Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache

Band 5

Jost Gippert / Ralf Gehrke (eds.)

Historical Corpora

Challenges and Perspectives

narr VERLAG

Jost Gippert / Ralf Gehrke (eds.)

Historical Corpora

Challenges and Perspectives

narr
VERLAG

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Internet: www.narr.de
E-Mail: info@narr.de

Redaktion: Melanie Steinle, Mannheim
Layout: Andy Scholz, Essen (www.andyscholz.com)
Printed in Germany

ISSN 2191-9577
ISBN 978-3-8233-6922-6

Contents

JOST GIPPERT / MANANA TANDASHVILI

Structuring a diachronic corpus

The Georgian National Corpus project

Abstract

The paper deals with the structural premises of diachronic corpora that are meant to represent specimens of a given language throughout its historical stages and to provide a diachronic cross-century retrieval. On the basis of the Georgian National Corpus project, it discusses ways to cope with variation caused by the use of different scripts and by language change, as well as requirements of annotating the different layers (chronological, dialectal, sociolectal, genre-based, etc.) the text materials pertain to, including a critique of the concepts of the ISO 639-6 standard.

1. Introduction

Corpora that are designed to embrace a given language throughout its historical stages and to provide diachronic access to its features present special challenges as to their structuring. Among these challenges, we may mention the problem of linguistic variation with all its facettes, including phonetic change and its (ortho-)graphical representation, morphological, syntactical, and semantic change, but also the necessity of balancing between well and less well attested text genres. Until the present day, only a few projects have successfully attempted to establish corpora that cover a time-span of more than a few centuries. In the present paper, we discuss some of the peculiar requirements of a large-scale diachronic corpus on the basis of the Georgian National Corpus project,[1] which has to cope with most of the problems addressed above. After outlining the project and its background, the paper focusses first on the problem of the various scripts used for Georgian throughout its history and their handling, and second, on the question of how to annotate the linguistic varieties to be subsumed in the corpus with a view to differentiated retrieval.

[1] The initial spark of the project was the foundation of a coordination council in Tbilisi on July 19, 2011 (see http://geocorpus.blogspot.de/p/blog-page_21.html – all URLs quoted here were last checked on January 28, 2015). The project started, with kind support by the Volkswagen Foundation, in autumn, 2012.

2. The Georgian National Corpus project and its background

The plan to establish a Georgian National Corpus (hereafter: GNC) that covers the complete time range from the earliest attestations of written Georgian in the 5th century C.E. up to the present day has evolved from several corpus building initiatives that have been realized since the late 1980s, mostly in joint endeavours of German and Georgian partners. This is true, first of all, for the text database of the TITUS and ARMAZI projects in Frankfurt,[2] which covers nearly all published text materials from the periods of Old and Middle Georgian (roughly 5th-13th and 13th-18th cc.) as well as a minor collection of Modern Georgian texts (19th c.; mostly grammatical treatises and poetic works). These materials (ca. 6 Mio. tokens), most of which were electronically prepared since 1987 via OCR, with manual correction and formatting, have been thoroughly preindexed and are searchable via a word-form based retrieval system, which reflects the chronological order of the attestations in its output. For the time being, a lemmatization function has not yet been implemented; however, the retrieval engine provides a lexicon-based word analysis for nominal forms (cf. Figure 1).[3]

The second main pillar of the GNC is the GEKKO corpus run by Paul Meurer in Bergen / Norway,[4] which has been compiled, mostly via data harvesting, from free online resources in Georgian, among them many newspapers and journals, but also literary texts (both autochthonous and translated) as well as pages from several official and semi-offical websites in Georgia. The corpus thus established comprises ca. 125 Mio. tokens; about one fifth of it (20 Mio. tokens) has already been equipped with a full morphological annotation and a lemmatization function which includes the extremely complex verbal system of the Georgian language (cf. the sample output in Figure 2).[5]

A third pillar of the GNC is the extensive corpus of dialectal varieties of spoken modern Georgian ('Georgian Dialect Corpus', GDC) compiled under the direction of Marina Beridze at the Arnold Chikobava Institute of Linguistics

---

[2] Cf. http://titus.uni-frankfurt.de/texte/texte2.htm#georgant and http://armazi.uni-frankfurt.de.

