Parallel Corpora and Implementation Possibilities in Multilingual Education (Georgian-Abkhazian Parallel Corpus of "The Knight in the Panther's skin")

Leila Avidzba
Ivane Javakhishvili Tbilisi State ,
University, Georgia
Email: leila.avidzba@tsu.ge

Mariam Kamarauli
Centre for the Study of Manuscript
Cultures, University of Hamburg,
Germany
Email: Ma.Kamarauli@em.uni-frankfurt.de

Manana Tandashvili
Imstitute of Empirical Linguistics,
Goethe Frankfurt University, Germany
Email: Tandaschwili@em.uni-frankfurt.de

Leila Avidzba

Ivane Javakhishvili Tbilisi State University, Georgia

Mariam Kamarauli

Centre for the Study of Manuscript Cultures, University of Hamburg, Germany

Manana Tandashvili

Imstitute of Empirical Linguistics, Goethe Frankfurt University, Germany

Parallel Corpora and Implementation Possibilities in Multilingual Education (Georgian-Abkhazian Parallel Corpus of "The Knight in the Panther's skin")

ABSTRACT

The epic The Knight in the Panther's Skin by Shota Rustaveli is a well-known literary work inside and outside of Georgia, which has been translated into more than 50 languages. It comes as no surprise that the epic offers many topics for intradisciplinary as well as interdisciplinary researches, one of which will be discussed here, namely aphorisms found in the translations of the epic. The translations in question are in Abkhazian, of which two exist: one translated by Dimitri Gulia (published in 1941) and the other by Mushni Lasuria (published in 1978). For a methodological analysis, the parallel corpus of the Abkhazian translations GeAbCo (https://geabco.com/) was created. This paper will explain the planning and conception of the corpus as well as show a first attempt of an analysis considering the aphorisms found, whereat even the question about the actual number of aphorisms in the epic is a research question itself.

Keywords: Corpora, aphorisms, translation, abkhazian language

Introduction

The Knight in the Panther's Skin by Shota Rustaveli is the most significant epos of the Georgian intangible cultural heritage. The poem, which was created in the 12th century, is represented by more than 160 different manuscripts and is one of the most crucial components of defining the identity of the Georgian nation. 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 research of this unique literary work with modern methods is not only a challenge for the Kartvelology of the 21st century but will also contribute to the scientific research of Georgian intangible cultural heritage and to 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.

Scientific research of the poem begun in the 18th century, when King Vakhtang VI added a first scientific analysis to the first printed book from 1712. This formed the basis for further research of the epos, which gradually developed into a separate field of Kartvelology - **Rustvelology**.

The history of Rustvelology covers more than 3 centuries and can be divided into several developmental 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 latter implies, on the one hand, the creation of digital resources - **big data** in Rustvelology (the creation of a parallel corpus of poem manuscripts and printed editions, as well as a multilingual parallel corpus of epos translations in different languages), and on the other hand, the creation of a methodological framework that will allow us to effectively use this unique multilingual parallel corpus in research and in education.

Georgian-Abkhazian parallel corpus of *The Knight in the Panther's Skin*

The epos *The Knight in the Panther's Skin* has been translated into 56 languages with the number of foreign language editions of the epos reaching up to 350. Among them are two Abkhazian translations (by Dimitri Gulia und by Mushni Lasuria), which have been published 7 times.

The Georgian-Abkhazian parallel corpus of *The Knight in the Panther's Skin* (GeAbCo) was created within the framework of the project financed by Shota Rustaveli National Science

Foundation of Georgia and plays a significant role for various reasons:

1. Unlike the Georgian language, the Abkhazian language, as a state language, cannot really meet the modern standards of language technology, which significantly hinders the full functioning of the Abkhazian language as a state language both in state institutions and in the educational field;

2. The creation of the Georgian-Abkhazian parallel corpus of *The Knight in the Panther's Skin* will pave the way and form the basis for the technologization process of the Abkhazian language;

3. The Georgian-Abkhazian parallel corpus will allow us to develop a new didactic teaching model on the basis of the parallel corpus, which can be effectively implemented in multilingual education.

Abkhazian translations of the epos

The first Abkhazian translation of *The Knight in the Panther's Skin* translated by **Dimitri Gulia** was published in 1941, using the Georgian alphabet. He translated the long version of the poem, which contains 63 chapters (1664 stanzas). While the first two editions of Gulia's translation (1941 and 1953) were printed in Georgian script, the subsequent editions (1959 and 1984) of same translation were printed in an Abkhazian script based on the Cyrillic script.

