SlavVaria 1/2024. 99–115
DOI: 10.15170/SV.1/2024.99

GYÖRFI BEÁTA
(Szeged, Magyarország)

The Evolution of OES *ся*: a Corpus Investigation

Abstract: The present study focuses on the Old East Slavic reflexive pronominal enclitic *ся*. The research of the syntactic behaviour and historical evolution of this element thus far has been hindered by the vast amount of data. The application of corpus linguistics and statistical analysis, however, offer fresh avenues for understanding the complexities of this pronoun.

Keywords: OES, clitics, corpus linguistics, statistics, t-test, diachrony

0. Introduction

Ся is without doubt one of the most widely studied elements of Russian historical linguistics. It has drawn much attention due to its spectacular path of development: between the 10th and 17th centuries, it underwent a radical syntactic transformation from an independent word to a bound morpheme.

Due to its morpho-syntactic diversity *ся* is defined in various ways in historical grammars or by the taggers in the Russian National Corpus: it is either labelled as an accusative reflexive pronoun, a particle, or even subsumed within the verb itself, even when not fully merged.

The present study aims to re-evaluate existing claims about the nature and development of "*ся*," potentially revealing new insights into its syntactic behaviour. The investigation gains particular significance by the application of new methodological tools, particularly corpus analysis and statistics Thus far, owing to the considerable volume of examples and the gradual nature of this shift researchers had no reliable instruments to capture this linguistic change.

Consequently, in this investigation, *ся* will be considered from a different angle: its evolution from an independent word into a postfix will be examined employing the Russian National Corpus. First, we will present previous interpretations of *ся* in the scholarly literature. The second part will describe the corpus of research and the method of investigation. The investigation itself will be presented in the third section. The following aspects of the use of *ся* will be looked at:

- We will compare the frequency of *ся* as an individual word versus a postfix in Old East Slavic texts from the 11th and 15th centuries.
- We will assess the validity of Zaliznyak's (2008) hypothesis regarding the connection between clitic "*ся*" usage and other pronominal clitics.
- The distribution of these two *ся* functions in Old East Slavic and hybrid texts will be explored.

- An examination of *ся* placement within Old East Slavic monuments will be conducted.

The final section will present the study's conclusions.

1. Interpretation of *ся* in Russian historical linguistics

There appears to be a consensus regarding the etymology of *ся*. It traces back to the Proto-Indo-European pronominal stem **sve-* or **swé-* (ФАСМЕР), which evolved into Proto-Slavic **sę* (ФАСМЕР) carrying a reflexive meaning in the accusative case.

Borkovskij and Kuznetsov (БОРКОВСКИЙ, КУЗНЕЦОВ 1965: 213, 273) discuss *ся* within the categories of both pronouns and voice. Regarding pronouns, they claim that the distinction between short and long forms began to diminish early on. As for voice, they posit that *ся* initially indicated intransitivity. This element retained its mobility until the 18th century, when its merger with the verb became standardized.

Chernyh (ЧЕРНЫХ 1962: 276–77) mentions *ся* peripherally in his discussion of the verbal category of voice. Kolesov (КОЛЕСОВ 2005: 358–59) discusses *ся* under the heading of pronouns, focusing on the relationship between full and short forms.

Selishchev (СЕЛИЩЕВ 2001: 112) considers only the dative forms of the reflexive pronoun to be an enclitic. He argues that the accusative form, always stressed, has a more ancient origin and was not used enclitically.

Samojlenko (САМОЙЛЕНКО 1962) presents a distinct perspective on the origin and use of short form reflexives. He proposes that these forms were originally independent pronouns until the 11th-12th centuries, not attached to neighbouring words. He cites their use with prepositions, distance from the verb they later joined, and ability to express contrast as evidence. Initially, short forms outnumbered long forms. However, due to their multifunctional nature, lack of inflection, and association with the expression of voice, they gradually fell out of use.

YANOVICH (ЯНОВИЧ 1986: 191–192) suggests that the accusative and dative forms of short form reflexives merged as markers of voice. He also claims that *ся* participated in syntactically free phrases, allowing pre- and postpositional placement.

Recent studies (ЗАЛИЗНЯК, 2008: 28) categorize short form reflexives as enclitics. Zaliznyak establishes a ranking system for Old East Slavic (OES) enclitics based on their position within clusters. The first five ranks are taken by discourse clitics *же, ли*; *бо*; *ти* and *бы*. The sixth rank is for short form dative pronouns (*ми, ти, си, ны, вы, на, ва*). Accusative pronouns (*мя, тя, ся, ны, вы, на, ва, и, ю, е, э, я*;) occupy the seventh rank, while auxiliary clitics (*есмь есми, еси, есмъ есме, есмо, есмы, есте, есвѣ, еста*) constitute the final rank. According to this classification *ся* is perceived as a 7th rank enclitic.

