Statistics in Empirical Translation Studies
(Challenges of Rustvelology in the Digital Age)

Tandaschwili Manana, PhD
Goethe University of Frankfurt
Frankfurt, Germany

ORCID: 0009-0005-7812-9124

DOI: https://doi.org/10.62343/cjss.2023.246

ABSTRACT

The present paper addresses the problem of data processing in multilingual parallel corpora. It focuses on the difficulties that can arise in the statistical processing of linguistic data in a multilingual parallel corpus, such as Rustaveli Goes Digital and the solutions that may be useful for overcoming the challenges of empirical translation research. During the statistical analysis of the corpus *Rustaveli Goes Digital*, we encountered certain problems that we will discuss in this article, namely the reliability of the statistical analysis in creating the index. Although many different ready-made tools are successfully used in linguistics for statistical analysis, the data processing of texts can still be very inaccurate without considering the grammatical characteristics of the languages. As empirical material, the text of the epic *The Knight in the Panther's Skin* was chosen in three languages: Georgian, Abkhazian, and Megrelian. The paper will show why ready-made tools such as KWIC and Voyant are not suitable for Caucasian languages and what problems the use of such tools can lead to.

Keywords: Caucasian languages, digital Rustvelology, translation studies, data processing

INTRODUCTION

The origins of Quantitative Linguistics date back to ancient Greece and India. One strand of tradition consists of the application of combinatorics to linguistic objects (Biggs, 1979); another is based on elementary statistical surveys, which are referred to under the keywords colometry and stichometry (Pawłowski, 2008). A thematically broader and more continuous development of quantitative linguistics (QL) began in the 19th century. Among other things, this involved sound and letter statistics as preparatory work for the development of stenographic systems and as a basis for language comparisons, the different forms of verse and the duration of sounds in relation to word length, and even the exact dating of an author's works. The studies on sound length and ideas on the interaction of other linguistic characteristics presented the first concepts that led to the development of language laws in the 20th century, most famously Zipf's Law. In the 20th century, several other topics were added: identification of anonymous authors, action quotient, language structure, language change law, type-token relation, development of children's language skills, dynamic aspects of text structure, etc. The objective of QL in the 21st century is more demanding – the formulation of language laws and, ultimately, of a general theory of language in the sense of a set of interrelated language laws. The present paper focuses on the questions of what kind of difficulties can arise in the statistical processing of linguistic data in a multilingual parallel corpus such as Rustaveli Goes Digital, and what solutions can help overcome the challenges of empirical translation research.

Parallel corpus *'Rustaveli goes digital'*

Shota Rustavelis Epos *The Knight in the Panther's Skin*. The Epos by Shota Rustaveli is the most significant literary work on Georgian intangible cultural heritage. The Epos was created in the 12th century and has been handed down in over 160 different manuscripts. Its significance has gone far beyond Georgia's borders and now has a prominent place in the history of world literature: the collection of manuscripts of the Epos is included in the UNESCO World Intangible Cultural Heritage Register. The Knight in the Panther's Skin is an excellent literary work and one of the most crucial components of defining the identity of the Georgian nation. The research of this unique literary work with modern methods is not only a challenge for the Kartvelology of the 21st century. However, it will also contribute to the scientific research of Georgian intangible cultural heritage and the internationalization of modern Kartvelology. The creation of a parallel corpus of the Epos' translations is

an important step for conducting interdisciplinary research. In addition, the multilingual parallel corpus can be successfully used in bilingual/multilingual education.

Research the history of the epos and modern challenges. Scientific research on the epos began in the 18th century when King Vakhtang VI added a scientific analysis to the first printed book from 1712. This formed the basis for further research on the epos, which gradually developed into a separate field of Kartvelology – into **Rustvelology**.

The history of the Rustvelology covers more than three centuries and can be divided into several stages:

1. Textological research;
2. Textological-lexicological research;
3. The Soviet stage of Rustvelologian studies;
4. Interdisciplinary research;
5. Internationalization of Rustvelologian studies;
6. Digitization of Rustvelology.

The digitalisation of Rustavelology began in 2018 at the University of Frankfurt with the project *Rustaveli goes digital*, led by Prof. Manana Tandashvili, since 2023 by Dr. Mariam Kamarauli. The project aimed to create a **big data** in Rustvelology - a multilingual parallel corpus of translations of Shota Rustaveli‘s epic in 58 languages. This goal required the solution of the following tasks (Tandaschwili, 2022, p.53):

I. Technical tasks:
   - conceptualization of the structure and design of the corpus and preparation of a technical framework;
   - digitization of the original text and its translations in 58 languages (including the digitization of several translations that co-exist in one language);
   - structural preparation of the texts for their inclusion in the parallel corpus;
   - connecting the digitized and structured texts with each other in accordance with chapters and stanzas;

II. Methodological tasks:
   - conceptualization of the methodological framework for the study of translation strategies in the corpus;
   - development of a methodological framework for creating a basic concept of automatic processing of a poetic parallel corpus;

III. Theoretical tasks:

- aligning the multilingual parallel corpus and preparing the texts for interdisciplinary research (philosophical, religious, sociological, cultural-specific, astrological, etc. terms);
- verification of capabilities of automated translation strategies research.

METHODS

Statistical processing of the corpus

The multilingual parallel corpus *Rustaveli Goes Digital* (Beta version led by Dr. Mariam Kamarauli) currently contains 32 parallel translations of the full text of the epic in 20 languages (Georgian, German, English, Spanish, French, Italian, Turkish, Azerbaijani, Kyrgyz, Russian, Belarusian, Ukrainian, Greek, Arabic, Persian, Armenian, Ossetian, Lithuanian, Mingrelian, Svan).

We have already used statistical processing to analyze address formulas in the parallel corpus to determine and compare the strategies used by the translators. The analysis of the address formula in the translations revealed the following structures (Tandashvili & Kamarauli, 2023, pp. 99-101):

1. The addressee of the communication is lexically given in the address formula (sun); it acts as a vector of the communication channel and ensures the accuracy of the reference. The addressee of communication is often named directly before direct speech, in the initial position of the sentence.
2. An interjection in the address formula (o, sun) serves to open the communication channel and ensures its activation.
3. Using the second-person pronoun or possessive pronoun in the address formula (you, sun; my sun) expresses the speaker's status in the communication act.
4. Using both indicators of expressiveness (an interjection and a second-person pronoun or possessive pronoun) at the same time, "'O, my sun," increases the degree of expressiveness and gives more power to the information following in the direct speech.

We compared in 20 translations the statistics of equivalence degree of "sun" (as a denotative or connotative equivalence) and the address formulas in terms of the level of expressiveness. As it turned out, the frequency of use of denotative equivalents of "sun" is directly proportional to the degree of expressiveness (Tandashvili & Kamarauli, 2023, p. 101):

1. Those translators who have systematically chosen the denotative equivalent for "sun" in the address formulas are rendering them with a higher degree of expressiveness. This correlation is confirmed by a lower number in the "difference" column (especially in the case of Wardrop, de la Torre, Barea, and Martinez).
2. The correlation, established as the result of statistical analysis, is relevant from the point of view of a complex evaluation of the quality of a given translation because it clearly shows the translators' efforts to preserve as much as possible of the original – not only the artistic language of the author but also his philosophical-religious and aesthetic worldview.
3. The results obtained using the corpus linguistic method indicate that the quality of the translation can be "measured" empirically. This, in turn, allows us to determine the strategies selected by the translator and the expediency and appropriateness of their application in the target text.

During the statistical analysis of the corpus *Rustaveli Goes Digital*, we encountered certain problems that we would like to discuss in this article, namely the reliability of the statistical analysis in the creation of the index.