[3] For the example see http://titus.fkidg1.uni-frankfurt.de/database/titusinx/titusinx.asp? LXLANG=517&LXWORD=uplisatws&LCPL=1&TCPL=1&C=A&T=0&LMT=100&K=0&MM=0&QF=1; the search engine (see http://titus.fkidg1.uni-frankfurt.de/search/query.htm) accepts both Georgian and Romanized input.

[4] Cf. http://clarino.uib.no/gekko.

[5] For the example see http://clarino.uib.no/gekko/simple-query (subcorpus: Georgian – disambiguated (ქართული-დის.); query input: [lemma="ყოფნა"]).

in Tbilisi.[6] This corpus, which also includes varieties of Georgian spoken outside of Georgia (in Iran, Turkey and Azerbaijan), has recently been made accessible for online retrieval via a word-based search engine (cf. the sample output in Figure 3).[7]

The GNC project further builds upon an extensive amount of recordings of spoken varieties of Georgian that were prepared within the project 'The sociolinguistic situation of present-day Georgia' (2005-2009).[8] Many of these materials have been fully transcribed (ca. 1.5 Mio. tokens) and are, for the time being, accessible via the TITUS server,[9] the Language Archive at the MPI Nijmegen,[10] and the GDC project (cf. Figure 4).[11]

The integration of all these data and functionalities, which is the main object of the two-year start phase of the GNC project begun in 2012, will bring together an unparalleled diachronic corpus extending over a time-span of about 1600 years and including chronological as well as dialectal and sociolectal variation.[12]

---

6 Cf. www.mygeorgia.ge/gdc/About.aspx and Beridze/Lortkipanidze/Nadaraia, this volume.

7 For the example see www.mygeorgia.ge/gdc/Default.aspx (query input: არის).

8 The project was kindly supported by the Volkswagen Foundation from 2005 to 2009.

9 Cf. http://titus.fkidg1.uni-frankfurt.de/ssgg/ssgg.htm.

10 Cf. http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI663243%23. To access the data users will have to register with the Language Archive; see http://dobes.mpi.nl/access_registration/ for instructions and http://corpus1.mpi.nl/ds/RRS_V1/RrsRegistration for the required form.

11 For the example (recording in the Atcharian dialect by N. Surmava, 2006) see http://corpus1.mpi.nl/ds/imdi_browser?openpath=MPI696092%23.

12 The project web site will be www.gnc.ge; for the time being cf. http://armazi.uni-frankfurt.de/gnc/gnc.htm and http://clarino.uib.no/gnc.

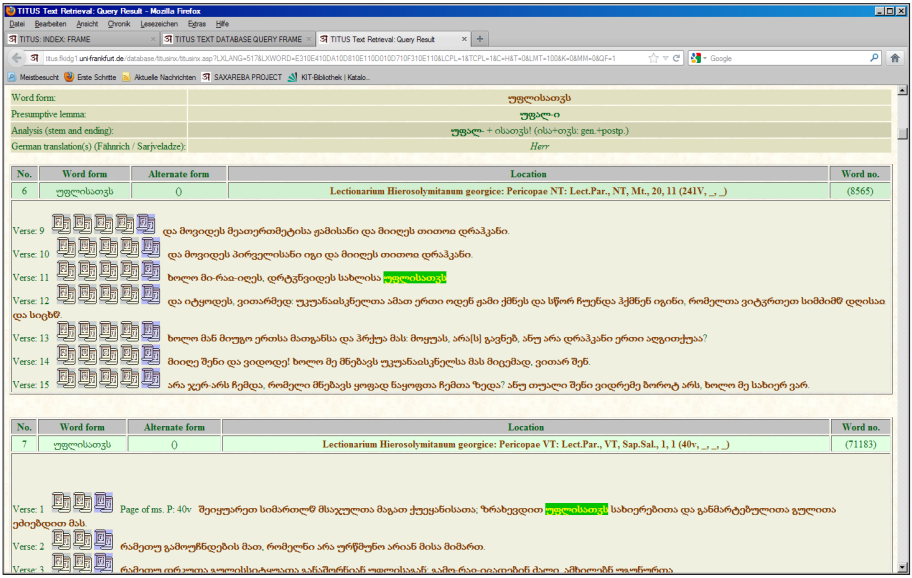

Figure 1: Query output of the TITUS search engine (*uplisatws* 'for the Lord'; cf. footnote 3)

