Attention!
This is a special internet edition of the article
“Towards a Corpus Caucasian. Building a corpus from unstructured data”
It should not be quoted as such. For quotations, please refer to the
original edition in
Jikia, Marika (ed.), *Tipologiuri dziebani / Typological Investigations* VII,
Tbilisi 2015, 149-165.

Alle Rechte vorbehalten / All rights reserved:
Jost Gippert, Frankfurt 2015
Towards a Corpus Caucasian
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-20\textsuperscript{th} 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 17\textsuperscript{th} to the end of the 19\textsuperscript{th} 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.\textsuperscript{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,\textsuperscript{2} while a few others were adapted

\textsuperscript{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.

\textsuperscript{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 Gülデンstädt 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).

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 /?openpath=MPI663243%23.
II: 258b, 11

Fig. 1: Specimen from Evliya Çelebi’s travel account: original manuscript and transcript (extract)³

---

151
<table>
<thead>
<tr>
<th>S. 496</th>
<th>Kartvelisch</th>
<th>Mingreliisch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Güldenstaedt</td>
<td>Klaproth</td>
</tr>
<tr>
<td>Gott</td>
<td>ḡ'w'j'w'j'w'</td>
<td>Gmerti</td>
</tr>
<tr>
<td>Herr</td>
<td>ʰkʰʰʰʰʰʰʰʰ</td>
<td>Upali, Batoni</td>
</tr>
<tr>
<td>Ich</td>
<td>ʰjʰj</td>
<td>Me</td>
</tr>
<tr>
<td>Mir</td>
<td>ʰjʰj</td>
<td>Me</td>
</tr>
<tr>
<td>Mich</td>
<td>ʰjʰj</td>
<td>Me</td>
</tr>
<tr>
<td>Mein</td>
<td>ḡ'j'j'j'j</td>
<td>Tičemi</td>
</tr>
<tr>
<td>Du</td>
<td>ḡ'j'j</td>
<td>Schen</td>
</tr>
<tr>
<td>Dich</td>
<td>ḡ'j'j</td>
<td>Schena</td>
</tr>
<tr>
<td>Dir</td>
<td>ḡ'j'j</td>
<td>Schena</td>
</tr>
<tr>
<td>Dein</td>
<td>ḡ'j'j</td>
<td>Schemi</td>
</tr>
<tr>
<td>Er</td>
<td>ʰrʰ</td>
<td>Is</td>
</tr>
<tr>
<td>Wir</td>
<td>ḡ'g'g'</td>
<td>Tschvani</td>
</tr>
<tr>
<td>Ihr</td>
<td>ḡ'g'g'</td>
<td>Tkhvani</td>
</tr>
</tbody>
</table>

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

**Præsens.**

<table>
<thead>
<tr>
<th>S. 1.</th>
<th>besazu</th>
<th>exzu</th>
<th>essa</th>
<th>tæza</th>
<th>uzgesa</th>
<th>bizteza</th>
<th>tiztesa</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>besanu</td>
<td>exnu</td>
<td>essa</td>
<td>tæna</td>
<td>ungesa</td>
<td>binesa</td>
<td>tienesa</td>
</tr>
<tr>
<td>3.</td>
<td>besane</td>
<td>exne</td>
<td>enesa</td>
<td>tænesa</td>
<td>unegsa</td>
<td>binetesa</td>
<td>tinetesa</td>
</tr>
</tbody>
</table>

**Imperfectum.**

<table>
<thead>
<tr>
<th>S. 1.</th>
<th>besazui</th>
<th>exzui</th>
<th>essai</th>
<th>tæzai</th>
<th>uzgesai</th>
<th>biztesai</th>
<th>tiztesai</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>besanui</td>
<td>exnui</td>
<td>essai</td>
<td>tænai</td>
<td>ungesai</td>
<td>binetsai</td>
<td>tinetesai</td>
</tr>
<tr>
<td>3.</td>
<td>besanei</td>
<td>exnei</td>
<td>enesai</td>
<td>tænesai</td>
<td>unegesai</td>
<td>binetessai</td>
<td>tinetessai</td>
</tr>
</tbody>
</table>

---

**Fig. 3:** Specimen from Schiefner’s Grammar of Udi

---

**Fig. 4:** Specimen from Uslar’s Grammar of Chechen

---

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](http://armazi.uni-frankfurt.de/armaziII/material/tabelle.htm).

арі  bgcolor = "lightblue";\text{\textit{Svan word-list from ZKOIRGO}}\textsuperscript{1}.

\begin{tabular}{|l|}
\hline
арі  bgcolor = "lightblue";\text{\textit{Svan text from SMOMPK}}\textsuperscript{2}.

\hline
\end{tabular}

\begin{itemize}
\item A. N. Gren, ‘\textit{Svanetskie Teksty’}, SMOMPK = Sbornik Materialov dlja Opisanija 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 acces-
sible 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).

¹ Petr Karlovič Uslar, Ėtnografija Kavkaza: Jazykoznanie. T. 7: Tabasaranskij jazyk (ed.
Aleksandr Magometov). Tbilisi 1979 (see http://books.google.de/books?id=SS_QQ-
gAACAAJ and ?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 down-
load the PDF with no involvement from us.

But because copyright status is often hard to determine, we have settled on the follow-
ing 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 Sprachkunde*, 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 itself\(^5\) and one from *archive.org* (cf. Fig. 11). The latter is not 100% identical, however, although it depends on the same Google digitization\(^6\) – 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 *bixaǯuğ*, 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.