Tokenization and accuracy of the statistical processing

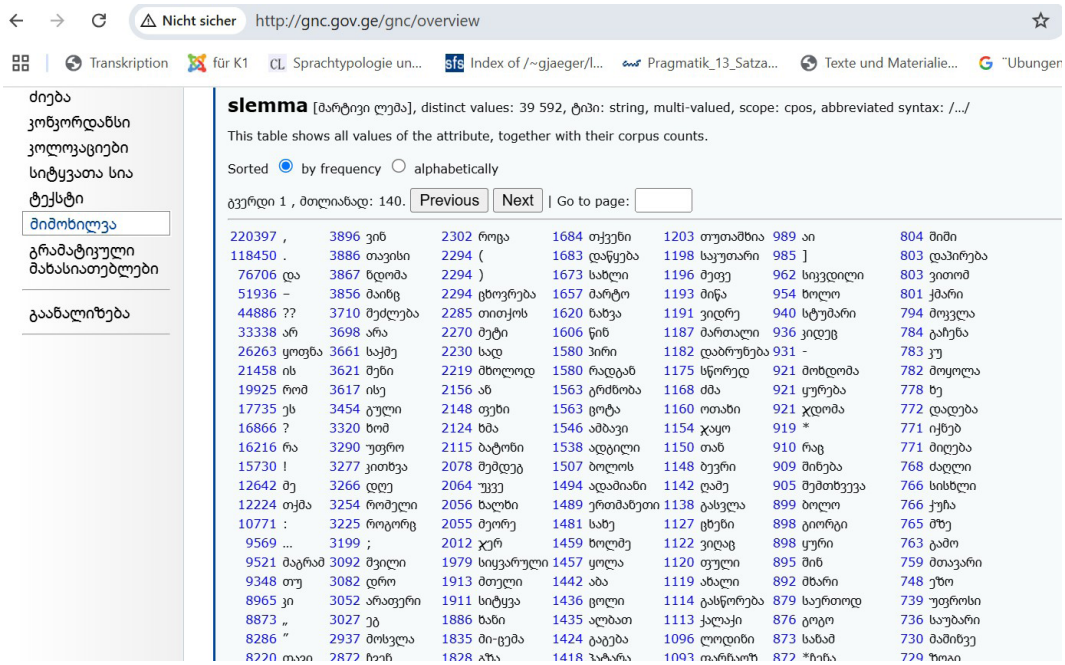
In linguistics, a frequency class is a statistical measure of the frequency of use of a word in a natural language. Frequency classes can be considered on two linguistic levels: a single word form (token) or an entire lexeme with various grammatical forms. The most common statistical analysis is carried out by the type-token relation (TTR), used in quantitative linguistics and quantitative stylistics to measure linguistic diversity in a text. It is defined as the relation of unique tokens divided by the total number of tokens. When tokenizing a text, a list of tokens is created without considering its grammatical representation. In the case of inflected languages, an annotation is required not only to statistically record individual forms of the word but also to assign the various forms to the corresponding lexeme. The accuracy of the frequency of lexemes in a corpus depends heavily on how precisely the grammar of this language is mapped in the annotation system. Compare the frequency of words and lexemes in Vefxistqaosani in GNC.

Table 1

Frequency of word forms

Table 2

Frequency of lexemes



The most frequentative ten word forms (tokens) in GNC: *da* და *and*, *ra* რა *what*, *ar* არ *not*, *me* მე *I*, *tu* თუ *if*, *mas* მას *he/she/it (Dat.)*, *iqo* იყო *was*, *rom* რომ *that*, *ese* ესე *this (also as an definite article)*, *igi* იგი *he/she/it*.

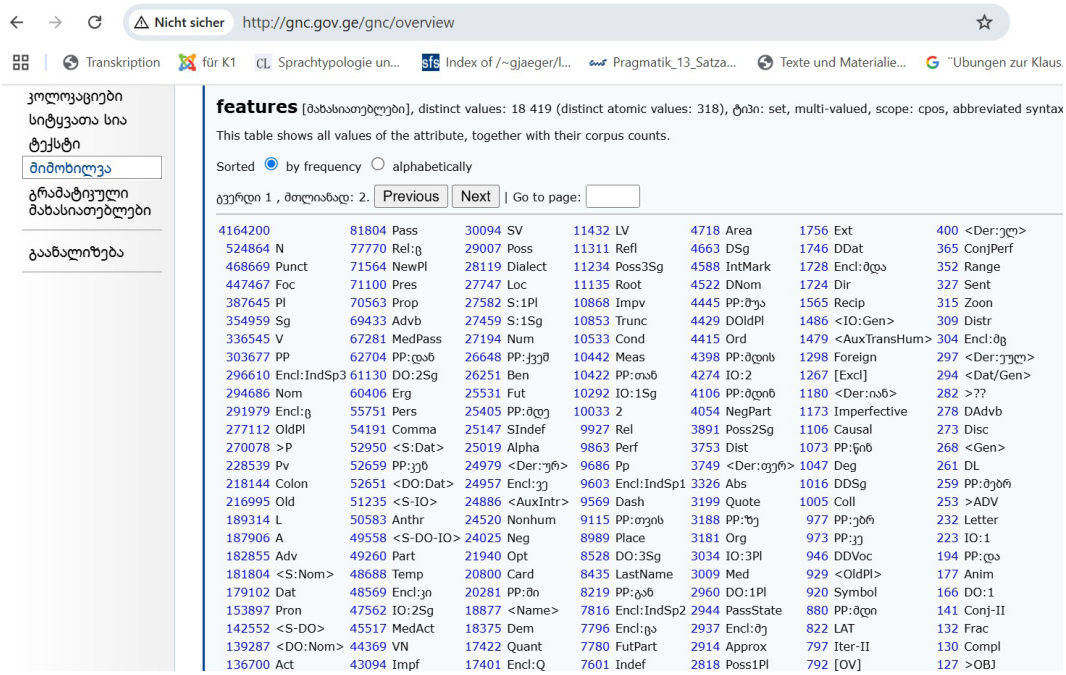
vs.

The most frequentative ten lexemes in GNC: *da* და *and*, *qopna* ყოფნა *to be*, *is* ის *he/she/it*, *ra* რა *an*, *ar* არ *not*, *es* ეს *this*, *misi* მისი *his/her/its*, *me* მე *I*, *kaci* კაცი *man*, *čemi* ჩემი *my*.

GNC can output statistics according to word class (noun, adjective..), semantical roles (subject, object), functionality (focus), as well as the grammatical features: case, person, TAM, genus verbi and so on.

Table 3

Frequency of grammatical features



The higher the annotation quality of a corpus, the more accurate and precise the results of the statistical processing of the linguistic data are.

DISCUSSION

For valuable evaluation of the statistical data, the specific characteristics of the respective languages should be taken into account; otherwise, the results of the statistical processing of the linguistic data will be inaccurate. This is particularly important for statistical analyses in parallel corpora. Below, we will present this on

the example of the multilingual parallel corpus Rustaveli Goes Digital in the case of Abkhazian and Megrelian translations of the epic The Knight in the Panther's Skin by Shota Rustaveli.

Tokenization in Abkhazian

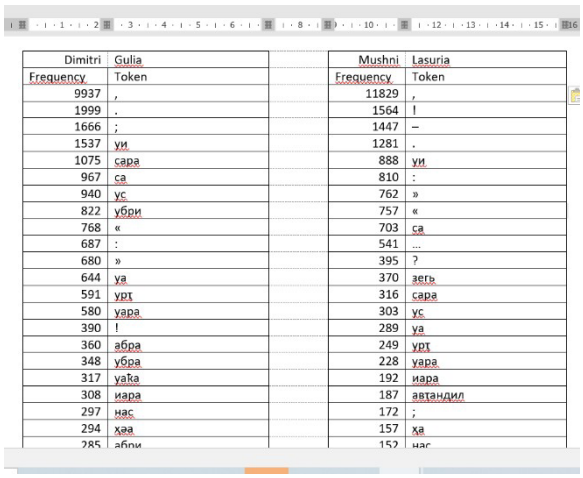
In computational linguistics, tokenization refers to the segmentation of a text into word-level units. A token is a character string that is assigned a type by a formal grammar. The token forms the basic lexical unit for the parser. As mentioned above, the accuracy of the frequency of lexemes in a corpus depends heavily on how closely the grammar of that language is mapped in the annotation system. We will show here what happens when tokenization a non-annotated corpus of Abkhazian. As a sample text, we take two Abkhazian translations of Rustaveli's epic, which are by Dimitri Gulia and by Mushni Lasuria.