This historical overview reveals that scholars viewed *ся* as a short pronoun, a marker of voice, and most recently, an enclitic. Consensus regarding the primacy of these functions remains elusive. Additionally, the syntactic positions these forms take, requires further exploration.

2. Method and corpus of investigation

The investigation is carried out on the Old East Slavic subcorpus of the Russian National Corpus, specifically, on the Old East Slavic subcorpus. At the time I accessed (08.04.2024), it contained 301 texts with 838,928 words. The RNC provides disambiguation and full part-of-speech (POS) and morphological tagging for its entries. The subcorpus encompasses a diverse range of genres, including chronicles, hagiography, legal acts, didactic tracts, pilgrimage accounts, and literature.

For this study, a subset of 45 texts totalling over 465,000 words was chosen. Selection criteria excluded translations, charters, and acts issued outside of Russia. The resulting corpus represents a variety of genres, including hagiography, literature and folklore, prayers, epistles, homilies, rules, and orations. Importantly, 32 texts are classified as Church Slavic (CS), while the remainder exhibit a hybrid nature, blending CS with elements of the spoken language.

As it has already been mentioned in the introduction, the sheer volume of data has historically been a major obstacle in studying *ся*. Modern technology (linguistic corpora) and related methodologies (statistical and distributional analysis) however, seem to cope with this challenge.

Corpus-assisted research allows the frequency and distributional analysis of large amount of data to be performed. It utilizes a large database of authentic texts (the historical subcorpus in our case). It is a quantitative method in that works with numbers which reflect the frequencies of words and phrases in corpora (BREZINA 2018: 3). Concordancing programs allow the user to research for specific target words in a corpus providing frequency information and a list of occurrences in context. They enable the investigation of grammatical constructions the given word takes part in (BIBER, CONRAD, RAPPEN 1998: 13). The application of corpus-based analytical techniques for studies in historical linguistics enables the investigation of language change. The insights gained can either support or contradict existing theories and thus enrich previous theoretical accounts.

To help us navigate in the maze of corpus data, corpus investigation is complemented by statistical survey. Statistics is a discipline which helps us make sense of quantitative data (BREZINA 2018: 3). Statistics in corpus linguistics is about mathematical modelling of a complex linguistic reality. It can help us discover and elucidate patterns and tendencies in the data that might otherwise remain hidden (BREZINA 2018: 5). Statistical methods, with

their mathematical tools, allow us to efficiently handle the vast amount of frequency data generated by corpus analysis.

4. Analysis

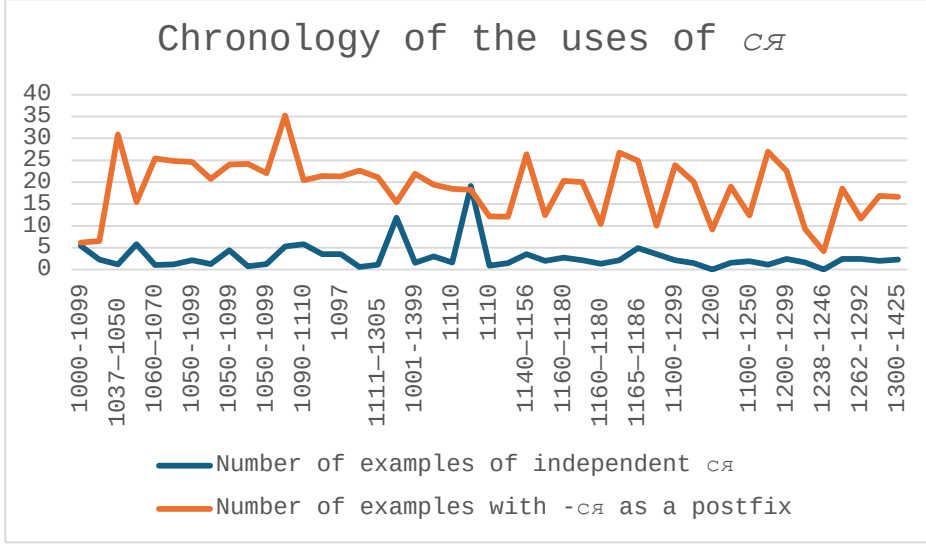
I started the quantitative investigation of *ся* by looking at its frequency in the investigated monuments. I distinguished between occurrences of *ся* as a postfix and as an independent word. The first column of table 1.[1]below contains the title and time of compilation of monuments in brackets. In the second column of the table I included information about the number of words in each text. In the third column the number of independent uses of *ся* can be found, while in the last column the number of examples of *ся* as a postfix.

