnius, Lithuania. https://www.aclweb.org/anthology/W15-1904.pdf

Leech, Geoffrey. 1993. "Corpus Annotation Schemes." *Literary and Linguistic Computing* 8, no. 4: 275–81.

López, Willelmira Castillejos. 2009. "Error Analysis in a Learner Corpus: What Are the Learners' Strategies?" *A Survey of Corpus-based Research*: 675–90.

Lüdeling, Anke, Maik Walter, Emil Kroymann, and Peter Adolphs. 2005. "Multi-level annotation error annotation in a learner corpora." In *Proceedings of Corpus Linguistics* 1, Birmingham, England, July 2005: 14-17.

Mathet, Yann, and Antoine Widlöcher. 2019. "Annotation, évaluation et mesure d'accord en linguistique de corpus." *Revue française de linguistique appliquée* 1, no. 24: 111–28. https://www.cairn.info/revue-francaise-de-linguistiqueappliquee-2019-1-page-111.htm

Rakhilina, Ekaterina, Anastasia Vyrenkova, Elmira Mustakimova, Alina Ladygina, and Ivan Smirnov. 2016. "Building a Learner Corpus for Russian." In *Proceedings of the joint workshop on NLP for Computer*

Assisted Language Learning and NLP for Language Acquisition at SLTC, Nov 2016, Umeå, Sweden. http://aclweb.org/anthology/W16-65

Rosen, Alexandr, Jirka Hana, Barbora Štindlová, and Anna Feldman. 2014. "Evaluating and Automating the Annotation of a Learner Corpus." *Language Resources and Evaluation* 48: 65–92. Springer.

Rozovskaya Alla, and Dan Roth. 2019. "Grammar Error Correction in Morphologically Rich Languages: The Case of Russian." *Transactions of the Association for Computational Linguistics* 7: 1–17. https://doi.org/10.1162/tacl_a_00251

Uetova, Ekaterina. 2019. "The Acquisition of Russian Agreement and Case Government by French-speaking Students: Study Based on RLC." Master 1 diss. Moscow: Research University "Higher School of Economics".

Uetova, Ekaterina, Irina Kor Chahine, Marina Zhukova, Valeriia Lelik, Arina Molchanova et al. 2019. "Nekotorye tendencii v usvoenii sintaksisa vo frankojazychnoj srede." In *Aprel'skaja mezhdunarodnaja konferencija, sekcija "Russian language in the multilingual world"*, April 2019, Moscow: Research University "Higher School of Economics". https://hal.archives-ouvertes.fr/hal-02099956.