Frequency of pronouns

The frequency of use of pronouns is one of the central common statistical indicators during the automatic processing of texts. The first and second person pronouns are generally characterized by the highest frequency among the personal pronouns. The word frequency via KWIC of the Abkhazian translations is given in the table 4 below:

Table 4

Comperison of the word frequency in both Abkhasian transitions by Gulia and Lasuria



Dimitri	Gulia		Mushni	Lasuria
Frequency	Token		Frequency	Token
9937	,		11829	,
1999	.		1564	!
1666	;		1447	–
1537	уи		1281	.
1075	сара		888	уи
967	са		810	:
940	ус		762	»
822	убри		757	«
768	«		703	са
687	:		541	…
680	»		395	?
644	уа		370	зегь
591	урт		316	сара
580	уара		303	ус
390	!		289	уа
360	абра		249	урт
348	убра		228	уара
317	уаҟа		192	иара
308	иара		187	автандил
297	нас		172	;
294	ҳаа		157	ҳа
285	абри		152	нас

The frequency of the personal pronouns (1st, 2nd and 3rd person in singular and plural) is listed individually in the corpus:

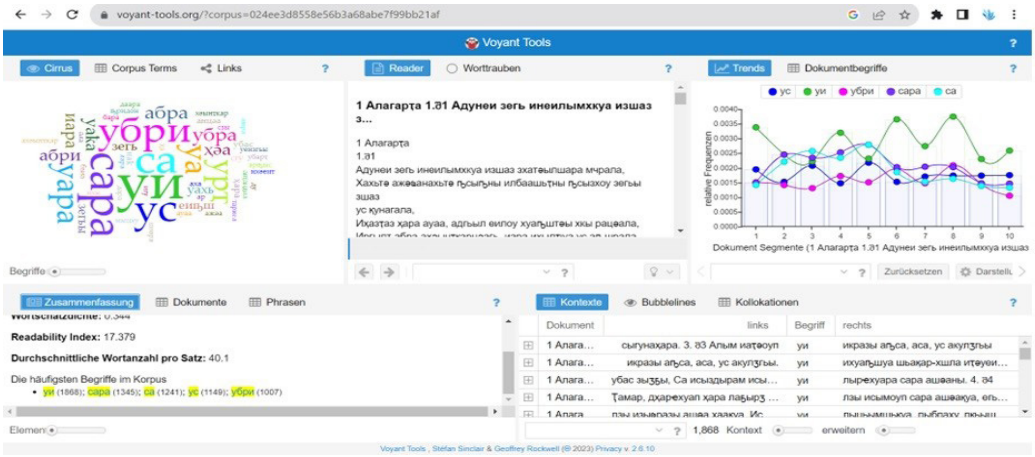
Gulia	*Lasuria*
сара 1075	са 703
са 967	сара 317
уа 644	уа 289
уара 580	уара 228
иара 308	иара 192
ҳа 176	ҳа 157

and so on…

The statistical processing of the texts via Voyant tools (https://voyant-tools.org/) can be visualized by five most frequently occurring words.

Figure 1

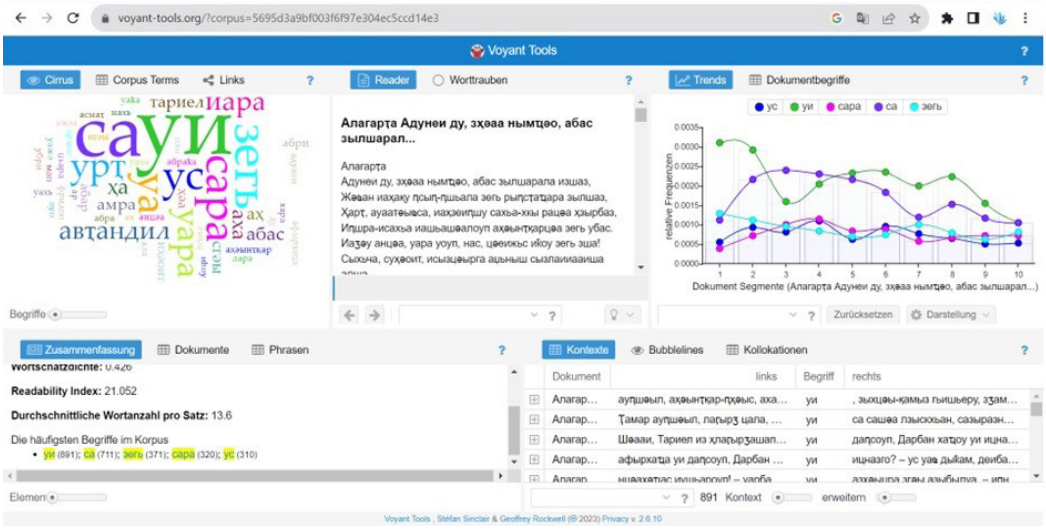
Visualization of the statistical processing of Dimitri Gulia's Abkhazian translation



The frequency of personal pronouns in the Georgian-Abkhazian parallel corpus of „Vefkhistqaosani" shows much more grammatical varietes in the Abkhasian translations than in the source text itself. This is due to the differentiation of personal pronouns according to form and semantic class in Abkhasian. There are two forms of personal pronouns in the Abkhazian language: long forms such as ***сара*** (1pers.), ***уара, бара*** (2. pers.) (marked with -pa) and short forms - ***са, уа/ба*** (without marking). The second personal pronouns in singular are additionally marked according to genus (***уара*** M and ***бара*** F) and the third personal pronouns in singular also (***иара*** M, ***лара*** F, ***иара*** N).

Figure 2

Visualization of the statistical processing of Mushni Lasuria's Abkhazian translation



The comparison of the frequency between the long and short forms of the personal pronouns is shown in the following diagrams:

Translation by Gulia

Figure 3

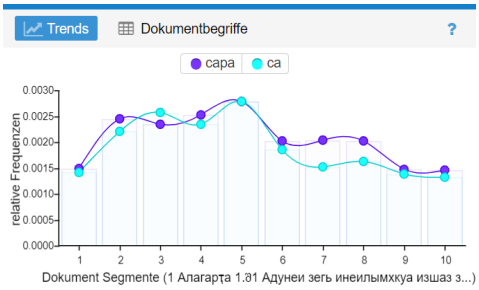


Figure 5

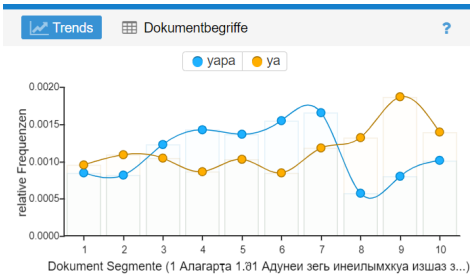
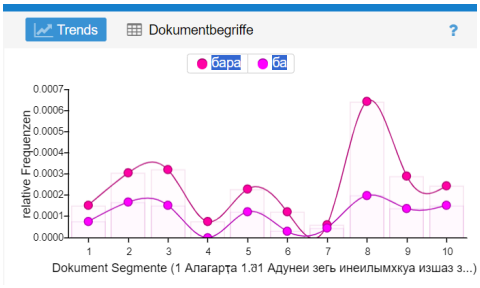


Figure 7



Translation by Lasuria

Figure 4

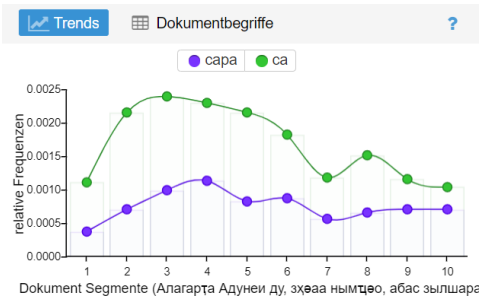


Figure 6

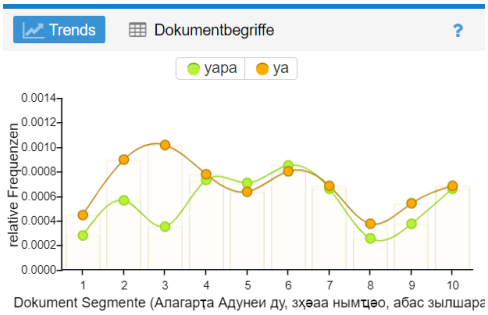
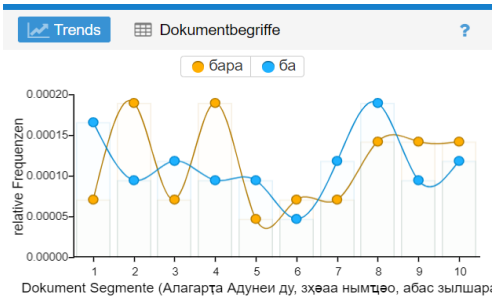


Figure 8



The more precise statistics in both translations show the differences in the use of long and short forms of personal pronouns according to the genus. As the comparison of personal pronouns differentiated by gender shows, the masculine personal pronoun occurs more frequently than the feminine personal pronoun:

2. Pers. Pron. M - 1224 (***уара*** 580 / ***уа*** 644), 3. Pers. Pron. M – 310 (***иара*** 308/ ***иа*** 2)

cf.:

2. Pers. Pron. F - 200 (***бара*** 138 / ***ба*** 62), 3. Pers. Pron. F – 183 (***лара*** 140/ ***ла*** 43)

Table 5

Comparison of long and short forms of personal pronouns in the Abkhazian Translations

Personal pronoun	*Genus*	*Gulias Translation*	*Lasurias Translation*
I person	-	сара 1075	са 703
		са 967	сара 316
II person	M	уа 644	уа 289
		уара 580	уара 228
	F	бара 138	бара 48
		ба 62	ба 48
III person	M	иара 308	иара 192
		иа 2	иа 1
	F	лара 140	лара 90
		ла 43	ла 56
I person, pl.	-	ҳара 201	ҳа 157
		ҳа 176	ҳара 58
II person, pl.	-	шəара 95	шəара 33
		шəа 53	шəа 15
III person, pl.	-	дара 90	дара 59

Regarding the frequency of use of long and short forms of personal pronouns, Gulia clearly favors the longer forms (except the 2nd personal pronoun masculine). However, Lasuria presents a different picture: the longer forms are favored only for the first personal pronoun in singular and plural, as well as for the second personal pronoun feminine. The question of what causes the high frequency of long or short forms in Lasuria and Gulia's translations requires additional corpus-linguistic and contextual analysis. This is a separate research topic, and we will not address this issue here.

This statistical analysis confirms the need to account for the grammatical features of the language when annotating the corpus in order to capture both the general (part-of-speech) features and the specific characteristics. For instance, these words

in Abkhazian should be annotated as personal pronouns, but also according to gender and form (long or short). We did this manually in our case, but in an annotated corpus, this should occur automatically.

Frequency of nouns in Abkhazian

The nominal morphology of Abkhazian differs from Kartvelian and the East Caucasian languages: Abkhazian has no declension, only the category of number, definiteness and possessiveness. This phenomenon is illustrated by the nouns in Abkhazian: the nouns are often marked by definiteness or possessive markers, which appear as prefixes to the nouns. The lexeme *მზე* *sun*, which occurs 309 times in the original text, corresponds to several inflected forms in the Abkhazian translations, which are marked by possessiveness differentiated by gender and thus result in the content of a noun phrase.

Table 6

Comparison of the "sun" in source language and target language

	Rustaveli	*Gulia*	*Lasuria*	*Grammatical category*	*Content*
Word	მზე „*sun*“	амра 134	амра 111	Definiteness	*the sun*
		рымра 3	рымра 4	Possessiveness, number (3.Pl)	*their sun*
		сымра 17	сымра 11	possessiveness (1. Sg)	*my sun*
		умра 5	умра 10	Possessiveness, genus (2.Sg.M)	*your (M) sun*
		ҳамра 3	ҳамра 10	Possessiveness, number (1.Pl)	*our sun*
		имра 6	имра 9	Possessiveness, genus (3.Sg.M)	*his sun*
Frequency	309	169	176		

For clarity, we will cite some examples from the source text of the epos and the Abkhazian translation by Gulia:

1.51.Ҭинаҭингьы амра иаҧылган, иаҵахын амра ҭинаҭинзарц!
*Tinatin was more beautiful than **the sun**, **the sun** wanted to be Tinatin.*

Cf.

თინათინ **მზესა** სწუნობდა, მაგრა **მზე** თინათინებდა.
*Tinatin resented the **sun**, but the **sun** was shining.*

38.920 Авҭандилгьы дигәалашәеит **имра** лаша, дызбылуа,

*Avtandil also remembered **his** bright **sun**, which burns him.*

Cf.

ავთანდილსცა მოეგონა **მისი მზე** და საყვარელი.
*Avtandil also remembered **his sun** and lover.*

34.820 Аҳәынҭқар: „Иаҳцәыʒма **ҳамра**, имзахама, нарха змам?“

*King: „Do we have lost **our sun**, it has become a moon without life?“*

Cf.

მეფემან ჰკითხა: „წასრულა **მზე** დაუდგომლად, მთვარულად?“
The king asked: „Has :gone, quietly disappeared like the moon?“

The Georgian does not have a grammatical category of definiteness. In the case of the lexical item „*sun*“, however, the reference is clearly definite (on the semantic level). This is morphologically marked in the Abkhazian translation by the **а-**prefix: **а-мра** (1.51).

Table 7
Expression of definiteness in Abkhazian

Language	*Lexem*	*form of language expression*	*Level*
Georgian	მზე	Implicative expression of reference	Semantic level
Abkhasian	а-мра	explicative expression of reference	Morphologic level

The following example demonstrates the ability of the Abkhazian language to indicate the category of possession in nouns by means of prefix morphemes, which are additionally differentiated by gender in the 2nd and 3rd person singular. In our case, it is the noun „*sun*“, to which the masculine possessive prefix of the 3rd person **и-** is added: **и-мра** (38.920). This noun in Abkhasian corresponds to the noun phrase **მისი მზე** *„his sun“* in the source text:

Table 8
Comparison by expression of the possessiveness in Georgian and Abkhazian

Language	*Lexeme*	*Structure*	*Form of language expression*	*Level*
Georgian	**მისი მზე**	NP	Explicative reference expression (with person and deixis specification)	Morphosyntaktic level
Abkhasian	**и-мра**	N	Explicative reference expression (with the specification of person and genus)	Morphologic level

Unlike the previous example, in this case, the reference is explicitly expressed in both languages, however, in addition to the difference in grammatical categories, they also differ from a structural point of view, which is both from the point of view of quantitative processing of the text (statistical analysis, e.g. during tokenization) and from the qualitative point of view (in translation studies, when parallelizing the text in establishing equivalence purpose) creates certain problems:

Avtandil also	*remembered*	***his sun***	*bright*	*which burns him*
Авҭандил-гьы	дигәалашәеит	**и-мра**	лаша	дызбылуа

ავთანდილსცა	მოეგონა	**მისი**	**მზე**	და	საყვარელი
Avtandil also	*remembered*	***his***	***Sun***	*and*	*lover*

The third example differs significantly from the two previous cases:

34.820 Аҳәынҭқар: „Иаҳцәызма **ҳамра**, имзахама, нарха змам?“

*King: „Do we have lost **our sun**, it has become a moon without life?“*

Cf.

მეფემან ჰკითხა: „წასრულა **მზე** დაუდგომლად, მთვარულად?“
*The king asked: „Has **the sun** gone [from us], quietly disappeared like themoon?“*

In the Abkhazian translation, the noun sun **ҳамра** (34.820) is accompanied by the **ҳ-** prefix of the 1st person plural. A two-person verb renders the predicate in the Abkhazian sentence:

иа - ҳ - цәызма	*vs*	წასულ-ა
$DO_{3Sg.} – S_{1pl.}\text{-}V_{tr.}$		$V_{Int.}\text{-}S_{3Sg.}$

The grammatical and pragmatical modification of the Georgian verb in the Abkhazian translation ($V_{Int.}\text{-}S_{3Sg.} > DO_{3Sg.} – A_{1pl.}\text{-}V_{tr.}$) is conditioned by the context: the departure of Avtandil causes the regret of the king Rostevan and also the royal court of Arabia. Accordingly, in the Abkhazian translation, the translator changes the perspective of king Rostevan‘s statement: Avtandil‘s departure is told from the

perspective of the king, which causes a grammatical change in the predicate of the Abkhazian sentence: an additional actant enters the verb иа-**х**-цәыʒма (1st person plural), which is also reflected in the noun through the possessive marker: **х**-амра.