The columns with frequency data contain two figures separated by a slash. The first number denotes the actual number of occurrences, the so called raw frequency counts. Since texts vary in length, raw frequencies are not directly comparable. To address this, data normalization was performed by dividing the frequency counts by the word count in each text. However, the resulting values were very small. Therefore, following (BIBER, CONRAD, RAPPEN 1998: 261) they were multiplied by 1,000 to achieve a scale corresponding to text length. Consequently, the second number in the boxes shows these normalized, comparable counts.

Investigating the diachronic (historical) aspect of this change presents a challenge. Accurately pinpointing a monument's compilation time is often difficult. Only a limited number of literary texts have definitive dates. The majority of documents have a broad timeframe assigned, which can blur the data and hinder precise analysis.

The following chart visually represents the chronological development in the uses of *ся:*

[1] The tables used for the analyses are accesible at https://drive.google.com/drive/folders/1B-cHi-thWLwWzLVxA2XBB1qIcDoKhJHw?usp=sharing



Looking at the frequency of *ся* we find that examples of *ся* as a postfix outnumbered independent uses from as early as the 11th century. We can also observe a narrowing gap between the two usage patterns over time. Also, the uses of *ся* in both cases exhibit a decline in time, especially as regards independent uses. By the 13th century, some monuments (*Хожение Антония* or *Слово о погибели русской земли*) lack independent uses entirely. The chart also displays a few intriguing spikes (around 1090 and 1110 for independent *ся)*. These anomalies can be attributed to two factors: on the one hand, the imprecise dating of some documents and the uneven distribution of texts across the timeframe can contribute to these fluctuations. On the other hand, differences in language varieties (Church Slavonic vs. hybrid) between the texts might also play a role.

To understand the reasons behind these spikes, I conducted a statistical analysis using an independent-samples t-test. This test is a descriptive statistical method that helps determine whether the observed difference in the uses of *ся* between the two language varieties (Church Slavonic and hybrid) is statistically significant.

4.1. How can statistics interpret the results?

The aim of the statistical analysis was to prove, that there is a meaningful difference in the independent and postfixal uses of *ся* in hybrid and CS monuments. As part of the statistical procedure, I set up a null hypothesis (it is practically the negation of the actual hypothesis). My null hypotheses for both cases assumed that there is no difference in the uses of *ся* between the two language varieties.

Considering the considerable variation within the data samples, Welch's independent-samples t-test was employed for the analysis. This specific t-test compares the mean values of a linguistic variable (in our example, the relative frequency of *ся*) and takes into consideration the internal variation in each group expressed as variance (BREZINA 2018: 187).

The t-test has three basic assumptions about the data:

Independence of observations, which means that the observations in one sample are independent of the observations in the other sample.

Normality: ideally, both data samples should approximate a normal distrbution (bell curve).

Homoscedasticity or equality of variances: it requires both samples to exhibit roughly similar variance. (BREZINA 2018: 187–189, T-TEST).

One of the key concepts of t-tests is variance. Variance is perceived as the sum of squared distances of individual values from the group mean divided by the degrees of freedom. The degree of freedom (df) is a complex concept, which is used when dealing with calculations based on a sample (i.e. a corpus). It signifies the number of independent ('free') components in the calculation, i.e. components that are not predictable from the previous components. In practice, it is the number of cases minus one (in our case, 12 for hybrid and 31 for CS texts). Variance is calculated according to the following formula:

$$\text{variance}=\frac{\text{sum of squared distances from the mean}}{\text{degrees of freedom}}$$

For our investigation we got the following variances[2] :

In the case of independent uses of *ся*: s1=1,61, while s2=5,34.
For postfixal uses: s1=7,11, s2=5.95

The t-test is calculated according to the following formula:

$$t-test=\frac{\text{mean of group1}-\text{mean of group2}}{\sqrt{\frac{\text{variance of group1}}{\text{number of cases in group1}}+\frac{\text{variance of group2}}{\text{number of cases in group2}}}}$$

As can be seen from the equation, there are three factors that have an effect on whether the test will be significant: 1) the size of the mean difference, 2) the

[2] The details of the calculation are accessible in an Excel file at https://drive.google.com/drive/folders/1B-cHi-thWLwWzLVxA2XBB1qIcDoKhJHw?usp=sharing

variance in each of the two groups and 3) the sample size (number of texts in both groups).

A high t-test value indicates a large mean difference, low variance within groups, and a large sample size. This combination suggests strong evidence in the data to reject the null hypothesis and conclude that the two groups are statistically distinct regarding the use of the linguistic variable under investigation.

The next section will present the summarized results of the t-test analysis in tables.

Independent uses of *ся*

Register	Group mean	Number of cases/texts	Variance	t-value	p-value
CS	2,26063	33	1,61	3,01	0,043
Hybrid	4,2876	12	5,34		

Postfixal *ся*

Register	Group mean	Number of cases/texts	Variance	t-value	p-value
CS	19,31	33	7,11	2,05	0,045
Hybrid	17,63	12	5,95		

As a result for the independent uses of *ся* I got a t-value 3,01 and for the uses of *ся* as a postfix, the test gave a lower t-value: 2,05.

