

Achtung!

Dies ist eine Internet-Sonderausgabe des Aufsatzes
„Towards a Corpus Caucasianum. Building a corpus from unstructured
data“

von Jost Gippert (2015).

Sie sollte nicht zitiert werden. Zitate sind der Originalausgabe in
Jikia, Marika (ed.), *T'ip'ologiuri dziebani / Typological Investigations*
VII,

Tbilisi 2015, 149-165
zu entnehmen.

Attention!

This is a special internet edition of the article
“Towards a Corpus Caucasianum. Building a corpus from unstructured
data”

by Jost Gippert (2015).

It should not be quoted as such. For quotations, please refer to the
original edition in
Jikia, Marika (ed.), *T'ip'ologiuri dziebani / Typological Investigations*
VII,

Tbilisi 2015, 149-165.

Alle Rechte vorbehalten / All rights reserved:

Jost Gippert, Frankfurt 2015

**Towards a Corpus Caucasicum
Building a corpus from unstructured data**

Abstract

The paper deals with the prospects of compiling a corpus of Caucasian languages from the random attestations of linguistic materials of these languages in pre-20th century publications that have been digitised within the Google Books project. It discusses three major problems that are met with in coping with this task, viz. data harvesting and compilation, character encoding, and unification of data. The main issues addressed are a) the accessibility of data on the background of copyright regulations, b) the unreliability of data provided via OCR, c) the lack of language assignment in mixed data, d) gaps in the encoding facilities provided by Unicode, and e) the requirement of a unique script-independent representation of data. The issues under concern are illustrated with examples taken from publications of the 17th to the end of the 19th century.

0. Introduction

The Caucasus has been renowned since Antiquity as an area with an extreme density of languages. Depending on the principles of distinction applied, investigators arrive at between 50 and 70 languages spoken in the area today, many of them with less than 5,000 speakers and many of them endangered of extinction.¹ For only three of the languages, viz. Armenian, Georgian, and Caucasian Albanian, a written standard was developed as early as the 5th c. C.E.,² while a few others were adapted

¹ The map provided in the Endangered Languages Project website (<http://www.endangeredlanguages.com>) gives a clear picture of the amount of endangered languages in the Caucasus area.

² In the case of Caucasian Albanian, the script was given up in the Middle Ages by consequence of the Arab conquest of the South-Eastern Caucasus; the only manuscript remains, concealed in a Georgian palimpsest in St. Catherine's monastery on Mt. Sinai, have recently been deciphered and published for the first time (see Jost Gippert, Wolf-

to using Arabic, Cyrillic, or Latin alphabets in more recent times. For all those languages that have not developed a written standard, linguistic material other than audiovisual recordings can only be found in secondary literature, i.e., grammars, dictionaries, text books, or stray notes provided by travellers, scholars, and other people interested. Collecting these materials with a view to compiling an exhaustive corpus of Caucasian languages is a challenging task indeed. In the present paper I intend to discuss three major problems met with in coping with this task, viz. data compilation, encoding, and unification of data.¹

1. Compilation of data

The attestation of materials from ‘minor’ Caucasian languages starts with the travel report by Evliya Çelebi, an Ottoman writer of the 17th c., who provided the first specimens of Abkhaz, Adyghe (Circassian), and Megrelian in Arabic script (cf. Fig. 1). In the late 18th c., European scholars such as Johann Anton Güldenstädt or Julius Heinrich von Klaproth began to establish the first more comprehensive word-lists, noting their materials down in Latin script (cf. Fig. 2). Since the second half of the 19th c., Russian developed to be the prevailing language of investigation, with authors such as Anton Schiefner, Peter von Uslar, or Adolf Dirr providing the first extensive grammars of languages such as Abkhaz, Chechen, or Udi (cf. Figs. 3–4). By the same time, the first scientific journals appeared that were devoted to the study of the Caucasus, including its peoples and languages (cf. Figs. 5–6).

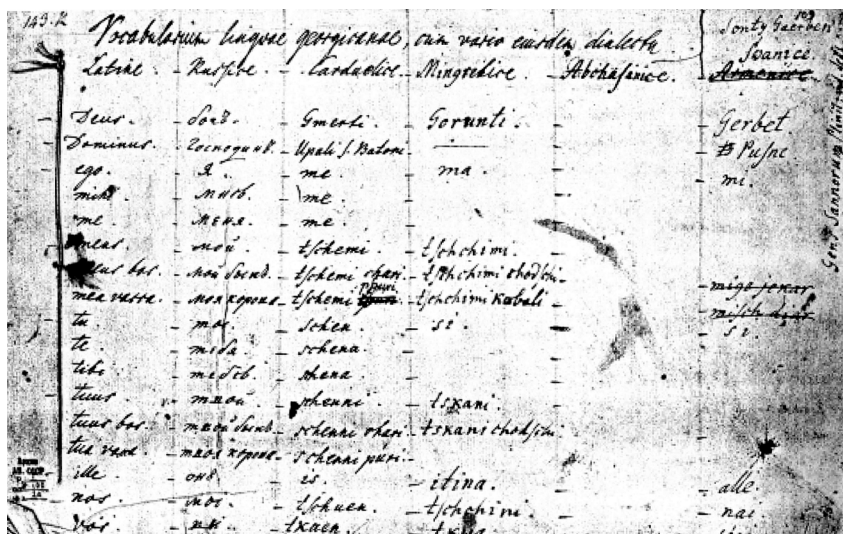
gang Schulze, Zaza Aleksidze, Jean-Pierre Mahé, *The Caucasian Albanian Palimpsests of Mount Sinai*. 2 vols. Turnhout: Brepols 2009; Monumenta Palaeographica Medii Aevi / Series Ibero-Caucasica, 2). The modern successor of Caucasian Albanian is the Udi language (see below).