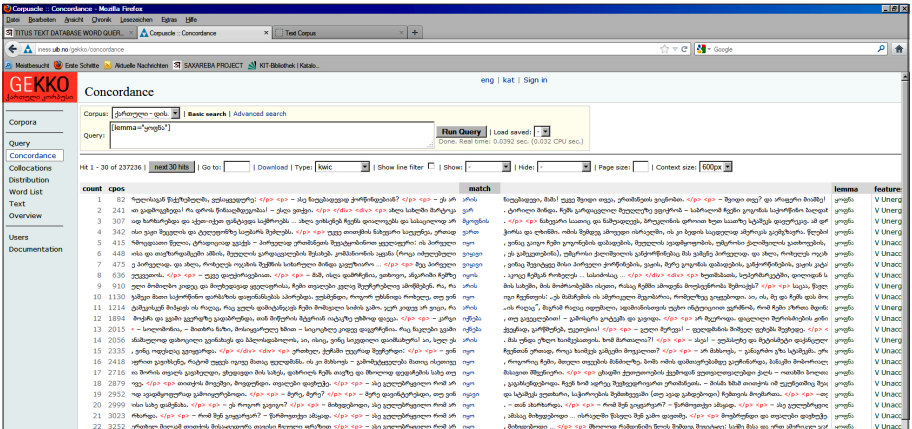

Figure 2: Search output of the GEKKO retrieval engine (lemma *ɣopna* 'to be'; cf. footnote 5)

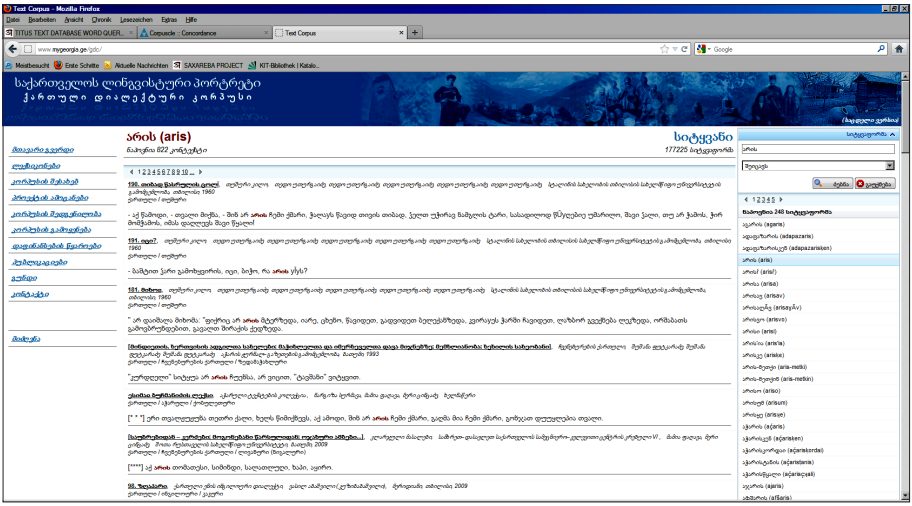

Figure 3: Search output of the GDC retrieval engine (word-form *aris* 'he/she/it is'; cf. footnote 7)

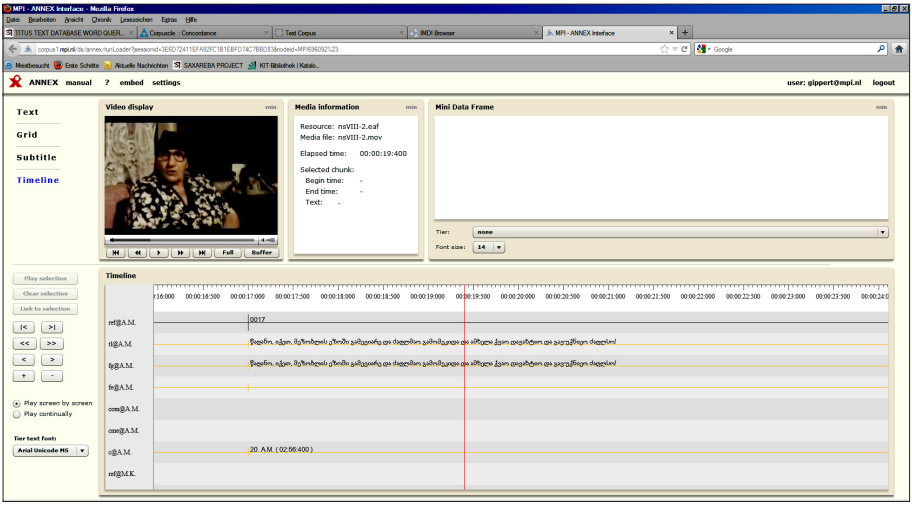

Figure 4: Dialect text from the SSGG project (via the ANNEX interface of the Language Archive at the MPI Nijmegen; cf. footnote 11)