The final stage of interpreting the t-test results involves assessing their level of significance. This is achieved using a statistical indicator, the p-value (probability value). The p-value represents the likelihood of obtaining the observed results if the null hypothesis was true. The test is considered to be significant if $p<0,05$. Statistically significant results are considered unlikely to have arisen solely due to chance (STATISTICAL ANALYSIS: URL).

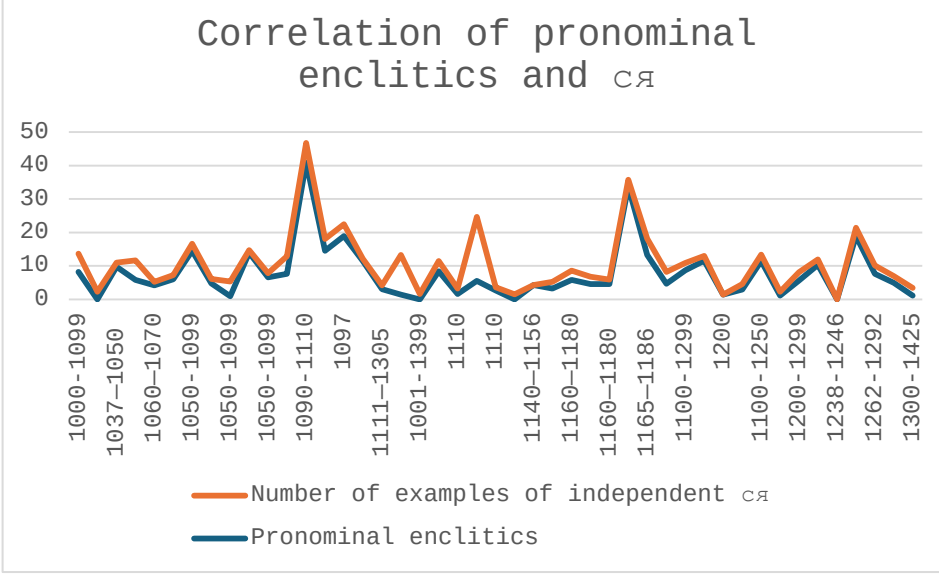
The p-value for the first investigation was 0,043 and 0,046 for the second. Since both p-values are lower than 0.05, we have sufficient evidence to reject the null hypothesis. These values suggest that the difference in the uses of *ся* between the language varieties is statistically significant. We can conclude, therefore, that *ся* was more widely used in the hybrid language than in CS.

4.2. Correlation between the loss of pronominal enclitics and the frequency of *ся*

Zaliznyak (ЗАЛИЗНЯК 2008: 219) proposes a link between the development of the clitic *ся* and the fate of other pronominal enclitics. These

accusative and dative enclitics were replaced by their stressed full-form counterparts and eventually lost. The same occurred to *ся* in combination with prepositions. However, the full reflexive form could not substitute the enclitic for reflexive meaning. Consequently, the enclitic *ся* gradually merged with the verb, ultimately becoming a postfix.

Based on this hypothesis, I investigated whether such a correlation exists in my corpus. Since I already had frequency data for the accusative reflexive clitic, I needed to collect frequencies for other pronominal enclitics (accusative and dative). Due to the unreliability of the RNC parser, I manually checked each enclitic, specifying its form and grammatical features for the search. I then combined the counts and normalized the data by multiplying by 1000 (see Table 2).



Our analysis, visualized in the following diagram, reveals a near-perfect correlation between the two groups of enclitics.

4.3. The position of short form accusative reflexives

The syntactic distribution of short-form accusative reflexive clitics can initially appear erratic: they can precede, as well as proceed their verbal host, or can occupy a medial position inside the clause, often at a distance from the verb they refer to.

Slavic scholarship offers two main approaches to analysing clitic positions: Zaliznyak (2008) focuses on the clitic's proclitic and enclitic nature, while a group of Czech linguists proposes a system with four distinct clitic positions.

4.3.1. The position of *ся* – pre- or postposition

Zaliznyak (ЗАЛИЗНЯК 2008: 174–219) observes that the reflexive particle *ся* appeared in both preverbal (proclitic) and postverbal (enclitic) positions in historical texts. He investigates the frequency of proclitic *ся* across different time periods and text categories (bookish vs. non-bookish) by analysing the ratio of preverbal and postverbal usages. Zaliznyak employs coefficients (percentages) for his analysis. His findings reveal a significant distinction in the evolution of proclitic *ся* between the two text categories. In early non-bookish texts, a high initial frequency of proclitic *ся* steadily declines throughout the 11th and 12th centuries. Conversely, proclitic *ся* was very uncommon even in the earliest bookish monuments.