The second Abkhazian translation of the poem was done by **Mushni Lasuria**. Excerpts of the poem were printed in the magazine "Alashara" in 1972, and the complete translation of the poem was published as a book in 1978. Unlike Gulia's translation, Lasuria's translation is based on a shorter version of the poem (61 chapters, 1587 stanzas). For his efforts of translating *The Knight in the Panther's Skin* into Abkhazian, Mushni Lasuria was awarded the Rustaveli Prize in 1981.

In the Georgian-Abkhazian parallel corpus of *The Knight in the Panther's Skin* (GeAbCo), both Abkhazian translations by D. Gulia and by M. Lasuria are presented.

The Georgian-Abkhazian parallel corpus *The Knight in the Panther's Skin* (GeAbCo)

The Georgian-Abkhazian parallel corpus *The Knight in the Panther's Skin* is a bilingual corpus that includes the original text of the poem and both Abkhazian translations of the poem (by Dimitri Gulia and by Mushni Lasuria). The Georgian source text of the corpus is the published version of the poem by Akaki Shanidze in 1975. This version of the poem comprises

the prologue and the main part (63 chapters, including an epilogue) and consists of 1669 stanzas.

The construction of the parallel corpus was executed in several stages:

1) digitization of texts,

2) alignment and

3) tagging.

Stage I: Digitization of texts

In the first stage, the Abkhazian translations were digitized and parallelized with the original in EXCEL format. The texts were parallelized both according to chapters and verses.
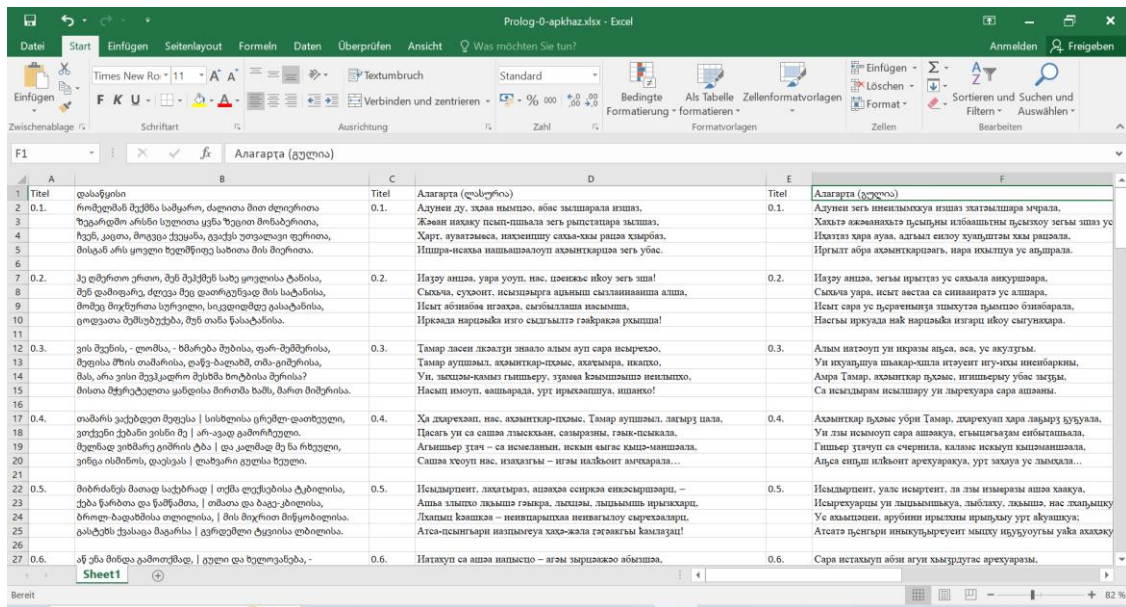

Figure 1: Excel file with both Abkhazian translations parallelized to the Georgian text

Here we encountered two types of problems: a) graphemic and b) structural. In particular:

a) the last two printed versions of Gulia's translation are published in Abkhazian script, in which the old orthography is used, while Lasuria's translation follows the new orthography.