¹ Some text materials of this type have been electronically prepared for online retrieval in the TITUS (Thesaurus Indogermanischer Text- und Sprachmaterialien) and ARMAZI projects (Caucasian Languages and Cultures: Electronic Documentation); see <http://titus.uni-frankfurt.de/texte/texte2.htm#lazica>. – Audiovisual materials of some of the unwritten languages have been collected and annotated in the ECLinG (Endangered Caucasian Languages in Georgia) and SSGG (The Sociolinguistic Situation of Present-day Georgia) projects funded by the Volkswagen Foundation between 2001 and 2010; they are available for online access at the Language Archive of the Max Planck Institute for Psycholinguistics, Nijmegen, see http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI533677%23 and http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI663243%23.

لسان غریب و عجیب کبازہ
 آق و با اخیبا بشبا خوبا فبا بزبا عبا ژربا زوبا اق زوبا و بانوبا
 ۱۲ ۱۱ ۱۰ ۹ ۸ ۷ ۶ ۵ ۴ ۳ ۲ ۱
 و امله اوچیت اوستویه اوقل اوچین ازیش بیجا اب انخوش
 کل کیت اطلر قلق کتده اوغلان کیده بر عورت
 بیجا ام اوز مجور ویت ازیش سیره ازیرویت اوزره یودیز و ا و خا قسی
 کیم نبود کمنزک اوغلان بن بلیروز سن بلیزک جانم کوزم
 سیره ازیرویه سیر ازیرویت ایش ازیرویه و ازیرویه اشیو و ازیو
 بن بلیزک ناکه یتر بویله نین سونلرک صایقلمین نه سونلر یوزر
 سیره ازیرویه اوزره یودیز و ا ازه اوزره اوقه خوب آچکیر
 بن بلیزک سونلرک سن بلیزک اما سیک عتک یوقدر التعم
 اوچکیر ازیو س اکی سیر ویر اخیزیش اوشن قوچ سر شخب سیره اقز اوشتون
 و خلق حقین بر شو یلمر و التعم الحقه بکه یاز قدر یازیشیویس یالویور
 اخیزیش امله سینه خوب سچاب یضطه یوزر میز یوزر و اندیش قوش
 و التعم کبازه قریم جدر کیده رر یضطه برر طشاغم یه انکی سیکیم

Turkish		meaning	BLEICHSTEINER	today	phonologically	reading	
(bir)	۱	1	ak'a	акы	akʷə	aqı	آق
(iki)	۲	2	ʿw-ba	qba	ʿo(ə)b'a	w{ü}ba	وبا
(üç)	۳	3	h-p'a, hə-p'a	xpá	(ə)xp'a	{t}xpa ?	اخیبا
(dört)	۴	4	p's'a-ba	pyšybá	p's ʿ(ə)b'a	b{t}šba	بشبا
(beş)	۵	5	hva-ba	xybá	x ^o (ə)b'a	xuba	خوبا
(altı)	۶	6	f-ba	fbá	f(ə)b'a	f{t}ba	فا
(yedi)	۷	7	bž'-ba	bžybá	b(ə)ž' b'a	b{t}zba	بونا
(sekiz)	۸	8	ā-ba	aaá	a: b'a	'āba	عبا
(doquz)	۹	9	ž'v-ba	žəybá	ž ^o (ə)b'a	j{t}ba	ژبا
(on)	۱۰	10	žva-ba	žəəybá	ž ^o ab'a	*ju(a)ba ?	زونا

Fig. 1: Specimen from Evliya Çelebi's travel account: original manuscript and transcript (extract)¹



S. 496 ↓		Kartvelisch				Mingrelisch			
		Gueldestaadt	Klaproth	Pallas	Peacock	Gueldestaadt	Klaproth	Pallas	Peacock
Gott	ღმერთი	Gmerti		Гмерти	Ghmerti	Gorunti			Ghoronti
Herr	ბატონი	Upali, Batoni							
Ich	მე	Me		Me	Me	Ma		Ma	
Mir	მე	Me							
Mich	მე	Me							
Mein	ჩემი	Tschemi			Tchemi	Tschschimi			Tchkim
Du	შენ	Schen		Шень	Shen	Si		Си	Si
Dich	შენ	Schena							
Dir	შენ	Schena							
Dein	შენი	Scheuni			Sheni	Tskani			Tkwan
Er	ის	Is		Иса, Иги	Ts	Itina		Итина	Tishi
Wir	ჩვენ	Tschuen		Чуень	Tchven	Tschschimi		Чхини	Tkwa
Ihr	თქვენ	Tkuen		Тхуень		Tkwa		Тква	

Fig. 2: Specimen from Güldenstädt's word-lists: original manuscript and transcript (extract, edited)²

Indicativ.

Praesens.