While Zaliznyak's research focused solely on the distribution of proclitic *ся*, it provided the impetus for my investigation into the frequency of pre- and postverbal usages in my corpus. This approach allows me to examine the validity of Zaliznyak's theory within the context of my own data relying on the instruments of corpus analysis.

In my version of the investigation I considered three positional varieties of *ся*: 1) postposition immediately after the verbal host (1); 2) postposition after the verbal host followed by a higher ranking enclitic (2); 3). preposition, preceding the verbal host (3).

(1) *и*	*повелѣ*	*кр(с)тити*	*сѧ. /*	*и*
Conj.	order.Aor.sg.3.	christen.Inf.	Cl.	Conj.
єпп(с)ъ	*жє*	*корсуньскыи*	*с*	*попы*
bishop.Nom.sg.	Encl.	Korsun.Adj.Nom.sg.masc.	Prep.	priest.Instr.pl.
ц(с)р[ц]ины	*wгласивъ*	*и /*	*и*	*кр(с)ти*
Tsaritsyn.Adj.Instr.pl.	announce.Part.act.Past.sg.Nom.	him.Acc.sg.	Conj.	christen.Aor.sg.3.
володимѣра.	ПВЛ по И. списку			
Volodymer.Acc.sg.				

„And he ordered to get christened. And the bishop from Korsun with the priests from Tsaritsyn announced him and the christened Volodymer."

(2) *тако*	*рекѧ. /*	*кобь*	*ми*	*нє*
so	say.Part.act.Pres.Nom.sg.	omen.Nom.sg.	I.Dat.sg.	not
дасть.	*с*	*вами*	*поити. /*	*воротивъ*
give.Pres.sg.3.	Prep.	you.pl.Instr.	go.Inf.	turn.Part.act.Past.Nom.sg.

же	сѧ	назадъ.	и	погна
Encl.	Encl.	back	Conj.	chase.Aor.sg.3.
вборзѣ.	Волынская летопись			
quickly				

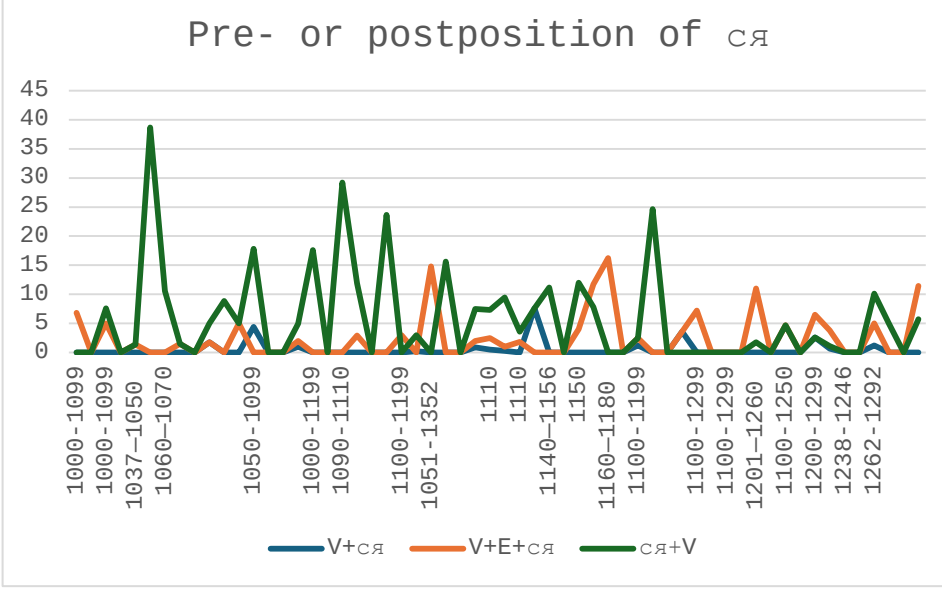
„And he said: the omen does not let me go with you. So, he left quicly turning around."

(3) посемь	бывшу	вечеру /	и	повелѣ
Conj.	be.Part.act.Past.Dat. sg.	evening.Dat.sg.	Conj.	order.Aor.sg.3.
изьнести	сѧ	на	дворъ.	ПВЛ по И. списку
take out.Inf.	Pron.refl.Acc.sg.	Prep.	yard.Acc.sg.	

„As it was already evening, he ordered to move out to the yard."

The results are summarized in table 3.

The chronological development of the position of the clitic *ся* relative to its verbal host is depicted in the following diagram:



The diagram shows that the preposition of *ся* was prevailing till the 12th century. The second most frequent position became the postposition of the reflexive pronoun after another enclitic (in most cases *же* or *бо*).