Gulia	Lasuris	Geor. Translit.	Lat. Translit.	IPA
Гу гу	Гə гə	გ°	g°	g°
Ҕу ҕу	Ҕə ҕə	ღ°	γ°	γ°
Ку ку	Кə кə	კ°	ķ°	ķ'°
Қу қу	Қə қə	ქ°	k°	k°
Ҟу ҟу	Ҟə ҟə	ყ°	q̇°	q'°
Ху ху	Хə хə	ჰ°	x°	x°

L. Avidzba, M. Kamarauli, M. Tandashvili, *Parallel Corpora and Implementation*
Possibilities in Multilingual Education

23, 2023

pp. 88-101

This graphic difference necessitated the need for an additional converter in the next step (the alignment process).

b) Abkhazian translations, as already mentioned above, differ from each other in terms of volume: Dimitri Gulia has translated the long version of the poem (63 chapters, 1664 stanzas), while Mushni Lasuria used the short version (61 chapters, 1587 stanzas) as his source text. In addition, unlike in the source text, the epilogue (АЛҐАРҬА) is separated in both Abkhazian translations from the main text as a chapter on its own, while the Georgian text includes it in its 63 chapters. Accordingly, only Gulia's translation (translation of the long version of the epos) is parallelized completely, without any chapters missing.

At the last stage of digitizing texts, the parallelized texts were implemented in the corpus:
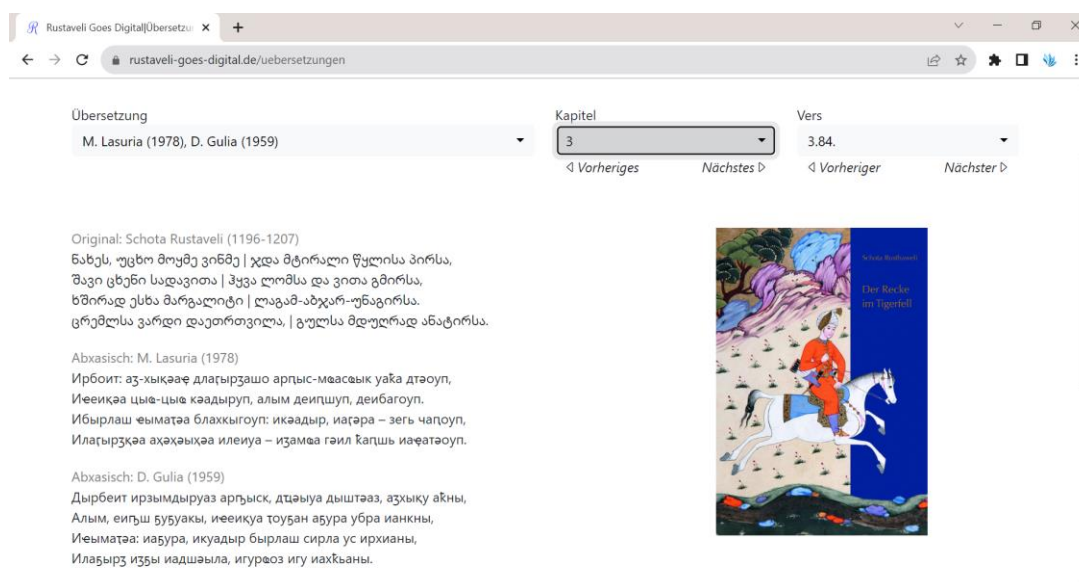

Figure 2: The parallel corpus "Rustaveli goes digital" with the Georgian source text and both Abkhazian translations

Stage II: Alignment

The main task of the second stage was alignment. Alignment is an important step to make parallel corpora linguistically usable. Alignment can be done at three different levels:

a) word level (lexical alignment),

b) phrase level (morphosyntactic alignment) and

c) sentence level (syntactic-semantic alignment), on which we will focus here.

The alignment process can be done manually or automatically; the former normally presupposes the latter. Naturally, the question arises whether the alignment for this project can

be automatized or has to be done manually. In order to answer this question, several aspects of sentence alignment need to be considered and defined.

Tschorn & Lüdelig describe in their conference paper "Morphological knowledge and alignment of English-German parallel corpora" (2003, 1) a **sentence level alignment**, which maps groups of L1-sentences (sentences from the bilingual parallel text) to corresponding groups of L2-sentences (sentences from the translated text) and call them "beads". They write further that "an alignment can be viewed as a sequence of beads that covers the entire parallel text. While most beads usually express the correspondence between a single L1-sentence and a single L2-sentence, other types of beads arise when sentences are split, merged, deleted, added or changed in order by the translator" (Tschorn & Lüdelig 2003, 1). In the case of poetic texts, these problems are less common: stanza Xn of L1 usually corresponds to stanza Yn of L2. For example, *if (in Xn of L1, n=15), then (in Yn of L2, n=15)* (as in Figure 1). However, even in poetic texts there are cases of switching sentences within stanzas or have some other irregularities.