S. 1. besazu	exzu	exsa	izsa	uzgesa	biziza	tizesa
2. besanu	exnu	ensa	iansa	ungesa	bintesa	tintesa
3. besane	exne	enesa	ianesa	unegsa	binetesa	tinetesa
Pl. 1. besajan	exjan	ejausa	iajana	ujangesa	bijantesa	tijantesa
2. besanan	exnan	enansa	ianansa	unangesa	binantesa	tinantesa
3. besaqun	exqun	equnsa	iaqunsa	uqungesa	biquntesa	tiquntesa

Imperfectum.

S. 1. besazui	exzui	ezsai	izsai	uzgesai	biztesai	tiztesai
2. besanui	exnui	ensai	iansai	ungesai	bintesai	tintesai
3. besanei	exnei	enesai	ianesai	unegesai	binetesai	tinetesai
Pl. 1. besajani	exjani	ejausai	iajansai	ujangesai	bijantesai	tijantesai
2. besanani	exnani	enansai	ianansai	unangesai	binantesai	tinantesai
3. besaquni	exquni	equnsai	iaqunsai	uqungesai	biquntesai	tiquntesai

Fig. 3: Specimen from Schiefner's Grammar of Udi¹

Неоконч. накл. лчар.
Повелит. 2-е лицо. лча.
Наст.-буд. условное лчах.
Наст.-буд. усл.-жел. лчахъра.

Наст. изъявит. лчу.
Повелит. 3-е лицо. лчул-а.
Прич. настоящее. { лчуң,
 лчуриг.
Сослагательное. . . . лчујла.
Буд. определенное. лчур.
Дьеприч. настоящ. лчуш.

Прош. недавнее лчич.
Прош. совершенное. лчир.
Дьепр. прошедшее { лччах или
 лччи.

Повѣствоват. . . лчина.
Давнопрошедшее. . . лчиніѣра.
Прич. прошедшее. { лчина,
 лчинариг.
Прил.-гл. ф. пр. вр. { лчиначуҳ,
 лчиначу,
 лчиначуриг.

Fig. 4: Specimen from Uslar's Grammar of Chechen²

¹ Transcript provided by Lana Ahlborn in the frame of the ARMAZI project (1999-2003), see <http://armazi.uni-frankfurt.de/armaziII/material/tabelle.htm>.

² Anton Schiefner, *Versuch über die Sprache der Uden*, St. Petersburg 1863 (Mémoires de l'Académie Impériale des Sciences de St.-Petersbourg, VIIe sér., t. VI/8, p. 30 (<http://books.google.com/books?id=OpIFAAAaAAJ>)).

Богъ <i>Гербетъ.</i>	Заря <i>Зиноръ.</i>
въ кн. Сван. <i>Герметъ.</i>	Утро <i>Ламъ.</i>
Небо <i>Деуъ.</i>	Вечеръ <i>Набоуъ.</i>
Ангелъ <i>Гкеръ.</i>	Сегодня <i>Лади.</i>
въ кн. Сван. <i>Англезъ.</i>	Завтра <i>Мыхаръ.</i>
Святый <i>Цхельнъ.</i>	Понедѣльникъ <i>Дештишъ.</i>
Церковь <i>Лахвамъ.</i>	Вторникъ <i>Тахашъ.</i>
въ кн. Сван. <i>Лахѣми.</i>	Среда <i>Джемашъ.</i>
Огонь <i>Лемескъ.</i>	Четвергъ <i>Цаашъ.</i>
Солнце <i>Мыжъ.</i>	Пятница <i>Вебишъ.</i>
Луна <i>Доштулъ.</i>	Суббота <i>Саптинъ.</i>
Звѣзды <i>Антхузгъръ.</i>	Воскресенье . . <i>Мыжмадехъ.</i>
въ кн. Сван. <i>Антхвасгъръ.</i>	Весна <i>Дунхѣ.</i>
Полдень <i>Исгмадехъ.</i>	Лѣто <i>Зай.</i>
Полночь <i>Исметъ.</i>	Осень <i>Мужгверъ.</i>

Fig. 5: Svan word-list from ZKOIRGO¹

арі бѣжан амѣчу бѣжан лі зход марѣ іѣштѣіаі тѣхара
лішіаіс тѣхарс тѣшхѣшіаі тѣхарс імѣ дѣхѣшіаі тѣхарѣш
дагра тѣхарс дѣшхѣшарі імѣ дѣшхѣшгѣр іѣшт дѣуаі чѣіаі
хвѣ іѣру ладііш лѣзпѣс ѣшху ладѣй урі ѣшху амѣхпѣс сгѣла
мѣхѣрнѣд ѣсбін лізі рѣдіаі міндор аѣхуед ѣѣзгі мѣліѣна
агріташ муштуаі ѣнкіѣла кѣл чѣмпіаі рѣуіш калѣаі ахсѣа
і калѣуіц тѣуім хѣхѣд тѣхмѣс лѣхт ѣѣл чѣдрінкѣа аѣхѣ сі бѣ
жан ѣзаллі марс хѣшѣуеліаі хѣа марѣ ѣхѣ сі кѣімѣаѣм хі
чіаіа хѣа ѣх маѣу хѣа ѣх дѣшкѣуід хѣла мѣ мамаѣу
ісгві лѣнхурі мѣ мѣѣх маруѣан аѣрасуан мѣж і дѣшѣуіс
хѣпѣш чіѣшѣаѣа сѣѣбін чѣхѣг ѣкѣрѣш лѣхура жіѣусгур
ѣкѣрѣш лѣсѣгѣар жіах ѣкѣрѣш кѣравіса арі ѣкѣрѣш ѣѣр
вал хѣѣб ѣѣі ѣх сі ѣр мѣмаѣкѣа мѣ чѣмѣтра марѣ нашіаіхуід
хѣаі.