Table 9

Language	*Lexeme*	*Syntaktische Funktion*	*Structure*	*Features*
Georgian	**მზე**	Subject	N	$N_{Nom.Sg.}$
Abkhasian	**ҳ**-амра	Direct object	PossPron+N	$N_{Sg.}$+PossPron.$_{3Pl.}$

This strategy used by the translator creates certain problems when parallelizing the text (in order to establish equivalence):

King	*Do we have lost*	***our sun***	*it has become a moon*	*life*	*having without*
а-ҳәынҭқар	**иаҳцәыʒма**	**ҳамра**	имзахама	нарха	змам

მეფემან	ჰკითხა	წასრულა	**მზე**	დაუდგომლად	მთვარულად
The king	*asked*	*to have gone*	***sun***	*disappeared*	*like the moon*

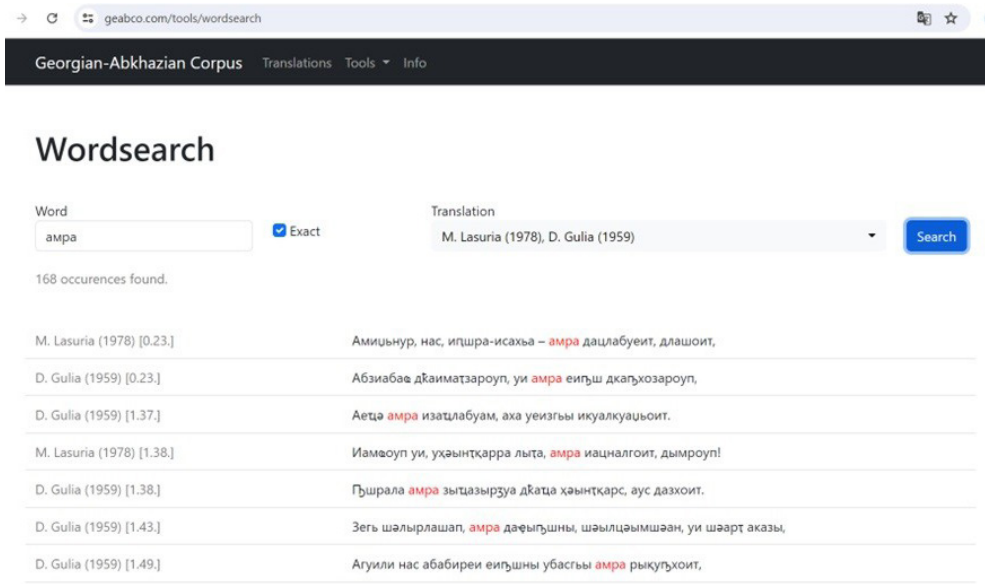
A few examples given here are only a hint of the problems that can arise during the statistical processing of the Georgian-Abkhazian parallel corpus via simple statistical analyses due the texts are not annotated. Today, only a simple search in the corpus is possible:(See Figure 9)

The above problem was solved by Paul Meurer in Abkhazian National Corpus (*The Abkhaz National Corpus*, n.d.) The AbNC was developed in the years 2016–2018 in a project financed by USAID, with participants from Sukhumi, Tbilisi, Frankfurt and Bergen. It comprises more than 10 million tokens of texts from various genres and is morphologically annotated. The corpus is hosted in the Corpuscle corpus management tool, which has advanced possibilities for searching and viewing the corpus texts. Simple search allows the search to word forms, but the advanced search allows you to search by word, lemma, slemma, stlemma or grammatical features.

The search can be limited to certain subcorpus or text, as in the given case: word form is searched only in the Abkhazian translation of Shota Rustaveli‘s epic.

Figure 9

Search result in Georgian-Abkhazian corpus



The difference between statistical analysis of raw texts and annotated texts is enormous. Precise statistical processing of the data in the parallel corpus Rustaveli goes digital requires a high quality of annotation for all languages integrated into the parallel corpus so that the parallel corpus can be used efficiently for statistical data processing.

Figure 10

Advanced search of lemma ***а́-мра*** *in Abkhazian National Corpus*

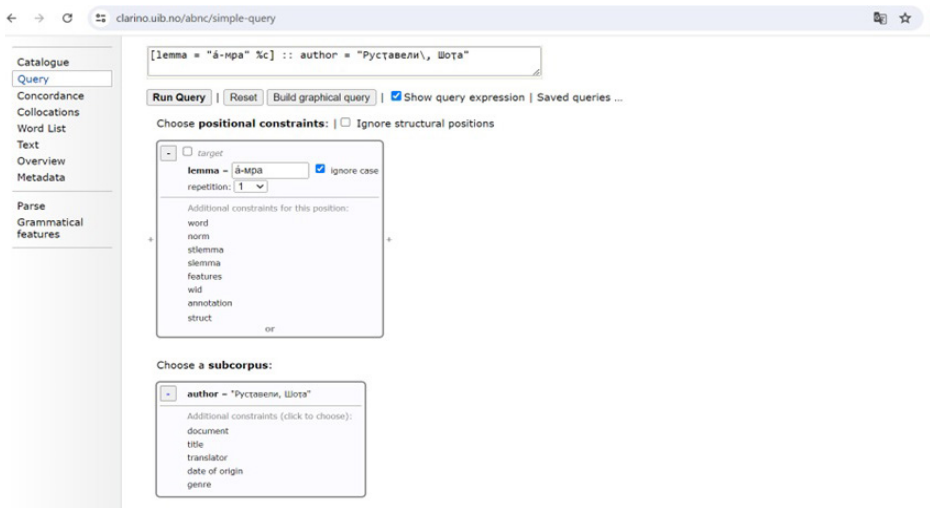
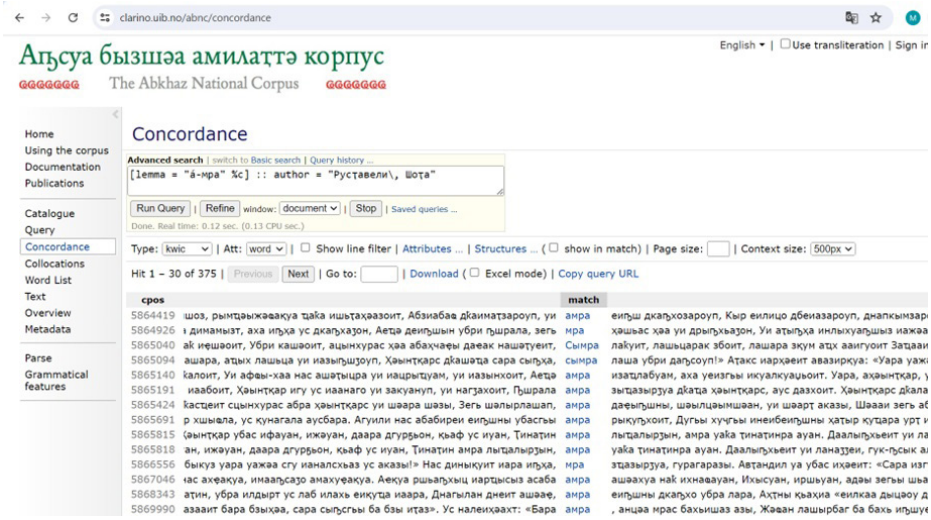


Figure 11

Concordance of lemma ***а́-мра*** *„the sun“*



Special features of the rendering of negation in Megrelian

In this section of the paper, I will further discuss the problematic aspects of statistical data processing in the multilingual parallel corpus *Rustaveli goes digital* using the example of the Megrelian translations of the epic. In particular, I will address the issue of how the category of negation is rendered in Georgian and Megrelian and the challenges of tokenization in the Georgian-Megrelian parallel corpus.

In Georgian, the category of negation is conveyed through both verb and noun morphology. The particles used in verb morphology form a three-member system: არ (not), ვერ (can‘t), and ნუ (don‘t). According to scientific literature, their functional-semantic distribution is as follows: არ expresses categorical negation, ვერ indicates the negation of possibilty, and ნუ denotes prohibition. However, the intensity of the semantic function of these negation particles can be modified by combining them with verbs in different screeves. For example, in the screeves of the third series of the tense-aspect-mood (TAM) system, the particle არ loses its categorical nature and conveys a neutral negation (Kurdadze et al., 2022, p. 208). In some TAMs, the particle ნუ expresses a threat (usually in combination with the particle **აბა** *aba*) or a wish; it is also used in curse formulas.(See Table 10)

Functional semantics of the negation particles become much more complicated when considering semantic groups of verbs or syntactic constructions which they can build:

a) The particle არ does not express categorical negation in verbs that cannot combine with the particle ვერ**.** Cf.: არ მწყურია *I‘m not thirsty,* არ შემიძლია *I can not,*

Table 10
Distribution of negation particles in Georgian

	Categorical Negation	*Negation of Possibility*	*Prohibition*	*Neutral Negation*
I Series	**არ**	**ვერ**	**ნუ**	-
II Series	**არ**	**ვერ**	-	-
III Series	-	-	**(ნუ)**	**არ**

არ მესმის *I don't hear* in contrast to *ვერ **მწყურია**, *ვერ **შემიძლია**, *ვერ **მესმის** (Djorbenadze 1984: 141).

b) In addition, the negation particle ვერ is not used with statical verbs არ აწერია *it is not written on it*, არ ახატია *it is not painted on it* in contrast to *ვერ აწერია**,*** *ვერ ახატია, and inversive verbs (verbs with a dative construction in the present tense): არ მშია *I'm not hungry*, არ მიყვარს *I don't love it* in contrast to *ვერ მშია**,*** *ვერ მიყვარს (Chumburidze, 1970, p. 42), with potential: არ იჭმევა *not edible,* არ ისმევა *not drinkable* in contrast to *ვერ იჭმევა**,*** *ვერ ისმევა (Machavariani, 2002, p. 100) and with verbs that express not having or lacking a property: არ გააჩნია**/**არ მოეპოვება *he/she/it does not possess/does not own* in contrast to *ვერ გააჩნია**,*** *ვერ მოეპოვება (Chumburidze, 1970, pp. 42-43).