3. Scripts and encoding

As a matter of fact, Georgian is a near-to ideal showcase to develop and test a "true" diachronic corpus, even though it has changed much less than other languages since it was first written; consider, e.g., a common word-form of today like *gmadlob* 'I thank you' which has not changed at all since its first attestation in a palimpsest manuscript of about the 6th c. C.E.,[13] in spite of the peculiar consonant clusters it contains. However, the literary history of Georgian was anything but homogeneous, the language having been written with three different scripts in the course of time: *Asomtavruli*, the Old Georgian majuscule script (ca. 5th-10th cc.), *Nusxa-Xucuri*, the minuscule script used in manuscripts of ecclesiastical content (ca. 9th-19th cc.), and *Mxedruli*, the mi-

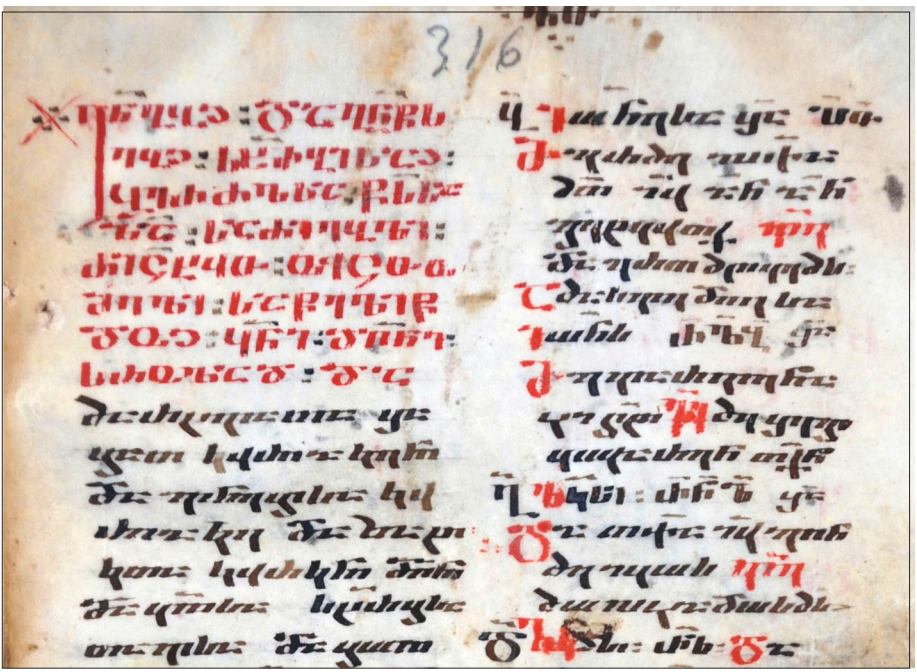

Figure 5: Old Georgian manuscript page (excerpt)

[13] In the so-called *Khanmeti* version (cf. 3.1 below) of the legend of St. Christina, preserved in the Codex georgicus no. 2 of the Austrian National Library, Vienna; see http://titus.fkidg1.uni-frankfurt.de/texte/etcc/cauc/ageo/xanmeti/vienna/vienn010.htm#Coll.Hag._Mart._Christin._II_8___022r-019v_22rb_1_(1)_186,19_f360,27 (users who are no members of the TITUS project will have to register via the form provided on http://titus.uni-frankfurt.de/titusstd.htm).

Figure 6: Modern Georgian press page (excerpt)

nuscule cursive used since the Middle Ages until the present day (cf. Figure 5 showing an Old Georgian manuscript written partly in *Asomtavruli* – in red ink – and partly in *Nusxa-Xucuri* script,[14] and Figure 6 showing a modern print in *Mxedruli* script).[15]

All three Georgian scripts have been assigned separate code points in the Unicode standard[16] so that it is possible today to encode the text materials of all periods as they were written originally. Under these conditions, an integrative approach towards diachronic retrieval across scripts presupposes the establishment of equivalences, which is straightforward for most letters. In a few cases, however, there are systematical discrepancies.