Given the presence of outliers in the overall data, I opted to compare the distribution variations between the two registers (CS and hybrid) using the median of each group. The median is a valuable statistical measure in situations with skewed data. Unlike the mean, the median is not significantly affected by extreme values at the fringes of the distribution. The median is calculated the following way: in case of an odd number of values it is the middle value; in the case of an even number of values, it is the mean value of the two central values. (BREZINA: 2018: 10)

	Median
∑ V+ся	0*
∑ V+E+ся	1,77
∑ ся+V	4,95
CS V+ся	0*[3]
CS V+E+ся	1,5
CS ся+V	4,825
Hybrid V+ся	0,1
Hybrid V+E+ся	2,25
Hybrid ся+V	7,4

Based on the analysis, we can draw the following key observations: 1) The most frequent position for *ся* across the entire corpus and within both language varieties (CS and hybrid) was prepositional, preceding the verb. 2. The results for the hybrid register partially support Zaliznyak's hypothesis. The frequency of proclitic *ся* was indeed highest in this register, suggesting a potential correlation with specific text styles. 3. instances of direct postposition of *ся* to the verb were extremely rare. In both the overall corpus and in CS texts, the median frequency was so low that it registered as zero.4) When *ся* appeared in postverbal position, it was usually preceded by a higher-ranking discourse enclitic. In these cases, the verb typically occupied the first position in the clause, and the enclitics occupied 2P.

4.3.2. The Czech theory of clitic positions

As regards the position of enclitics in Slavic, in the scholarly literature two canonical positions can be distinguished: the traditional Wackernagel Position – the position after the first stressed word or phrase of the clause and the secondary contact position - directly next to its host, i.e. the verb (KOSEK et al.: 2019).

To navigate the apparent complexity of enclitic placement, a group of Czech linguists (KOSEK et al.: 2019) proposed a theory of four clitic positions based

[3] The median was so low, that Excel gave a 0 result.

on their observations of Old Czech enclitics. While observing the „competition“ and/or „cooperation“ between the two possible word order patterns of old Czech enclitics, they distinguish 4 positions. I have adapted this theory for the Old East Slavic (OES) system. We will explore the details of these four clitic positions in the context of OES:

a) The *postinitial contact position* (2PC position). In this position, the enclitic (E) follows directly after the initial word of the clause, which often coincides with the verbal host (H). As *ся* can appear pre- and postpositionally as well, the host in many cases can follow the pronoun. A peculiarity of OES: the initial word is usually followed by other, higher-ranking enclitics.

H+*ся*

H+E+ *ся*

[]+ *ся*+H

[] +E+ *ся*+H+[]*

(4) is an example for the 2PC position. Here *ся* stands after the initial verb and a second rank enclitic.

(4) изнемоглъ	бо	ѧ	бѣ. /	ходивъ
die.pqperf.sg.3.masc.	Cl.	Cl.	be. Aux.aor.sg.3.	go.Part.Act.Past.Nom.sg.
на	воину. /	а	король	оугорьскыи
Prep.	war.N.Acc.sg.	Conj.	king.N.Nom.sg.	Ugric.Adj.Nom.sg.masc.
иде	во	оугры	Киевская летопись	
go.Aor.sg.3.	Prep.	Ugorland-Acc.		

„As he had passed away when going to war, the Ugric king returned to the land of Ugors.”

b) The *post-initial isolated position* (2P position). The enclitic *ся* occurs after the initial phrase of the clause (or a higher ranking enclitic), but is crucially separated from its host (H), and it is not immediately followed by its host either:

[]+ *ся*+ ….. +H

[]+E+ *ся*+ ….. +H

(5) exemplifies the post-initial isolated position, where the reflexive pronoun takes the second position following a conjunction. It is separated from its verbal host *уладити* by an auxiliary verb.

(5) а	послалъ	ко	мнѣ /	ѿсюда
Conj.	send.Part.sg.masc.	Prep.	me.Dat.sg.	from here.Conj.
ся	бѣхо(м̂)	оуладили ·	Владимир Мономах Письмо	
Cl.	be.Aor.pl.1.	settle.Part.pl.		

„And he sent tom e from the place he had settled."

c) The *non-post-initial contact position* (NPC position). The enclitic occurs anywhere except directly after the initial phrase and is adjacent to its host.

[] []*+H+ *ся*

[] []*+ *ся*+H

In (6) *ся* takes the 4th place in the clause following its verbal host, occupying thus a non-post-initial contact position. It should be also mentioned, the verbal form itself is defective, as it is merged with a proclitical *ся*.

(6) и	начаша	сѧисповѣдати	сѧ.	шво
Conj.	begin.Aor.pl.3.	confess.Inf.	Cl.	Conj.
ко	игоуменемь. /	дроузии	же	к
Prep.	abbot.Dat.pl.masc.	others.pl.Nom.	Cl.	Prep.
попомъ.	и	дьякономъ.	Волынская летопись	
priest.Dat.pl.	Conj.	deacon.Dat.pl.		