Fig. 6: Svan text from SMOMPK²

¹ A. V. Bartolomej (Bartholomäi), ‘Poëzdka v Vol’nuju Svanetiju’, ZKOIRGO = Zapiski Kavkazskago Otděla Imperatorskago Russkago Geografičeskago Obščestva 3, 1855, 149–237 (here: p. 214; http://books.google.com/books?id=_UxDAAAACAAJ).

² A. N. Gren, ‘Svanetskie Teksty’, SMOMPK = Sbornik Materialov dlja Opisaniya Městnostej i Plemen Kavkaza 10/2, 1890, 76–109 (here: p. 76; <http://books.google.de/>).

Most of the printed sources indicated above have meanwhile been digitised in the course of the ‘Google Books’ project (<https://books.google.com>). However, the compilation of linguistic materials of the Caucasian languages from Google’s digitised files is anything but straightforward. First of all, Google does not provide access to all the digital books it hosts because it tries to respect the copyright regulations that are valid for the country of a given user. This means, e.g., that for users from the USA, all books that were printed before 1923 are accessible while for a German user, access presupposes that the author of a book died at least 70 years ago. By consequence, American users will have access to UsLAR’s grammars published in the 1880ies but not to the 6th volume (on Tabasaran) which appeared, as an *opus postumum*, as late as 1979.¹ For German users, UsLAR’s 19th century grammars should also be accessible as the author died in 1875; however, Google still refuses to give access to them, claiming that they cannot keep track of all authors’ lives whose works they store. As a matter of fact, full access to books digitised by Google is restricted to pre-1871 books for the time being for German users – unless they have access to an American IP-address.²

Second, the digitization provided by Google (via OCR) is anything but reliable so that the linguistic materials under concern cannot be used off-hand for compiling a corpus. This can be illustrated using the example

[books?id=zNs6AQAAIAAJ](https://books.google.de/books?id=zNs6AQAAIAAJ)).

¹ Petr Karlovič UsLAR, *Ėtnografija Kavkaza: Jazykoznanie*. T. 7: *Tabasarsanskij jazyk* (ed. Aleksandr Magometov). Tbilisi 1979 (see http://books.google.de/books?id=SS_QQ-gAACAAJ and [?id=qsgRAQAAMAAJ](http://books.google.de/books?id=qsgRAQAAMAAJ)).

² I quote from an e-mail by Mr Jon Orwant (Google Inc.) of May 2, 2011: ‘Each nation has their own copyright rules. Within the US, we are usually able to use 1923 as a cutoff date for determining whether books are in copyright, and so some of the seven books you identified are fully readable and downloadable inside the US. Outside the US we have to use the rules of the appropriate country. Books from former Soviet republics and from the pre-Soviet era have an unclear copyright status.

We are able to make the PDFs available within a country when the books are out of copyright in that country, and when we do they’re available directly from Google Book Search; if you had a US IP address you could just visit Google Book Search and download the PDF with no involvement from us.

But because copyright status is often hard to determine, we have settled on the following rule for countries that don’t have a cutoff like the US: either the book must have been published before 1871, or there must be clear and convincing evidence that all authors of the book died more than 70 years ago. Of the seven books you identified on April 19, I don’t have evidence that either condition applies to any of them. (While Peter von UsLAR does appear to have died long ago, he does not appear to be the sole author of any of the seven books).

of the first word-list of Udi provided by Klaproth in his *Beschreibung der Russischen Provinzen zwischen dem Kaspischen und Schwarzen Meere* of 1814 (cf. Fig. 7). Here, the word for ‘God’, *Bacha*,¹ is compared to Avar *Betschaß*.² Undertaking a search for the latter word in Google yields four hits from Klaproth’s works, three of them from another book of his (*Reise in den Kaukasus und nach Georgien*, 1814), where it appears in a comparative Avar-Andi word-list (cf. Fig. 8–9).³ As a matter of fact, the three contexts are just the same – Google simply digitised three copies of the book.⁴ However, the digital texts are anything but identical, and it is by mere accident that the word searched for (*Betschaß*) is rendered correctly throughout in all three cases, if we compare the rendering of other words in the context (e.g., the word for ‘not’, *hetscheu*, appearing as *hetscheu*, *Helschen*, and *het scheu*, resp.). Klaproth, now, had printed the same list in a different form in yet another book of his a few years before (*Archiv für Asiatische Litteratur, Geschichte und Sprachenkunde*, of 1810; cf. Fig. 10). Here, the Avar word for ‘God’ is written *Betschass*, with double *s*. Searching for this word form yields two instances of the given work by Klaproth, one from Google itself⁵ and one from *archive.org* (cf. Fig. 11). The latter is not 100% identical, however, although it depends on the same Google digitization⁶ – probably it was stored on *archive.org* when the OCR of the scan (of 2007) was still being corrected. A third hit is found, astonishingly enough, in an ‘Instruction for the fabrication of cigars’ by one Heinrich Schütte; inspecting the context in question (cf. Fig. 12),⁷

¹ The actual form of the Udi word is *bixažug*, most probably a compound meaning the ‘Creating Lord’; see *The Caucasian Albanian Palimpsests of Mount Sinai*, vol. 1, p. IV-10 (proposal by W. Schulze).