A special method of analysis contained in the sentence level alignment, namely **lexical distance functions**, is used for alignment at the sentence level, which "try to find corresponding words in the two halves of the parallel text" (Tschorn & Lüdelig, 2003, 2). There have been suggestions *for automatic extraction of an ad hoc dictionary* from parallel texts that have been aligned and those that have not, as well as during alignment (Tschorn & Lüdelig 2003, 3). These methods often rely on a concept of co-occurring frequency and assume that translation links only exist between single words. However, it is well known that words in texts have a Zipfian distribution, which means that texts primarily contain a small number of very common words (such articles and prepositions) and a larger number of rare words (Tschorn & Lüdelig, 2003, 3; Baayen 2001). Highly frequent words are used in almost every sentence, which makes them practically useless for determining how two sections of a parallel text are related. Rare words are a much better measure of relatedness, but co-occurrence counting methods struggle capturing them (Tschorn & Lüdelig, 2003, 3).

Aside from lexical distance functions, Tschorn & Lüdelig (2003, 3) also describe **dictionary-based distance functions**, which rely on an existing machine-readable dictionary. Due to the fact that one word can have multiple translations and that multi-word units can be considered, dictionary-based distance functions outperform distance functions that rely on less knowledge and depend on the quality and the scope of the used dictionary (Tschorn & Lüdelig, 2003, 3). But as Tschorn & Lüdelig (2003, 3) quote Schultink (1961, 113), when we have "in

principle unlimited number of new formations", any dictionary will be lacking because of morphological productivity, irrelevant of its size or quality.

This becomes particularly true in languages like Georgian or Abkhazian languages, which have one the most complex inflection system for nouns and verbs. The declension system of Modern Georgian includes 7 cases and, taking into account postpositions and particles, produces up to 700 inflectional forms. The Georgian verb is polypersonal and conjugation paradigms consist of several thousand inflectional forms. Because of this polypersonality, the Georgian language is considered a radical pro-drop language, omitting subject and object pronouns. Abkhazian has no declension system, but it has enough other linguistic phenomena: different types of number (singular, dual, plural, associative plural, INCLUSIVE vs. EXCLUSIVE, +HUMAN vs. -HUMAN (Hewitt, 1979, 152, 156-158)), an elaborate category of definiteness (definiteness vs. indefiniteness), which is marked in the nouns and many more. Complex verbal structures are also characteristic for the Abkhazian language, which can mark various grammatical categories such as non-volitive modality, reciprocity, reflexivity, sociativity etc. additional to person and number markers.

Therefore, at this stage, the possibility of automatic alignment in the Georgian-Abkhazian parallel corpus is excluded yet, as online dictionaries and automatic analysis and synthesis programs are needed. While it is true, that several English-Abkhazian dictionaries but because the Abkhazian language is characterized by complex morpho-phonematic processes, an automatic analysis of Abkhazian is still at the beginning stage. Therefore, the alignment in the corpus was done manually by tagging.

Stage III. Tagging

The manual tagging was executed classified as follows:

1. Tagging of thematic lexical groups in the poem (social status, astronomical objects, gemstones, weapons, clothes etc.) and finding their equivalents in the both Abkhazian translations.

2. Tagging of figures conveying literary style (metaphors, similess, hyperboles, trope, etc.) in the poem and finding their equivalents in the both Abkhazian translations.

This tagging revealed similarities and differences between the source text and the Abkhazian translations. For example, the word for sun ∂ซ$\jmath$ is confirmed 309 times in the poem with different meanings and functions:

a) the sun as an astronomical object;

b) the sun as an stylistic device (for comparisons, metaphors, epithets);

c) the sun as a component of a compound;

d) the sun as a component of a swearing.