„And they began to confess their sins to the abbot, and others to the priest and the deacon."

d) The *non-post-initial isolated position* (NPI position). In this case, the enclitic occurs anywhere except in the post-initial position and it is not adjacent to its host:

[] []*+H+[]* + *ся*+[]*

[] []*+*ся* +[]*+ H+[]*

The non-post initial isolated position was the least frequent position for *ся*. This position was characteristic of *ся* functioning as particles or pronouns. In (7) *ся* occupies a middle position in the clause following a pronoun and preceding an adverb.

(7) ТЪІ	НАДѢЄШИСѧ	БѢЖАТИ	В	ПОЛОВЦѢ. /
you.Nom.sg. 2.	hope.Pres.sg. 2.	run.Inf.	Prep.	Cumania.Acc. pl.
А	ВОЛОСТЬ.	СВОЮ.	ПОГУБИШИ. /	ТО
Conj.	county.Acc.sg.	own.Acc.sg.fem.	destroy.Pres.sg.2.	Conj.
К	ЧѢМОУ	Сѧ	ВПѧТЬ	ВОРОТИШЬ.
Prep.	what.Dat.sg.	Partic.	again	turn.Pres.sg.2.
Киевская летопись				

„You are hoping to flee to Cumania and destroy your own country. So why are you turning again?”

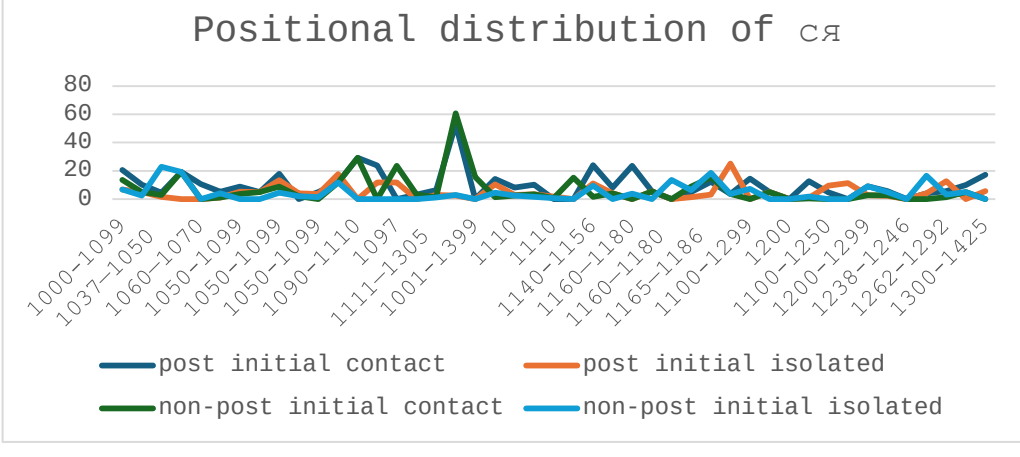
In the next section I will leverage the Czech theory of clitic positions, adapted for OES, to examine the distribution of *ся* in my corpus.

The following search parameters were set for the analysis: [word = “ся”]. The respective positions were determined „manually”.

Table 4. summarizes the findings. The first column lists manuscript titles. Next, I included a column for the time of compilation of the text and another one for the number of words. The following four columns represent the counts for each clitic position (2PC, 2P, NPC, NPI). An additional column captures instances where the context was insufficient to determine the position of *ся*.

Each box in the table displays two figures separated by a slash. The first number represents the raw frequency count, while the second number is the normalized count (multiplied by 10,000 in this case).

The following diagram visualises the chronology of this distribution:



Two facts can be deduced from this quite skewed diagram: 1) a general decline in the overall number of syntactic constructions containing "free" *ся* starting from the 12th century 2) a decrease in the uses of *ся* in the post initial contact position approximately from the 13th century.

Since the data distribution was skewed again, I calculated the median values to compare the distribution of *ся* in the overall corpus and across the two language varieties to explore potential language-specific variations.

	median
∑ post initial contact	6,905
∑ post initial isolated	3,07
∑ non-post initial contact	2,68
∑ non-post initial isolated	2,01
CS post initial contact	5.59
CS post initial isolated	2,415
CS non-post initial contact	3,25
CS non-post initial isolated	2,265
Hybrid post initial contact	8,07
Hybrid post initial isolated	3,4
Hybrid non-post initial contact	1,59
Hybrid non-post initial isolated	1,76

By analysing the medians for each clitic position across the CS and hybrid language variety, we can make the following insights:

1) The post initial contact position emerges as the dominant one as evidenced by its higher median value. The second most frequent position in the overall corpus and in hybrid texts was the post initial isolated, aligning with Wackernagel's law. 2) At the same time CS texts show a tendency for *ся* to attach directly to its verb host (non-post initial contact position), making it the second most dominant position based on the median.