[14] Manuscript no. 16 of the (old) collection of Georgian manuscripts of St. Catherine's monastery on Mt. Sinai, fol. 316r (photograph J.G., 2009).

[15] From an article by H. Kurdadze on the Georgian alphabet in the Inflight Magazine of Georgian Airways 5, 2006/7: 4.

[16] Asomtavruli: U+10A0 – U+10C5; Nuskha-Khutsuri: U+2D00 – U+2D25; Mkhedruli: U+10D0 – U+10F5; see www.unicode.org/charts/PDF/U10A0.pdf and www.unicode.org/charts/PDF/U2D00.pdf, resp.

3.1 The notation of *u* and *v*

One discrepancy is determined by the fact that the *Asomtavruli* script inherited a peculiarity from its model, the Greek alphabet of Hellenistic times, in that it had no character for the vowel [u], which was written with a digraph <ႭჃ> equivalent to Greek <ΟΥ> instead. This digraph developed to a single letter, <ⴓ>, in the minuscule script and is still a single letter, <უ>, in the *Mxedruli* script of today. In the rendering of Old Georgian manuscripts, it has been usual practice for long to transcribe *Asomtavruli* into *Mxedruli*, and most scholarly editions of Old Georgian texts are printed in the modern script. The <ow> digraph is usually replaced by the single <უ> letter in these editions, albeit it could as well be represented by the corresponding digraph, <ოვ>, in *Mxedruli* script. Thus, the sequence ႱႭჃႪႬႤႪႤႡႨ' ႣႠ ႱႾႭჃႠႬႨ <sowlnelebi, da sxowani> 'aromatic spices, and others'[17] is usually transcribed as სულნელები, და სხუანი <sulnelebi, da sxuani>, not transliterated as სოვლნელები, და სხოვანი <sowlnelebi, da sxowani>. This, now, is problematic in a diachronic perspective as both nouns have slightly changed meanwhile, *sulnelebi* having been replaced by *surnelebi* while *sxvani* 'others' is written with the <v> character, <ვ>, instead of <უ> today (სხვანი). It is true that the replacement of <ow> by <v> is near-to regular in the given environment (between a consonant and a vowel); however, there are cases where <u> has been maintained in the same constellation (e.g., in Modern Georgian ჭკუა <č̣ḳua> 'intellect') or, conversely, <v> was used in the same position already in Old Georgian (e.g., in ႵႥႠჂ <kvay> 'stone') so that the application of an automatic substitution rule may fail. For the change leading from *sulnel-* to *surnel-*, there is no "automatic" rule at all as the dissimilation involved here is sporadic, not regular.[18]

17 From Lc. 24.1 in the so-called "Pre-Athonian" redaction of the Old Georgian NT (9th-10th cc.), first attested in the so-called "Graz Lectionary" (manuscript Gr. 2058/2 of the Graz University Library, fol. 8r), a Sinai codex of ca. the 8th c. mixing Khanmeti and Haemeti features; see http://titus.uni-frankfurt.de/texte/etcs/cauc/ageo/xanmeti/grlekt/grlek.htm?grlek017.htm#Gr._2058/2_8r_3_Lk._24_1.

18 In contrast to the regular dissimilation rule of Modern Georgian which changes a sequence of *r* – *r* into *r* – *l* as in the adjective formation suffix *-ur-* (see, e.g., *čex-ur-i* "Czech") appearing as *-ul-* in *rusuli* "Russian" or *german-ul-i* "German". Modern Georgian does admit of sequences of *l* – *l* as in *alubali* "cherry". – The Old Georgian stem *sulnel-* can still be found used (as an obsolete form) in religious contexts today.