In the Abkhazian translations, only the first, second and third type is attested but the fourth is not, cf. ST 494 and ST 563[1]:
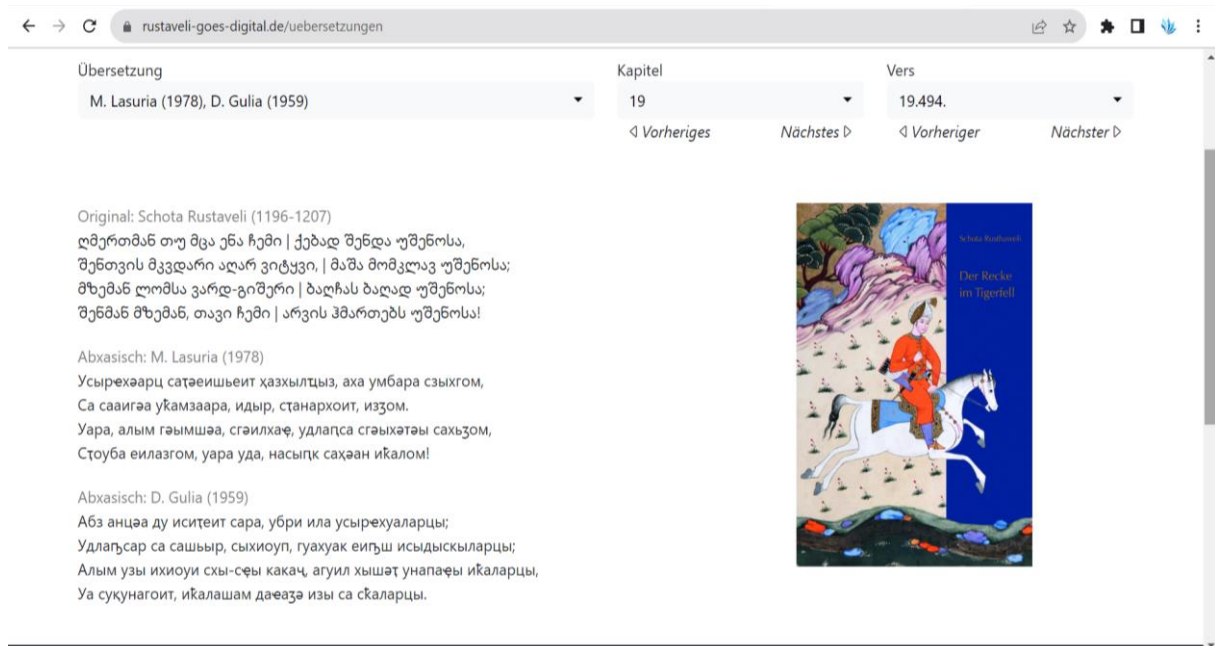

Figure 3: Chapter 19, stanza 494

(1a) GEO

შენმან	მზემან,	თავი	ჩემი	არვის
šen-man	***mze-man***	*tav-i*	*čem-i*	*arvi-s*
your-ERG.SG	sun-ERG.SG	head-NOM.SG	my- NOM.SG	noone-DAT.SG

ჰმართებს	უშენო-ს-ა!
hmartebs	*ušenosa!*
own.s3sg.o3sg.PRES	aside from you-DAT.SG-EXT.V

'*By thy sun, my self pertains to none save thee.*'

---

[1] We thank Prof. George Hewitt and his wife Ms. Zaira Khoba for their assistance in the process of analyzing the Abkhaz sentence examples.

(1b) ᴀʙᴋʜ (D.G.)

Уа,	суқунагоит,	иқалашам
ua	*s-u-ḳu-na-go-uo-i̇*	*i-q̇a-la-ša-m*
you.MASC	I-you.MASC-befit-DYN-FIN.PRES	it-PREV-happen-ꜰᴜᴛ ɪɪ-not

дaᶜᴇаӡə	изы	ca	cҡаларцы
dač̑ₐa-ӡᵒ	*iz-ə*	*sa*	*s-q̇a-la-r-c-ə*
other-person	him-for	I	I-PREV-become-ᴘᴜʀᴘ

'I am worthy of you, it will not happen if I become someone else's.'

(1c) ᴀʙᴋʜ (M.L.)

Сҭоуба	еилазгом,	уара	уда,
s-touba	*ø-ei-la-z-g-o-m*	*uara*	*u-da*
my-oath	it-PREV-I-break-DYN-not.PRES	you.MASC	you.MASC-without

насыпк	саҳәан	иқалом!
nacəp-ḳ	*s-a-hᵒa-n*	*i-q̇a-l-o-m*
happy.destiny-one	I-it-request-ᴘᴀꜱᴛ.ᴀʙꜱᴏʟᴜᴛᴇ	it-PREV-happen-DYN-not.PRES