² The given form is the ergative case form, the (absolutive) lexicon form being *bečed* (cf. Klaproth’s „Khundzakh“ form *Bedshet*).

³ <http://www.google.de/search?q=%22Betscha%C3%9F%22>; all searches quoted here were undertaken on May 21, 2011, and a second time on June 30, 2012, with similar results.

⁴ Julius von Klaproth, *Kaukasische Sprachen. Anhang zur Reise in den Kaukasus und nach Georgien*, Halle and Berlin 1814 (<http://books.google.de/books?id=cpkbTqZhJnKC>; ?id=Obk4AQAIAAJ; ?id=EU8PAAAAQAAJ).

⁵ https://www.google.de/search?q=%22Betschass%22&gws_rd=ssl. Meanwhile (May 29, 2015), there are two more hits from Klaproth’s *Archiv*, which has obviously been digitised several times.

⁶ See <http://archive.org/details/archivfrasiatis00klapgoog>, linked to <http://books.google.de/books?id=8QMJAAAAQAAJ> just like the Google-internal link.

⁷ Heinrich Schütte, *Anleitung zur Fabrikation von Cigarren*, Bremen 2010 (Reprint of the edition Quedlinburg and Leipzig 1846), p. 34 (<http://books.google.de/books?id=o2kQjP08pJwC>). In the scan of the original print of 1846 (published under the title *Die Cigarrenfabrikation*;

we see that this is a mere OCR error, German *verschaffen* in Gothic print having been misread as *betschass*. Leaving this disturbing effect aside, we may summarise that a) the fact that Google digitises several copies of one and the same book does not necessarily improve the reliability of the data the company makes available, and b) Google provides no means to distinguish languages in a given search: Avar *Betschaf* is not at all determinable as being Avar – which would be important indeed in the given context as it is mentioned in an Udi word-list, in its turn embedded in a German context –, and is even mixed up with a quasi-German *betschass* in the results. This means that an automatical compilation via harvesting of the relevant data from Google Books is not yet possible.

<p>Im Gebiete von Schechl liegt auch noch das Dorf Waratschin, dessen Einwohner Christen sind, die aus Georgien herzukommen vorgeben, und von denen die eine Hälfte sich zur Georgischen, die andere zur Armenischen Sekte hält. An ihren Kirchen sollen viele Inschriften in Stein gehauen seyn, die sie aber selbst nicht lesen können. Sie bebielen sich Georgischer und Armenischer Väter, aus welchen sie die Jugend unterrichten. Dessen ungeachtet sprechen sie einen Lesghischen Dialect, den sie für ihre Muttersprache ausgehen, und aus dem ich hier einige Worte, mit andern Lesghischen verglichen, folgen lasse.</p>	<p style="text-align: center;">— 178 —</p> <p>Tochter — Chinat. Bruder — Witschi — Awarisch: <i>uaz, wam, waf.</i> Mädchen — Esengl. Knabe — Galt — Tatarisch: <i>ogit, ul, bala.</i> Wasser — Chee — Mizdshagisch und Turkschisch: <i>Ehl.</i> Wein — Gieh. Brod — Schum — Awarisch: <i>Tsched.</i> Apfel — Desch. Mein Bruder — Dis witschi. Ich Brod mit uns, mein Bruder — <i>Miecke arza schum uka bis witschi.</i></p>
<p>Gott — Dacha — Awarisch: <i>Betschaf.</i> In Kundsach: <i>Wedsher.</i> Weib — Schumuf — Awarisch: <i>Tschuschu.</i> Sohn — Gari</p>	

Fig. 7: Udi word-list in Klaproth 1814a¹

<http://books.google.de/books?id=v8g-AAAACAAJ>), *verschaffen* was OCRed correctly (see <http://books.google.de/books?id=v8g-AAAACAAJ&q=verschaffen>). – Meanwhile (May 29, 2015), a similar effect can be seen when searching for the spelling *betschaf*, for which there is one more hit now, from Diedrich Westermann's *Wörterbuch der Ewe-Sprache* (vol. I, Berlin 1905, p. 313); here, the original print has the German word *botschaft* (under the lemma *künyà* 'Todesnachricht').

¹ Julius von Klaproth, *Beschreibung der Russischen Provinzen zwischen dem Kaspischen und schwarzen Meere*, Berlin 1814, pp. 177–8 (<http://books.google.de/books?id=Hq1KAAAACAAJ>).



Fig. 8: Google search for 'Betschaß'

A u á r i s c h.	A n d i s c h.
Gott, Betschaß	Gott jowob
Tod, adshal	Tod adshal
nicht; hetscheu;	nicht suw
Mensch adam	Mensch adam
Leben jômryr	Leben jomuru
viel jemere	kurz. botscho.

Fig. 9: Avar *Betschaß* in Klaproth 1814b¹

¹ Klaproth, *Kaukasische Sprachen*, p. 35.