Table 11
Distribution of negation particles in Georgian by different verb types

Type of verb	*Categorical Negation*	*Negation of Possibility*	*Neutral Negation*
Inversive verbs	-	**ვერ**	**არ**
Statical verbs	-	**ვერ**	**არ**
Verbs with potentialis		**ვერ არ**	-
Verbs of existence	-	**ვერ**	**არ**

The distribution of negation particles gives an interesting picture in different grammatical moods, in particular, the particle ვერ cannot be confirmed with imperative. It is usually used with indicative and conjunctive. The particle ნუ**,** on the contrary, is used with imperative and optative forms (Chumburidze, 1970, p. 42). (See Table 12)

The use of the particles ნუ and არ on a pragmatic level shows an interesting picture: these particles can convey identical functional content by combining with different TAM forms of the verb. For example, the negative verb form in the conjunctive II with ნუ particle - ნუ დაწერდა - has the same pragmatic content as the negative verb form in the perfect II with არ particle - არ დაეწერა.

Table 12

Distribution of negation particles in different moods

Mood	**არ**	**ვერ**	**ნუ**
Indicative	+	+	(+)
Imperative	+	=	+
Conjunctive	+	+	(-)

The semantic function of the particles არ and ვერ concerning the act of communication is also interesting: Although **ვერსად(აც) ვერ** წახვალ (*you cannot go anywhere*), formally should convey the negation of possibility: in a particular context it is used to convey the semantics of a categorical prohibition: არ წახვალ (*you will not go anywhere*).

Furthermore, this semantics can be seen more clearly in the idiomatic expression **ფეხსაც ვერ მოიცვლი,** which, despite the presence of the particle **ვერ,** expresses a clear prohibitive - „*You will not change your foot under any circumstances*“ = I forbid you to move from the spot. So, the particle ნუ can express a categorical negation in the present tense if the action has already begun. In such a case, არ გააკეთო (NegPart არ +Imperative) and ნუ აკეთებ (NegPart ნუ+Present) convey the same thing functionally and semantically *don‘t do it*.

The three-part system of negative particles (არ *ar***,** ვერ *ver***,** ნუ *nu*) presented in the Georgian language corresponds to the two-part system in Magrelian: **ვა(რ)** and ნუ. The particle **ვა(რ)** in Megrelian conveys both functions of არ and ვერ particles in Georgian. ვა is not an independent element and is not written separately, it is attached to the verbal form and creates synthetic morphological forms of the negative verb. Writing the negation particle ვა together with the verb is also facilitated by the fact that it is included as an infix in verb forms which has a complex preverbs: **დოთ**-ვა-**დო-ხოდუ** *dot-**va**-do-doxu* (**დოთე**-ვა-**დო-ხოდუნ** *dote-**va**-do-doxun*) *he/she/it does not sit down „**არ ჯდება**“* (Khubua, 1942, p. 744).

With the forms of potentialis, the ვა particle corresponds to the Georgian ვერ particle in its function and conveys the negation of the possibility. The fact that Megrelian does not and cannot differentiate between არ and ვერ particles is compensated for in the verb form (Kiria et al., 2015, p. 623).

Cf.:

ვა-ჭარუნს *he/she/it does not write* (Pres., Act.) vs **ვე-ეჭარე<ვა-იჭარ** *it cannot be written* (Fut., Pass.), **ვა-აჭარე** *he/she/it cannot write* (Fut., Act.)

For the statistical analysis of Megrelian texts, the negation particle ნუ is irrelevant, as it is always written separately. Therefore, we will not discuss it here and will

instead return to the main issue: the problem that arises during tokenization due to the negation particle ვა being written with the verb.

As is known, functional elements stand out with the highest frequency in the statistical processing of data. Among them is the negation particle არ**.** The table shows the highest frequency words in the Georgian National Corpus diachronically:

Table 13

Tokens with the highest frequency in comparison

Old Georgian	*Middle Georgian*	*Modern Georgian*	*GRC*
და	და	და	და
იგი	ყოფნა	არ	ეს
ყოფ(ნ)ა	ის	ყოფნა	ის
რომელი	რა	ის	რომ
ის	არ	რომ	არ
არ	ეს	ეს	ყოფნა
რამეთუ	მისი	რა	რომელი
ყოველი	მე	მე	რა

As the statistical analysis of parallel texts of Georgian and Megrelian proverbs revealed, the Georgian negation particle არ takes the second place in terms of frequency, and in the statistical analysis of Megrelian texts, the particle conveying the category of negation is not found separately at all (Jgharkava, 2024, p. 25). (See Table 14)

The same issue arises with the **Rustaveli corpus**: it is impossible to accurately measure the statistics of the negation particle **ვა** in the Megrelian translation of the epic. Unlike in Georgian, the Megrelian negation particle **ვა** forms a token only when combined with the verb, resulting in an inaccurate count from the perspective of statistical processing of negation.

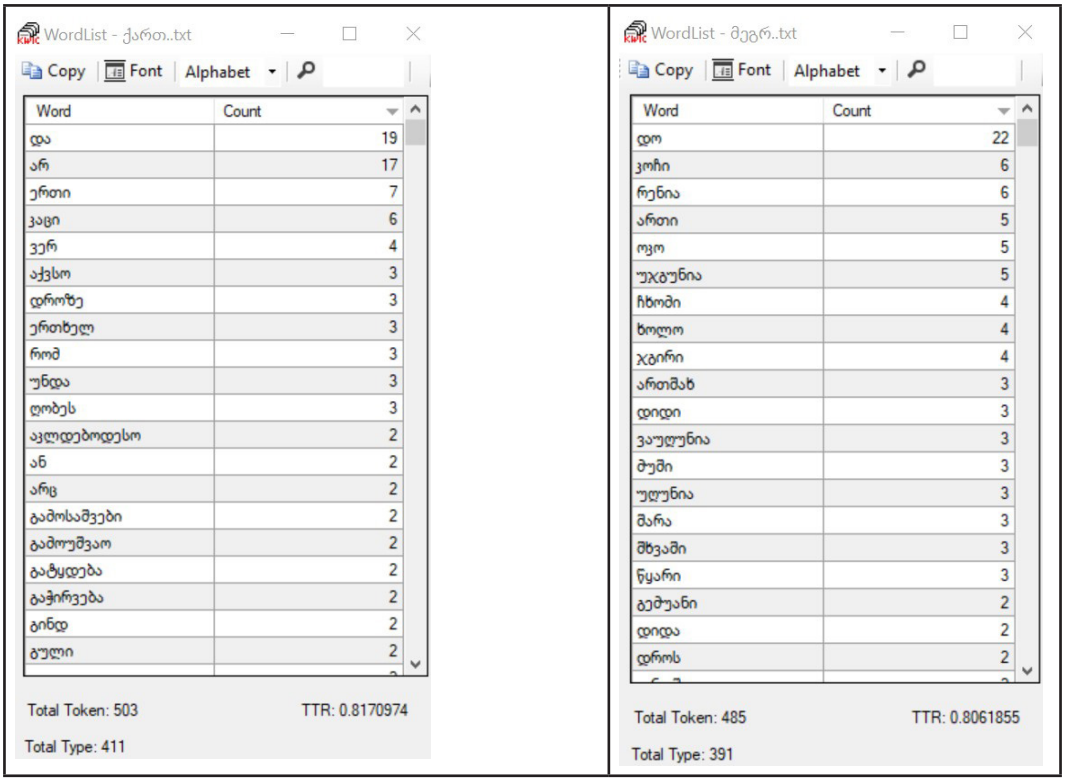
We present the mentioned problem on the example of the stanza 3.90:

რა **პასუხი არა გასცა**, მონა გარე შემობრუნდა,
როსტანს ჰკადრა: „შემიტყვია, იმას თქვენი **არა უნდა**;
თვალნი მზეებრ გამირეტდეს, გული მეტად შემიძრწუნდა,
ვერ ვასმინე საუბარი, მით დავყოვნე ხანი მუნ, და-“.