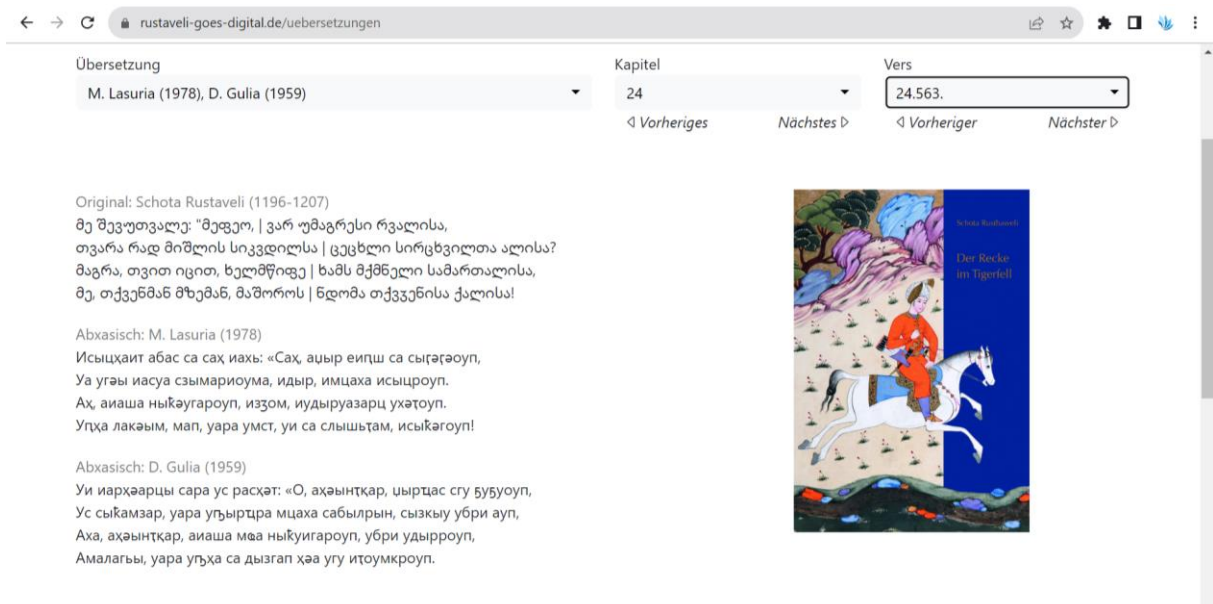'I won't break my oath, I won't ask for any fate other than yours.'

Figure 4: Chapter 24, stanza 563

(2a) GEO

მე,	თქვენმან	მზემან,	მაშორჰოს	ნდომა
me,	*tkven-man*	*mze-man*	*mašoros*	*ndoma*
i.NOM.SG	your.PL-ERG.SG	sun-ERG.SG	distance.S3SG.O1SG.OPT	desire.NOM.SG

თქვენისა	ქალისა!
tkven-is-a	kal-is-a
your.PL-GEN.SG-EXT.V	woman-GEN.SG-EXT.V

'By your sun! I am far from desiring your daughter!'

(2b) АВКН (D.G.)

Амалагьы,	уара	унҧха,	са	дызгап,
amala-gjə	*ua-ra*	*u-pha*	*sa*	*də-z-ga-ṗ*
just-and	you.MASC	your-daughter	I	her-I-take-FUT I

хәа	угу	иҭоумкроуп
hᵒa	*u-gu*	*i-ɸ-to-u-m-ḳ-r-o-u-ṗ*
saying	your.MASC-heart	it-it-in-you.MASC-not-hold-if-it-be-STAT.FIN.PRES

'But don't believe that I will take your daughter.'

(2c) АВКН (M.L.)

Уҧха	лакәым,	маҧ,	уара
u-pha	*l-a-ḳᵒə-m*	*maṗ*	*uara*
your.MASC-daughter	she-be-not.STAT.FIN.PRES	no	you.MASC

умст,	уи	са	слышьҭам,
umsṫ	*ui*	*sa*	*s-lə-šta-m*
you.MASC-not-die-AOR-FIN	that.one	I	I-her-pursue-not.FIN.STAT.PRES

'Your daughter, no, you don't die, I won't chase her'

In these two examples above, the swearing was not translated into Abkhazian. However, in other stanzas of the poem, swearings characteristic of the Abkhazian language is used:
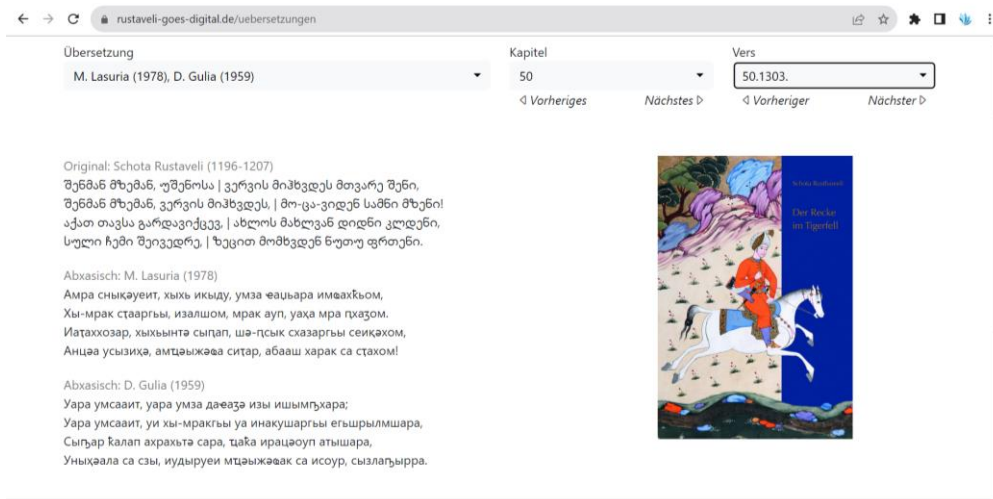
Figure 5: Chapter 50, stanza 1303