Sprachproben in Auárischer und Andischer Sprache.

I. Gott stirbt nicht ; der Mensch lebt nicht lange.

Auárisch.

Gott *Betschass*
Tod *adshal*
nicht *hetscheu* ;
Mensch *adam*
Leben *jómyr*
viel *jemere*
nicht dauernd. *kwanaljari.*

Andisch.

Gott *zowod*
Tod *adshal*
nicht *ssuw*
Mensch *adam*
Leben *jomuru*
kurz. *botscho.*

Fig. 10: Avar *Betschass* in Klaproth 1810

"Betschass"

Search

About 4 results (0.08 seconds)

Advanced search

[Archiv für asiatische Litteratur, Geschichte und Sprachkunde - Google Books Result](#)
Julius von Klaproth, *Akademiya nauk* - 1810 - History - 224 pages
Gott **Betschass**- \- Gott' zowod Tod adshal Tod adshal nicht hetscheu ; nicht ssuw Mensch adam . Mensch adam Leben jomyr Leben jomuru viel jemere kurz, ...
books.google.com/books?id=imEoAAAAYAAJ...

[Anleitung zur Fabrikation von Cigarren - Google Books Result](#)
Heinrich Schlütte - 2010 - History - 42 pages
... nicht geben können, die präk tische DeMönstraM, «inen ausreichenden Ersah zu **betschass**/n, um d'adtirch llntr« Anweisuiig noch praktischer zu machen. ...
books.google.com/books?isbn=3867414823...

[Full text of "Archiv für asiatische Litteratur, Geschichte und ... \[Translate this page \]](#)
Gott **Betschass** Gott zowod Tod adslial Tod adshal nicht hetscheu; nicht ssuvü Mensch adam Mensch adam Leben jómyr Leben jomuru viel jemere kurz« **botscha**. ...
www.archive.org/stream/.../archivfrasiatis00klapgoog_djvu.txt - Cached - Similar

In order to show you the most relevant results, we have omitted some entries very similar to the 3 already displayed.

If you like, you can [repeat the search with the omitted results included](#).

Fig. 11: Google search for 'Betschass'

then, daß sich solche weniger durch schriftliche Beschreibung, leicht aber durch eigene Anschauung erlernen läßt. Da die letztere manchem unserer Leser unzugänglich sein möchte, und gerade diese zuerst nach unserer Anweisung greifen dürften, so wollen wir versuchen, ihnen für das, was wir hier nicht geben können, die praktische Demonstration, einen ausreichenden Ersatz zu **verschaffen**, um dadurch unsere Anweisung noch praktischer zu machen. Wir empfehlen Ihnen demzufolge, sich in eine gut renommierte Tabakhandlung zu begeben, sich mehrere Sorten zeigen zu lassen und von allen verschiednen gestalteten Cigaretten etwa ein halbes Duzend

Fig. 12: German *verschaffen* misread as *betschass* in H. Schütte's Anleitung

2. Scripts and character encoding

We have seen in the previous example that the distinction of printing types may be connected with orthographical variation, which may be crucial for the retrieval. This is all the more true in the given context as the language material under concern was noted in different scripts in the printed sources (Arabic, Latin, Cyrillic, later also Georgian etc.). Digitizing these sources for a world-wide retrieval presupposes the existence of an encoding standard that covers the phonetic comprehensiveness of Caucasian languages throughout these scripts. As a matter of fact, Unicode, the encoding standard prevalent in the World-Wide Web today,¹ is not yet sufficiently elaborated to fulfil this task.

One problem that is obvious here is the usage of accents and other diacritical marks in the rendering of Caucasian languages. Since its invention in the early 1990ies, Unicode has provided code points for many combinations of base characters with diacritics such as, e.g., *ä*, *á*, *ğ*, or *ü*, mostly in connection with existing national orthographies. In addition to this, it contains a large set of diacritics for free combination with base characters, a sequence of *a* + acute accent (´) being equivalent to ‘precomposed’ *á*. In recent years, the Unicode Consortium has been reluctant to add ‘precomposed’ charac-

¹ See the official website of the Unicode consortium, <http://www.unicode.org>.

ters to the standard, users being forced to use free combinations instead;¹ and the software industry is still trying to cope with the correct rendering of such combinations on the screen and in printed form.

Be that as it may, neither the inventory of diacritics nor that of base characters in Unicode matches all the requirements of encoding older specimens of the Caucasian languages digitally. One example may suffice again to show the effects of this. In the 10th volume of the *Sbornik Materialov dlja Opisani-ja Městnostej i Plemen Kavkaza* (SMOMPK), a journal devoted to the exploration of the Caucasian ‘Tribes’ and languages, several authors provided specimens of the Svan language, a cognate of Georgian. Fig. 6 above shows a specimen of A.N. Gren’s collection of Svan texts, written in Cyrillic with quite a lot of diacritics and a few extra characters. In Fig. 13, the different degrees of encodability of these elements are indicated by colours. Black indicates (Cyrillic!) characters that are encodable as such, being part of the Unicode ‘Cyrillic’ block because they are also used in (modern or older) Russian, Serbian, Abkhaz, or other languages using the Cyrillic script (e.g., *a, ô, c, m, μ, v, ž*). This also comprises some precomposed characters such as *ä*, which in a Cyrillic context would have the code point U+04D3, not U+00E4 as in a Latin context.² Green indicates combinations of base characters with diacritics that are not encodable as precomposed characters but can be encoded as free combinations (e.g., *q, ñ, ð, ŷ, ǰ, ǵ*). As there is no extra-set of diacritics to be used in combination with Cyrillic characters, the rendering of such encodings depends on the facilities of the operating system and / or software used, and the results may still look odd (as in the example given here).³ The case is even worse with elements marked with blue colour as these imply a special treatment in the combination of a base character with a diacritic as in the case of *ĩ* where the dot should be suppressed when the breve is added. Using the Turkish dotless *ı* as the base character instead is no recommendable solution here as this is not part of the Cyrillic block of Unicode.⁴ In cases of items rendered