\(^5\) 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=8QMJAAAQAAJ 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=02kQjP08pJwC). 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 Betschaß 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.

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 *betschaß*, 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 *kúnyà* ‘Todesnachricht’).

Did you mean: “Betscha”

Reise in den Kaukasus und nach Georgien unternommen in den Jahren ... - Google Books Result
Julius von Klaproth - 1814 - Foreign Language Study
Gott, Betschaß. Gott zowod Tod, adshal Tod adshal nicht; hetscheu, nicht fuw > Mensch
e dam’ > Mensch e dam Leben jomyr Leben jomuru viel jemere turj ...
books.google.com/books?id=D8MAGAAAMAAJ...

Beschreibung der Russischen Provinzen zwischen dem Kaspischen und ... - Google Books Result
Julius von Klaproth - 1814 - History - 263 pages...
... Lesghischen von gleichen, folgen lasse. Gott — Bach — Avarisch: Betschaß. Z
Kchundsach: Bedesh. Weib — Schuwit — Avarisch: Tschushu. Sohn — Ban. ...
books.google.com/books?id=Ew0FAAAAYAAJ...

Reise in den Kaukasus und nach Georgien unternommen in den Jahren ... - Google Books Result
Julius von Klaproth - 1814
Gott, Betschaß. Gott zowod’ Tod, adshal Tod adshal nicht. Helschen; nicht fuw Mensch
adsM Mensch abam Leben jomyr’ ’ > Leben jomuru viel jemere turj ...
books.google.com/books?id=cpxB1gZlygWc...

Reise in den Kaukasus und nach Georgien ... 1807 und 1808. [With ... - Google Books Result
Henrich Julius von Klaproth - 1814
Gott, Betschaß. Gott zowod Tod, adshal Tod adshal nicht; hetscheu; nicht fuw Mensch ab
Mensch abam Leben jomyr Leben jomuru viel jemere kurz. botsche ...
books.google.com/books?id=EUBPAAAIQAQJ...

Fig. 8: Google search for ‘Betschaß’

Fig. 9: Avar Betschaß in Klaproth 1814b

---

1 Klaproth, Kaukasische Sprachen, p. 35.
Sprachproben in Auarischer und Andischer Sprache.

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

Auarisch. Andisch.

Gott Betschass Gott zowod
Tod adshal Tod adshal
nicht hetscheu; nicht ssuw
Mensch adam Mensch, adam
Leben jomyr Leben jomuru
viel jemere kurz. botscho.
nicht dauernd. kwañaljari.

Fig. 10: Avar Betschass in Klaproth 1810

Fig. 11: Google search for ‘Betschass’
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., \(ä, á, ģ,\) or \(ů\), 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; 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 Opisanija 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., а, б, с, т, џ, ћ). This also comprises some precomposed characters such as а, which in a Cyrillic context would have the code point U+04D3, not U+00E4 as in a Latin context. Green indicates combinations of base characters with diacritics that are not encodable as precomposed characters but can be encoded as free combinations (e.g., ā, ĕ, ř, ѣ, є, џ). 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). 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 ĭ where the dot should be suppressed when the breve is added. Using the Turkish dotless ĭ as the base character instead is no recommendable solution here as this is not part of the Cyrillic block of Unicode. 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 (ı, U+012D); if encoded as a sequence of i + (diacritic) breve (U+0049 + U+0306), the result is as well rendered correctly (as ı) using Microsoft Office 2007.

161
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 ņ in აწკჰდა or ᴢ in ძახადჰაჸ; 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 ŭххъ ჯვრაჰშ; replacing this character by ø (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 undispensable prerequisite before the data can be systematically integrated into a corpus.

---

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 Mkheidruli 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,

---

\(^2\) Curiously enough, a Google search for the word Рустамаха̀л (dat.sg. of the proper name Rusťam, 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 (Ресуэил де матэриаукс пур ла дескриптион дес контрéeс эт трибус ду Саукасе).
Corpus Caucasianum-ის შესახებ

კორპუსის დახმარებით ისტორიულმა მემორალმა

კორპუსი საქართველოს ექსპორტის კომისიის ბრძანების შესამართლის ბრძანების შესაძლო საქმიანობის თანამშრომლობის განვითარების გზებზე 20-იან საუკუნიდან დასავლეთიდან ჩრდილოეთიდან გვეწყვეტილობენ, რომელთა მიერ თანამშრომლობის გზებზე იქნება Google Books პროექტის ფართო გამოყენება. ბრძანების სახელი მოსახილველი პროექტის სახელი, რომელთა წარწერის შემთხვევაში სახელი, გერმანულ-გერმანულ გამარჯვების და გერმანულ-ყინულური, ნინგრძობის გამარჯვების, და სახელის შემართლება. სამყაროს გამოყენება გამოიწვის: ა) სახელის სტატუს-სლოგანის სახელმწიფო ზოგადობის ფართო, ბ) OCR-ის პერფორმანსს მონაცემთა ფონის დამუშავება, გ) სახელის დამუშავების სტეფანო გამარჯვების, დ) სახელმძღვანელო Unicode-ის ვესტ გამარჯვების, და ე) სახელის სტატუს-სატრალი-ინფორმაციური დამუშავების, ხაზის დამუშავების სატრალი-ინფორმაციური (XVII საუკუნეთან მიმდევრობის XIX საუკუნეთა მიმდევრობის) დამუშავების ხაზის დამუშავები პატივშენები. 165