(3a) GEO

შენმან	მზემან,	უშენოსა	ვერვი-ს
šen-man	***mze-man***	*ušenosa!*	vervis
your-ERG.SG	sun-ERG.SG	aside from you-DAT.SG-EXT.V	no one-DAT.SG

მიჰხვდეს	მთვარე	შენ-ო
mihxvdes	*mtvare*	*šen-i*
understand.S3SG.OPT	moon.NOM.SG	your-NOM.SG

'By the sun, without you no one can see your moon!'

(3b) АВКН (D.G.)

Ухаҭа	умсааит,	уара	умза	даҷаӡə
u-xa-ta	*umsaait*	*uara*	*umza*	*dač̣ₐ-ʒᵒ*
your.MASC-self	you.MASC-not.die-SUBJ	you.masc	your.MASC-moon	other-person

изы	ишымҧхара
iz-ə	*i-šᵒ-m-pxa-ra*
him-for	it-that-not-shine-NON.FIN.FUT I

'I swear to you, (don't die yourself) your moon won't shine for someone else'

(3c) АВКН (M.L.)

Амра	сныкәуеит,		хы-хь	икыду,
a-mra	*s-ɸ-nə-ku-ue-iṫ*		*xəxj*	*i-ḳəd-u*
the-sun	I-it-PREV.by-swear-DYN-FIN.PRES		above	which-suspended-NON.FIN.STAT.PRES

умза	ҽацьара	имѳахҟьом
u-mza	*č̣ə-ǯara*	*imwaxq̇jom*
your.MASC-moon	other-place	it-PREV.path-PREV.off-fall-DYN-not.FIN.PRES

'I swear by the sun, when it's hanging high, your moon won't turn off the road'

Repetition of poetic formula

In literary works, there are often poetic expressions, which the authors repeat in the text. This is one of the features of GeAbCo: any repetitions can be found automatically as shown in the figure below.
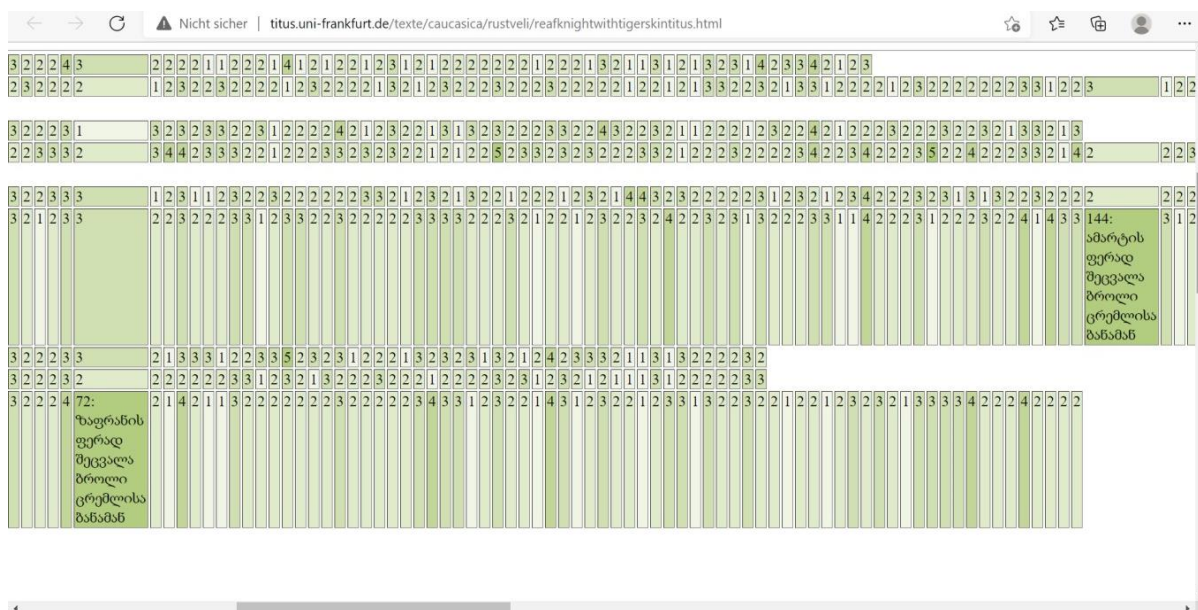