3.2 The notation of *wi*

The second element of the *Asomtavruli* digraph <ႭႣ>, the letter *vie*, <Ⴣ>, is problematic in other contexts, too. As the descendant of Greek <Υ>, it usually stands for a diphthong-like [wi] sequence (resulting from or replacing Greek [ü]); the same is true for its *Nusxa-Xucuri* equivalent, <ⴣ>. In such cases, modern transcriptions replace <Ⴣ> either by *Mxedruli* <ვი>, i.e. <vi>, in accordance with the modern pronunciation, or by transliterative <ჳ> = <w>, as in ⴂⴣⴐⴂⴣⴌⴈ <gwrgwni> 'crown' rendered by either გვირგვინი <gvirgvini> or გჳრგჳნი <gwrgwni>. Again, these replacements are not straightforward as they are not applicable when <Ⴣ> or <ⴣ> follows or precedes a vowel. What is more, the orthographic rules of Old Georgian manuscripts differ to a considerable extent in the use of the character. For instance, we often meet with the sequence [wi] being represented by the digraph <ႭႣ>, i.e. <ow>, instead of plain <Ⴣ> = <w>, or [v] in post-vocal position being rendered by <Ⴣ> = <w> or <ႭႣ> = <ow> instead of <Ⴅ> = <v>; cf., e.g., *Asomtavruli* spellings like <xowdodit> instead of "normal" <xwdodit>, <iṭq̇ows> instead of <iṭq̇ws>, <simšowidita> instead of <simšwdita>, or <moaowlina> instead of <moavlina>, all appearing in the lower layer of the palimpsest pages of the Kurashi Gospel manuscript.[19]

3.3 Diplomatic rendering vs. diachronic retrieval

All these graphical discrepancies must be taken into account if the corpus is meant to reflect the manuscript heritage of Old (and Middle) Georgian as neatly as possible (in the sense of a "diplomatic" rendering of hand-written sources) and yet to provide diachronic access to its linguistic contents. To cope with these demands, it is desirable to envisage a multilevel annotation format that is able to store authentic spellings, period-conformant normalizations, and diachronic surrogates side by side. A similar approach has been worked out for the project "Referenzkorpus Altdeutsch" (cf. Figure 7), which is to be diachronically aligned with corpora of later stages of German to yield a diachronic corpus of all periods of German.[20] In such an annotation system, an Old Georgian spelling variant like ႱႾႣႠႬႨ (= <sxwani>) should be stored as-is (i.e., in *Asomtavruli*) alongside its "normalized" Old Georgian equiva-

---

[19] Cf. Gippert (2013: 113).

[20] Cf. www.deutschdiachrondigital.de for the project of a "sprachstufenübergreifendes tiefenannotiertes Korpus historischer Texte des Deutschen".

lent, სხოვანი (= <sxowani>), as well as its "modern" adaptation, სხვანი (= <sxvani>), and its lemmatic basis, the stem სხვა- (= <sxva->), the latter representing the entry point for diachronic queries.[21]

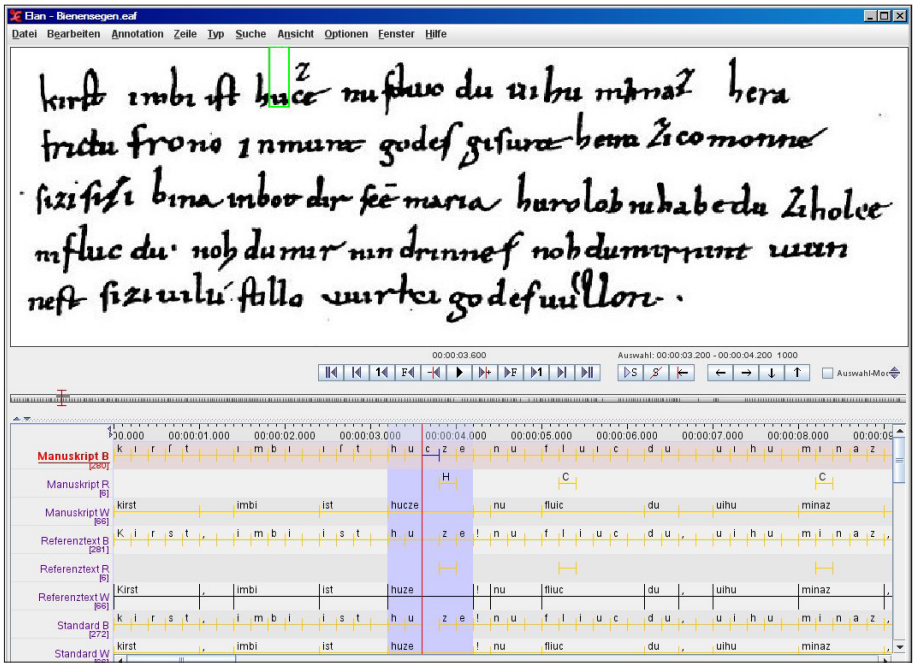

Figure 7: Old High German manuscript with multilevel annotation distinguishing diplomatically rendered and normalized spellings (from the "Referenzkorpus Altdeutsch" project)[22]