*Since he did not **answer**, the slave went back,*
*He said to Rosten: „I understood that he **will** listen to **nothing** more from you;*
My eyes were dazzled as by the sun; my heart was sorely troubled.

Table 14

Comparison of frequency of the Georgian and Megrelian proverbs



*I **could not make him** the conversation **to hear**, so I stayed there for a long time.*

Translation by Kaka Jvania (3.92):

მონას მუთუნქ ვაგურთუნი : დირთ დო უწუ ხენწიფეს
თაშ შევატყვი ი კოს თქვანი : მუთუნ ვაკო შხვას წუხენსჷ
გურქ შემეწუხ თოლქ მიდამირთ : ვაბხვალამუქ მა თენერსჷ
ოშ უწუენ ართ ვარჩქილე : შურო პასუხის ვერზენსჷ.
*When the slave **could do nothing**, he returned and said to the king.*
*I understand it this way, he **doesn‘t want anything** from you, something else was worrying him.*
*My heart was troubled, my eyes darkened: I‘ve **never experienced** anything like it.*
*If you‘ve told him a hundred times, he **doesn‘t** even **understand**: he **doesn‘t give** any answer at all.*

Translation by Gedevan Shanava (3.90):

მონას **მუთუნქ ვაგმაღინუ** მუკირთ დო თეში ქმორთუა,
ხენწფე თქვანი ის **ვარჩქილე** მონაქ ენა თაში თქუა,
თოლქ ქამისკიდ თიშ ჯინაშა, გურქუ დახე წამირთუა,
ვაგმაგონუ ნარაგადქ ადრეთ თიშენი **ვამმართუა**.

*The slave **could not find out anything**, turned out and came back.*
*King, he **does not understand** you - The slave said it like this.*
I could not take my eyes off, my heart was almost broken.
***I couldn't make him listen** to what was said, and that's why I **couldn't come back** in time.*

The statistics of the 10 most frequently occurring words in comparison are listed in the Table 14:

Table 15
Most frequently occurring words in comparison

Source text	*Translation by Jvania*	*Translation by Shanava*
და (25)	დო (23)	დო (23)
რად (9)	თეში (8)	მაფაქ (9)
რა (9)	ჩქიმი (7)	მუს (7)
არ (9)	მუში (7)	რდუ (6)
ესე (8)	ხენწიფექ (6)	მა (6)
მეფე (7)	მა (6)	ჩქიმ (5)
იყო (7)	ის (6)	უწუ (5)
იგი (7)	ვარ (6)	რე (5)
თუ (7)	გური (6)	მუთუნ (5)
ვერ (7)	აფუ (6)	კოჩი (5)

Cf.: Mapping the statistical processing of the chapter III in Voyant:

Figure 12
Georgian Text in Voyant

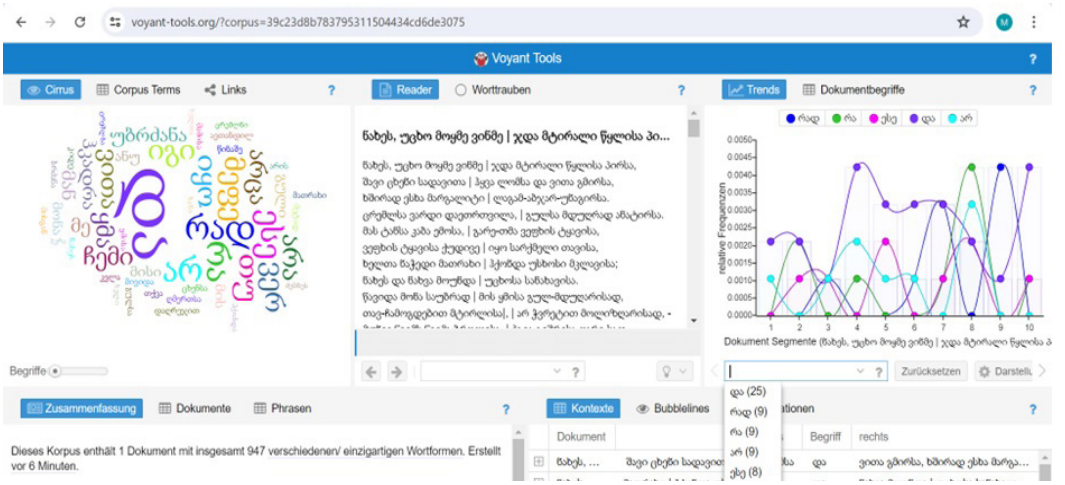
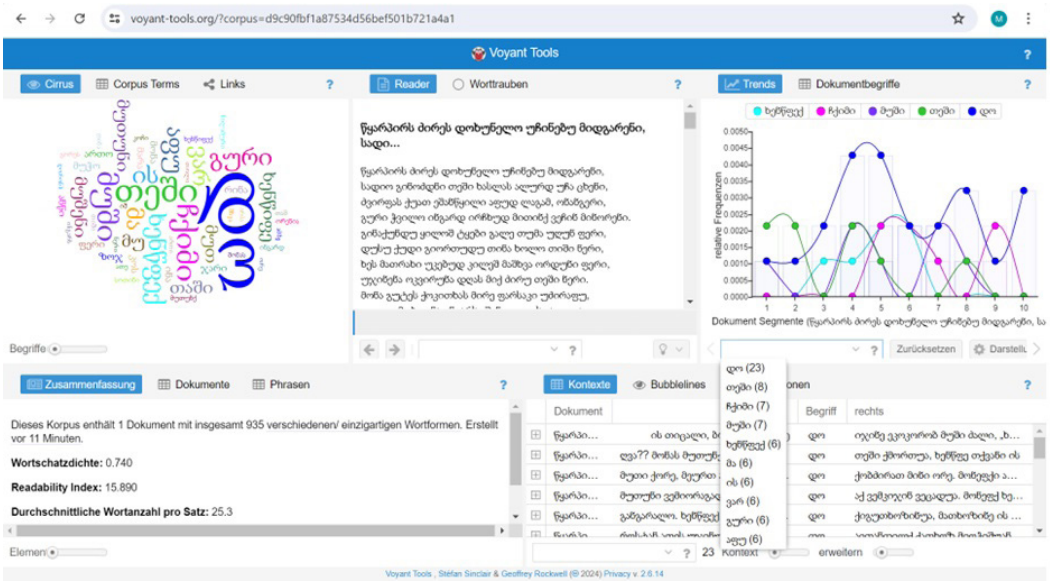