¹ As to ‘precomposed characters’ see the FAQ page, http://unicode.org/faq/char_comb-mark.html, and the mail exchange documented on <http://unicode.org/mail-arch/unicode-ml/Archives-Old/UML003/0481.html>.

² See <http://www.unicode.org/charts/PDF/U0400.pdf> as to the Cyrillic block.

³ It is true that Microsoft provides a special rendering engine for the positioning of diacritics in its Office suite for Windows since 2007; this, however, covers only Latin base characters, not Cyrillic ones.

⁴ Note that a Latin *i*-breve combination is encodable as a precomposed character (*ï*, U+012D); if encoded as a sequence of *i* + (diacritic) breve (U+0049 + U+0306), the result is as well rendered correctly (as *ĩ*) using Microsoft Office 2007.

in violet colour, it is unclear whether we have to deal with a diacritic at all or a mere flyspeck, as in the case of *ō* in *ახცხობა* or *ŕ* in *დაშხადჳარი*; to decide this, a thorough investigation of the individual attestations is needed. Finally, items printed in red colour are characters that cannot be encoded as such in any way, as in the case of the fourth and sixth character in the sequence *ჳუხორ თჳვრამ*; replacing this character by *o* (U+0277), an obsolete character denoting a semi-high back rounded vowel in the International Phonetic Alphabet, cannot be anything but a temporary makeshift. The compilation of necessary extensions of the Unicode standard thus remains an indispensable prerequisite before the data can be systematically integrated into a corpus.¹

არი ბაჟან ამაჩუ ბაჟან ლი ჳოდ მარა იბჳთუილ ოჳახარა
 ლიშიაღს ოჳახარს ოჳახაშიაღს ოჳახარს იმგა დოჳაშიაღს ოჳახარაშ
 დაგრა ოჳახარს დაშხადჳარი იმგა დოშოდოჳთგარი ნოშთ ლუზაი ყჩქინთ
 ხვი იჳრუ ლადგჳიშ ლაჳნაჳს აშხუ ლადაჳთ ური აშხუ ამჳჩნიკ სგაღა-
 მოჳარნად თსბინ ლიზი უდიღაჳთ მინდორ აჳხუედ კოჳტგი მილიონკა
 აგრითჳხ მუშტუაჳთ ანკიოლა კილ ჩამცილ დაჳყუიშ კალტაი ახცხობა

არი ბაჟან ამაჩუ ბაჟან ლი ჳოდ მარა იბჳთუილ ოჳახარა
 ლიშიაღს ოჳახარს ოჳახაშიაღს ოჳახარს იმგა დოჳაშიაღს ოჳახარაშ
 დაგრა ოჳახარს დაშხადჳარი იმგა. დოშოდოჳთგარი ნოშთ ლუზაი ყჩქინთ
 ხვი იჳრუ ლადგჳიშ ლაჳნაჳს აშხუ ლადაჳთ ური აშხუ ამჳჩნიკ სგაღა-
 მოჳარნად თსბინ ლიზი უდიღაჳთ მინდორ აჳხუედ კოჳტგი მილიონკა
 აგრითჳხ მუშტუაჳთ ანკიოლა კილ ჩამცილ დაჳყუიშ კალტაი ახცხობა

არი ბეჟან ამეჳუ ბეჟან ლი მხოღ მარე ებშთუიელ თახარა
 ლიშიაღს თახარს თეშეშეიაღ თახარს იმღა დოხეშეიაღ თახარეშ
 დაგრა თახარს დეშნადგარი იმღა დოშოთოთდაგარი ვოშთ ლუზაი ჳყინტ
 ხვი აერუ ლადღიშ ლეზნას ეშხუ ლადედ ური ეშხუ ამფხნიკ
 სგალამოხერნედ ოსბინ ლიზი ჳოდიათ მინდორ ათხუედ ქოთმგი მილიონკა
 ანგრითახ მუშტუათე ენკიოლე კალ ჩემცილ ჳაჳჳიშ კალთაი ახსგბა

Fig. 13: Gren's Svan text⁹ with a Latin transcript, with degrees of encodability indicated by colours, and a Georgian transcript

¹ The preparation of a Unicode extension proposal for the rendering of Caucasian languages in older prints was the actual task of the project 'Towards a Corpus Caucasicum', which was financed by Google Inc. in 2011. In the one-year runtime of the project, the requirements for encoding the South Caucasian (Kartvelian) languages as printed in Tzarist publications were established; further continuation of the project is pending.