4. The annotation of linguistic varieties

While unified access to linguistic elements across the history of a language is one fundamental task of a diachronic corpus, the differentiation of the individual varieties comprised in it is another one. For more detailed investigations into the historical diversification of a given language, it is necessary to distinguish the different layers the textual materials pertain to, especially

[21] It is true that the plural form *sxvani* is rare today, the regular plural form being *sxvebi*. The so-called "old" plural forms require special treatment in the corpus.

[22] The so-called "Lorscher Bienensegen" contained in the Vatican manuscript Cod.lat.Pal. 220, p. 58r; see http://titus.uni-frankfurt.de/texte/etcs/germ/ahd/klahddkm/klahd.htm?klahd091.htm#Kl.ahd.Dkm._Bienenseg._1_S396__Vat.lat._220_58r_47.

with a view to lexicographic analyses. This concerns not only the successive chronological layers – in the given case Old, Middle and Modern Georgian – but also other layers that are distinguishable in the data, among them dialects, sociolects, and registers determined by text genres (i.e., "styles") and communication modes (e.g., "spoken" vs. "handwritten" vs. "printed" vs. "electronic", etc.). In the case of Georgian, this is crucial indeed, as the diversification of discernible layers begins as early as the Old Georgian period.

4.1 Layers of Old Georgian

As a matter of fact, a large set of layers must be distinguished for Old Georgian with respect to chronological, regional, and other properties. Chronologically, the set begins with the so-called *Khanmeti* and *Haemeti* varieties, which represent the earliest strata of Georgian literacy (ca. 5th-7th and 7th-8th cc.), with a "mixed" variety attested in the famous Graz lectionary,[23] and which are clearly distinguishable by peculiar morphological features. Within the subsequent period of "standard" or "classical" Old Georgian (ca. 9th-12th cc.), we may distinguish several locally-based varieties mostly established by Georgian writers in the monastic diaspora, on Mt. Sinai, Mt. Athos, or in Palestine, but also within Georgia as in the case of the "Gelati" school of the 11th-12th c. C.E.[24] Albeit most of the textual material of Old Georgian is religious, there are still some genre-specific peculiarities that force us to distinguish authentic from translated sources, and among them Biblical, hagiographical, homiletic, historiographical, philosophical, documentary, and other styles. A peculiar layer of Old Georgian is met with in documents that emerged later than the 12th c., in an attempt to maintain the religiously determined Old Georgian standard of literacy alongside the developing "Middle" Georgian vernacular which mostly manifested itself in secular literature; this layer, which had its impact up to the 18th c., may be called "Late Old Georgian".

4.2 Layers of Middle and Modern Georgian

Different from Old Georgian, the Middle Georgian period was much less characterized by chronological or local differentiation. Instead, it was marked by greater genre-specific differences between, e.g., poetic, epic, historiographic, or documentary texts, manifesting themselves mostly in lexicographic fea-

[23] Cf. footnote 17 above.

[24] Cf. http://armazi.uni-frankfurt.de/armaz1m.htm.

tures (e.g., in an increasing impact of Persian) but also in the degree of grammatical conservativeness. Thus, the Old Georgian phenomenon of verbal tmesis (e.g. *mo-vinme-vida* 'someone came', lit. 'hither-someone-went', with the preverb *mo-* being split from the verbal root *-vid-* by the inserted indefinite pronoun *vinme* 'somebody') is still attested frequently in the 13th c. epics (prose and verse) but only exceptionally later. A similar distinction of genre-based registers is applicable to written Modern Georgian, too; here we would have to distinguish, right from the beginning, poetic and prose genres, the latter including belletristic literature as well as journalistic, juridical, scientific, or other styles. And of course, there were changes in the orthographical standards and the grammar within the period of Modern Georgian, too, which can roughly be divided into three subperiods in this respect, viz. Tzarist, Soviet, and contemporary.