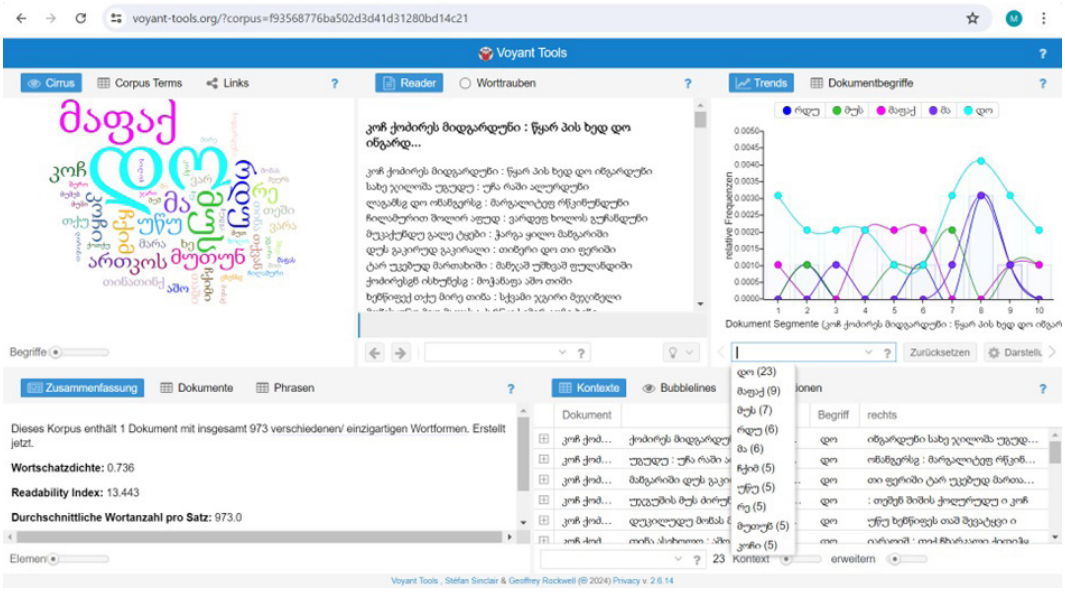
Figure 13

Megrelian translation by Jvania in Voyant



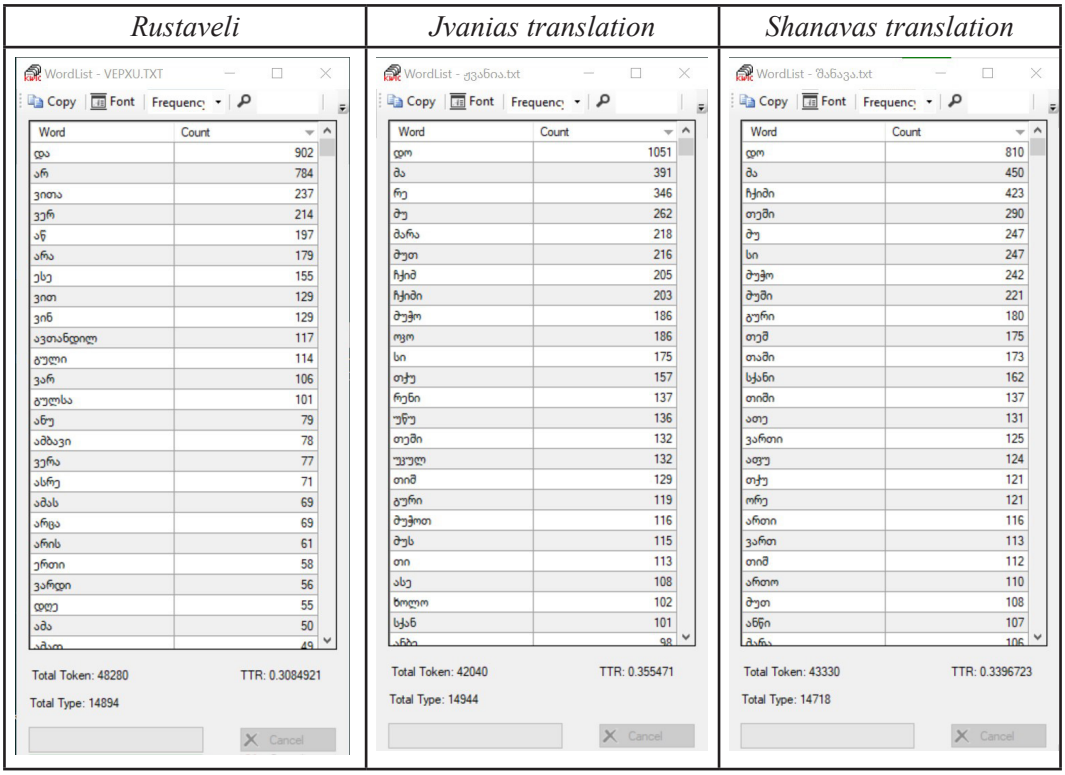
Figuer 14

Megrelian translation by Shanava in Voyant



Comparison of word frequency in source and target text of the epic:

Figure 15
Comparison of word frequency in source and target text in KWIC



In these examples, it is also interesting that the particle **ვა** *va-* sometimes preceded by the indefinite pronoun **მუთუნ** *mutun* 'someone'. In combination with the negative verb (verb with the negative particle), the indefinite pronoun **მუთუნ** becomes a negative pronoun. The negative pronouns are present in Megrelian and Laz (**მითა** *mita* 'nobody', **მუთა** *muta* 'nothing'), but they are less productive. It is also interesting to note that the verb in combination with the negative pronouns **მითა** *mita* and **მუთა** *muta* is always formed in the positive form (e.g. **mita murtumu** 'no one came' and not **mitas vamurthumu**). From the point of view of these two different ways of conveying the negation, it will be interesting to check statistically which translator chooses which strategy. This requires an annotated corpus of Megrelian, so that a precise statistical processing of negative verbs would be possible despite the peculiarity of the negation category in Megrelian.

RESULTS

The multilingual parallel corpus *Rustaveli goes digital*, which currently contains 32 parallel translations of the full text of the epic in 20 languages (Georgian, Ger-

man, English, Spanish, French, Italian, Turkish, Azerbaijani, Kyrgyz, Russian, Belarusian, Ukrainian, Greek, Arabic, Persian, Armenian, Ossetian, Lithuanian, Mingrelian, Svan) is an important digital resource for translation studies. Although nowadays, there are many different ready-made tools that are successfully used in linguistics for statistical analysis, the data processing of texts can still be very inaccurate without considering the grammatical characteristics of the languages. As shown in the article, it is necessary to develop a suitable tool for each language to be able to carry out a cross-linguistic analysis in a parallel corpus such as *Rustaveli goes digital*.

To use the multilingual parallel corpus for multidisciplinary research, it is necessary to incorporate a two-level statistical data processing: at the low level, statistical processing of the text must consider the linguistic features because inaccurate statistical results can lead to wrong statistics, and thus to wrong conclusions. In this way, we will get the accurate statistical data obtained at the low level of the statistical data processing respective languages as a result, which we can compare at the second level with the statistical results of the other languages that were also statistically processed at the low level.

ABBREVIATIONS:

AbNC	Abkhazian National Corpus	Nom	nominative
Act.	active	Pass.	passive
DO	direct object	Pers.	person
GNC	Georgian National Corpus	Pl	plural
F	feminine	PossPron	possessive pronoun
Fut.	future	Pres.	present
KWIC	Key Word in Context	S	subject
M	masculine	Sg	singular
N	neuter	TAM	tempus, aspect, mood
NP	nominal phrase	Vtr.	transitive verb
N	noun	Vint.	intransitive verb

REFERENCES

Chkheidze, M., & Taktakishvili, L. (2016). Vefxistqaosnis gamocemata bibliografia: 1712–2015 [Bibliography of "The Knight in the Panther's Skin": 1712–2015]. Tbilisi: Sezani.

Chumburidze, Z. (1970). Uarkhopiti nats'ilakebi kartulshi da mati khmarebis st'iluri taviseburebani [Negation particles in Georgian and stylistic features of their use]. School and Life, 2(17), 41-46.

Gippert, J. (2024). A zoological riddle from Medieval Georgia. In F. Mühlfried (Ed.), Languages and Cultures of the Caucasus: A Festschrift for Kevin Tuite (pp. 85–103). Wiesbaden: Reichert.

Jgharkava, G. (2024). Kartvelur enata andazebis kvleva ts'ifrul ep'ok'ashi – teoretiuli da teknologiuri charcho [Digital processing of proverbs in Kartvelian languages – theoretical and technological framework]. Millennium, 2, 5-43. https://doi.org/10.62235/mln.2.2024.7991

Koller, W. (1992). Einführung in die Übersetzungswissenschaft. Heidelberg.

Khubua, M. (1942). Uarkhopis nats'ilaki va megrulshi [The negation particle va in Megrelian]. Moambe (Journal of the Academy of Science of Georgia), 3(7), 743–745.

Kurdadze, R., Lomia, M., Margiani, K., & Tchumburidze, N. (2022). Uarkhopa kategorias kartvelur enebshi [The category of negation in the Kartvelian languages]. Tbilisi: Universali.

Machavariani, G. (2002). Kartvelur enata shedarebiti gramat'ika [The comparative grammar of the Kartvelian languages]. Kartvelologische Bibliothek, 9. Tbilisi: Publishing house of Tbilisi Staatliche University.

Sinz, J. (2017). Translationstheorien: Die Äquivalenz nach Werner Koller und die Adäquatheit in der Skopostheorie. Grin Verlag.

Tandaschwili, M. (2022). Dighitaluri rustvelologia [Digital Rustvelology]. Tbilisi: Iverioni.

Tandaschwili, M., & Kamarauli, M. (2023). Vepkhistqaosnis stiluri sakhelebis targmani (mts'eris p'irveli maghals) [Translating the stylistic devices of "The Knight in the Panther's Skin" (The case of the 'sun')]. Digital Kartvelology, 2, 82-105.

The Abkhaz National Corpus. (n.d.). *CLARINO Bergen Center.* https://clarino.uib.no/abnc/page