5. Conclusions

In the course of the investigation, I looked at the behaviour of *ся* in the OES subcorpus of the RNC applying the tools and methodology of corpus linguistics and statistical analysis. This approach enabled the analysis of a vast dataset which thus far hindered the research of this topic.

In the manuscripts *ся* appears as an independent word as well as a postfix. Frequency analysis demonstrated a clear dominance of postfixal *ся* over independent *ся* from the very beginning of the textual record. Notably, the use of independent *ся* began to decline from the 13th century onwards.

The OES corpus comprises two language varieties of texts: Church Slavic and hybrid. Interestingly, the frequency of *ся* (both postfixal and independent)

displayed a significant difference between these registers. The hybrid register exhibited a higher overall prevalence of *ся* compared to the CS register. This finding supports Zaliznyak's hypothesis, suggesting that *ся* was more widespread in the spoken language.

Next, I checked Zaliznyak's hypothesis regarding the potential correlation between the frequency of *ся* and that of other accusative and dative pronominal enclitics. The results of our corpus analysis revealed a remarkably strong positive correlation, suggesting that the decline of other pronominal enclitics mirrored the decrease in the use of independent *ся* over time.

At first glance, the potential positions for *ся* might appear unpredictable. To navigate this complexity, I carried out a twofold investigation of its positional varieties. First, I studied the pre- and postpositional distribution of *ся*. Analysis revealed a shift in preference over time: prepositional *ся* dominated until the 12th century, followed by a rise in postpositional *ся* occurring after a higher-ranking enclitic. Second, employed the framework proposed by Czech linguists, which distinguishes four clitic positions. This investigation indicated that the prevailing position for *ся* was 2P adjacent to its verbal host. However, the second most dominant position differed between language varieties: in hybrid texts isolated 2P, in CS the non-post initial contact position prevails. Finally, the non-post-initial isolated position – associated with pronominal and participial functions of *ся* – was the least common position across the whole corpus.

In conclusion, our corpus-based investigation of the reflexive element *ся* has yielded significant insights. While the potential positions for *ся* might initially appear complex, the study revealed a clear underlying order. This order was brought to light by applying the methodologies of corpus linguistics and statistical analysis. These tools facilitated the examination of a vast dataset, enabling us to observe diachronic trends and language-specific variations in the usage of *ся*.

Literature

BIBER, CONRAD, RAPPEN 1998 = BIBER D., CONRAD S., RAPPEN R. Corpus Linguistics. Investigating Language Structure and Use. Cambridge, 1998. DOI: 10.1017/CBO9780511804489

BREZINA 2018 = BREZINA V. Statistics in Corpus Linguistics. A Practical Guide. Cambridge, 2018. DOI: 10.1017/9781316410899

KOSEK et al. 2019 = KOSEK P., ČECH R., NAVRÁTILOVÁ O., MAČUTEK J. Wackernagel's Position and Contact Position of Pronominal Enclitics in Older Czech. Competition or Cooperation? // Jazykovedný časopis, 2019. 70, 267–275. DOI: 10.2478/jazcas-2019-0057

STATISTICAL ANALYSIS = The Beginner's Guide to Statistical Analysis | 5 Steps & Examples. https://www.scribbr.com/category/statistics/

T-TEST = https://www.statology.org/t-test-assumptions/

БОРКОВСКИЙ, КУЗНЕЦОВ 1965 = БОРКОВСКИЙ В.И., КУЗНЕЦОВ П.С. Историческая грамматика русского языка. Морфология. Москва, 1965.
ЗАЛИЗНЯК 2008 = ЗАЛИЗНЯК А.А. Древнерусские энклитики. Москва, 2008.
КОЛЕСОВ 2005 = КОЛЕСОВ В.В. История русского языка. Москва, 2005.
САМОЙЛЕНКО 1962 = САМОЙЛЕНКО С.Ф. Из истории основ и грамматических форм личных местоимений в славянских языках // Филологические науки. № 2. 3–15.
СЕЛИЩЕВ 1952 = СЕЛИЩЕВ А.М. Старославянский язык. Учебное пособие. Москва, 1952.
ЧЕРНЫХ 1962 = ЧЕРНЫХ П.Я. Историческая грамматика русского языка. Москва, 1962.
ФАСМЕР = ФАСМЕР М. Этимологический словарь русского языка. https://lexicography.online/etymology/vasmer/%D1%81/%D1%81%D1%8F
ЯНОВИЧ 1986 = ЯНОВИЧ Е.И. Историческая грамматика русского языка. Москва, 1986.