3. Unification of data

As we have seen above, the early specimens of Caucasian languages to be compiled into a Corpus Caucasicum were printed in several different scripts, among them Arabic, Latin, and Cyrillic. In addition to this, further scripts such as Georgian Mkhedruli have been adapted to write Caucasian languages in more recent times; cf., e.g., the Georgian rendering of the Svan sample text in Fig. 13. This now implies that the different sources must be encoded differently if the digital representation is meant to give a true picture of the original outline of the specimens. It further implies that special modes (e.g., a ‘Latin’ mode, an ‘Arabic’ mode etc.) depending on the original rendering must be designed for all kinds of queries concerning these materials. This, however, is in no way satisfactory for a ‘Corpus’ covering either one or several of the languages; instead, users should be provided with facilities that bridge between the different rendering systems. This presupposes, as we have seen, the compilation of inventories of encodable, composable and unencodable characters used in the printed materials, with a view to encode them ‘uniquely’ on the basis of an extension of the Unicode standard. Second, it presupposes the setting up of relationships between the different rendering systems used (Cyrillic, Latin, Georgian, Arabic etc.) as well as variants of these systems. This can, for the time being, best be achieved by adapting a multilevel annotation of the text specimens where all elements are tagged, in addition to their rendering in the original ‘spelling’, with a common, standardised representation accompanying it. This approach has been successfully tested with older specimens of Udi (cf. Fig. 14)¹ which can now be searched not only via the original script but also via a unified Latin transcription.² This,

¹ See <http://titus.uni-frankfurt.de/texte/etcs/cauc/udi/cput/cput.htm> (‘Corpus of published Udi texts’, entered and slightly corrected by Wolfgang Schulze, Gräfensteinberg, 2003-2005).

² Curiously enough, a Google search for the word *Русьтамаха́л* (dat.sg. of the proper name *Rusʹtam*, with focus particle) yields only the TITUS edition of the text passage but not the

however, has to be developed manually for each ‘orthographic’ representation, taking into account the actual phonological systems of the languages. The ‘automatic’ compilation of a reliable and usable Corpus Caucasicum from the materials provided by Google Books remains illusionary indeed until the problems indicated above have been solved.

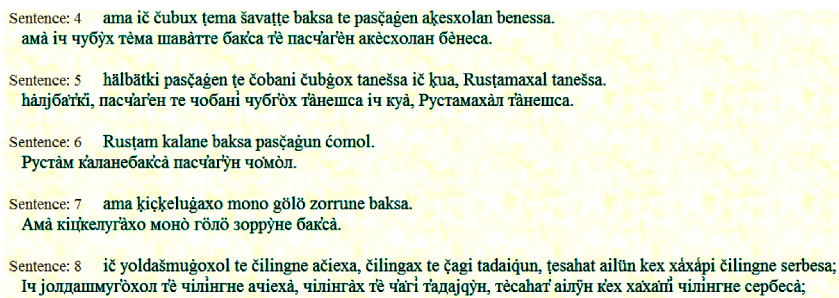


Fig. 14: Twofold rendering of Udi text specimens in the TITUS project

original print in SMOMPK 6/Pril., 1888, p. 7 (see <http://www.google.de/search?q=%D0%A0%D1%83%D1%81%D1%82%D0%B0%D0%BC%D0%B0%D1%85%D0%B0%CC%80%D0%BB>); it seems that the volume in question has not yet been OCRed (but digitised, see <http://books.google.de/books?id=OR7VAAAAMAAJ>). – It is also curious that the whole series of SMOMPK is only found in Google Books under a Russian transliteration of the French version of its title, *Recueil de matériaux pour la description des contrées et tribus du Caucase* (see <http://books.google.de/books?q=editions:LCCN55045301>); in 2012, this was still rendered in Cyrillic characters (*Ресуеил де матэриаукс пур ла дескрипцион дес контрées эт трибус ду Саукасе*).

Corpus Caucasianum-ის შესახებ
კორპუსის აგება არასტრუქტურული მასალიდან
რეზიუმე

სტატიაში განხილულია კავკასიური ენების კორპუსის კომპილაციის პერსპექტივები, რაც გააერთიანებს ამ ენებში დადასტურებულ მოუწესრიგებელ ლინგვისტურ მასალას მე-20 საუკუნემდე გამოქვეყნებულ პუბლიკაციებში, რომლებიც დიგიტალიზირებულ იქნა Google Books პროექტის ფარგლებში. წარმოვადგენთ სამ ძირითად პრობლემას, რომელიც უკავშირდება აღნიშნულ საკითხს, კერძოდ, მასალის მოგროვება და კომპილაცია, ნიშნების კოდირება, და მასალის უნიფიცირება. უმთავრესად ვეხებით: ა) მასალის ხელმისაწვდომობას საავტორო უფლებების ფონზე, ბ) OCR-ის მიერ მომზადებული მასალის არასანდოობას, გ) ენის დადგენის ნაკლებობას შერეულ მასალაში, დ) ხარვეზებს Unicode-ით კოდირებულ ერთეულებში, და ე) მასალის ერთიანი დამწერლობისგან-დამოუკიდებელი წარმოდგენის საჭიროებას. განსახილველი მასალა ილუსტრირებულია პუბლიკაციებიდან (XVII საუკუნიდან მოყოლებული XIX საუკუნის ბოლომდე) აღებული მაგალითების მიხედვით.