Figure 6: Repetitions found in the poem

ამარტის   [ფერად შეცვალა ბროლი ცრემლისა ბანამან.] (266)
ზაფრანის [ფერად შეცვალა ბროლი ცრემლისა ბანამან] (358)
Бырлыш хаҵы алаҳырӡ шы қаруа џшрала ишьақунаргылеит; (D.G.266)
Бырлаш хаҵы, амраш еисоз, лаҳырӡ иарѳажьт ашафран еиџш; (D.G.358)

Илеиуа рыла�224ырз̌қәа еихсыгьуам, бырлашушәа икаҧсоит, (M.L.266)
Абри нахыс ахаҿбырлаш алагырз̌қәа инадырⱷеижьит, (M.L.358)

As this example shows, a high-quality poetic formula is used in Georgian: in a sentence consisting of 6 words, 5 words are repeated. If we take into account that amart (precious stone) and saffron (plant) refer to the same color - red (here we should refer to our article), then we can say that according to the lexical content, we have a 6-component artistic formula here. As for Abkhazian, none of the translators has even attempted to convey this rare case of Rustaveli's poetic ability.

Idiomatic expressions

Rustaveli often uses idiomatic expressions in the poem. For example, the idiomatic expression *თხა და მგელი ერთად ჯოვს txa da mgeli ertad ჯovs* 'the goat and the wolf graze together', which stands for peace, mutual respect and love.

The exact same idiomatic expression is confirmed in Abkhazian, which was used by both translators:

Gulia 1658 /1664
Рхәынҭқарраҟуарҿ аҵьмеи қуҵьмеи џьара еицхәуан еицәшәазомызт.
In their kingdoms, goats and wolves grazed together and were not afraid of each other.
аҵьмеи (*goat and*) қуҵьмеи (*wolf*) џьара (somewhere) еицхәуан (*grazed together*)

Lasuria 1582 /1664
Убарҭ рахраҿ абгеи аҵьмеи еицыхәуан уа есқьынгьы.
In their kingdom, the wolf and the goat grazed together there all the time.
абгеи (wolf and) аҵьмеи (goat) еицыхәуан (*grazed together*)

Both translators use a word by word translation of this passage. Neither Gulia nor Lasuria use any functional equivalent for translating the above-mentioned idiomatic expression into Abkhazian.

The parallels that exist between the original and the Abkhazian translation of *The Knight in the Panther's Skin* make the Georgian-Abkhazian parallel corpus of the poem significantly interesting not only linguistically but also in terms of content. This corpus can benefit not only scientists and linguists but also teachers and other interdisciplinary scholars.

ABBREVIATIONS

1	1st person	EXT.V	extensional vowel	OPT	optative
3	3rd person	FIN	finite	PRES	present
ABS	absolute	FUT I/II	future I/II	PREV	preverb
AOR	aorist	GEN	genitive	PURP	purposive
DAT	dative	MASC	masuline	S	subject
DYN	dynamic	NOM	nominative	STAT	stative
ERG	ergative	O	object	SUBJ	subjunctive

References

Baayen, R. H. (2001). *Word frequency distributions* (Vol. 18). Springer Science & Business Media.

Hewitt, B. George. (1979). Abkhaz. (Lingua Descriptive Studies, 2.) Amsterdam: North-Holland.

Rustaveli (Abkhazian). 1984 - Шота Руставели. Абжьас-цәа зшәу / Еитеигеит Д. Гулиа. _ Акуа: Апхәынтшәкутыжьырта, 1959 (Абжьасцәа зшәу / Гулиа Д. Ивымтакуа реизга. _ Акуа: Алашара).

Rustaveli (Russian). Руставели, Шота. Абжьас-цәа зшәу / Акыртшәахьтә еитеигеит М. Лашәриа. Акуа: Алашара, 1978.

Schultink, H. (1961). "Produktiviteit als morfologisch fenomeen". In: *Forum der letteren* (Vol. 2, No. 1, pp. 110-125).

Tschorn, P. & Lüdeling, A. (2003). *Morphological knowledge and alignment of English-German parallel corpora.* Conference paper. Berlin: Humboldt-Universität zu Berlin, Philosophische Fakultät II (http://dx.doi.org/10.18452/13416)