4.3 Dialectal and sociolectal variation

As in other languages with a strong literary standard, dialectal and sociolectal variation comes into play mostly in spoken manifestations of Georgian. Roughly speaking, the Georgian dialects form two subgroups, a western and an eastern one, with Kartlian, the dialect of the central eastern part of the country and its capital, Tbilisi, being closest to the written language of today. Among noteworthy sociolects, we may mention that of the Georgian Jewry, which is characterized by a peculiar terminology (not necessarily of Hebrew origin) and a special intonation, or the argots of thieves or drug dealers, which have characteristic lexical features, too.

4.4 Annotation of layers

How, then, to account for all these divergent layers in a diachronic corpus of Georgian? Traditionally, the pertinence of a text to a given layer has been regarded as meta-information that is best stored in a (TEI) header. This, however, has a big disadvantage as it cannot account adequately for mixed texts such as, e.g., prose texts containing verse passages, journalistic texts containing quotations from argot speech, or even hagiographical texts containing quotations from the Bible or passages in foreign languages.[25] The annotation of information concerning chronological, dialectal, sociolectal and other layers would in

[25] A striking example of the latter is an Early New Persian sentence quoted, in Georgian script, in the "Life of St. Nino" (see Gippert 1992: 10).

such cases better be stored word-by-word in order to facilitate layer-based queries and indexation. This can again be achieved via a multilevel annotation scheme; cf. Figure 8 illustrating this with another example from the "Referenzkorpus Altdeutsch" project where Old High German and Latin words are annotated accordingly using the respective three-letter codes of the ISO 639-3 standard[26] ("goh" = "German-Old-High" and "lat" = "Latin").

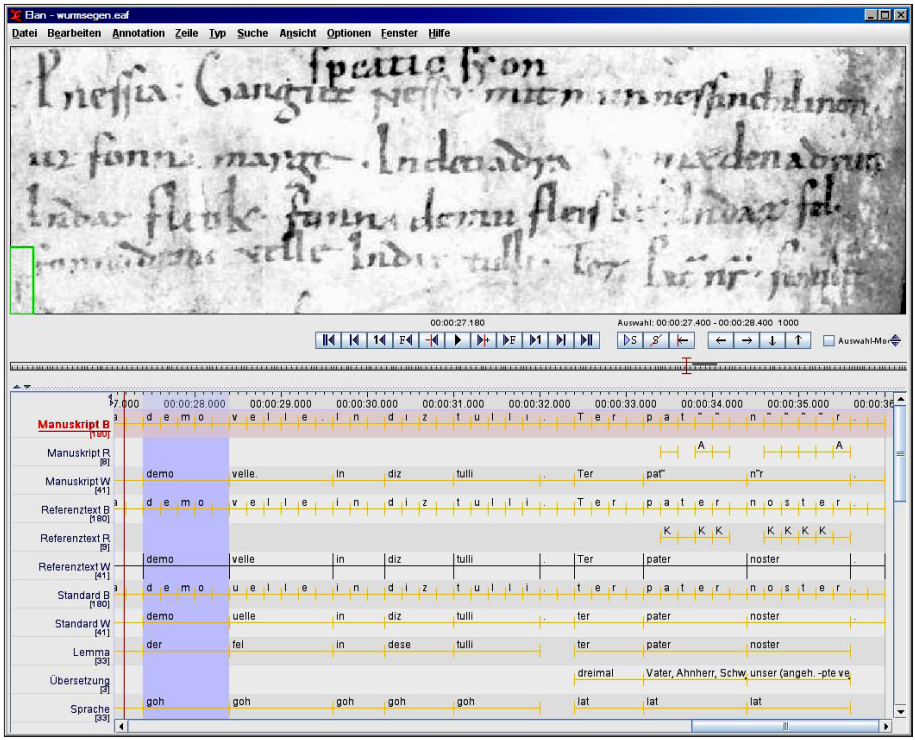

Figure 8: Old High German manuscript with multilevel annotation distinguishing Old High German ("goh") and Latin ("lat") words (from the "Referenzkorpus Altdeutsch" project)[27]

[26] For the standard see the official site of the ISO 639-3 Registration Authority at www.sil.org/iso639-3/.

[27] The so-called "Wurmsegen" (no. 1) contained in the Tegernsee manuscript Clm 18524b, p. 203v of the Bayerische Staatsbibliothek, Munich; see http://titus.uni-frankfurt.de/texte/etcs/germ/ahd/klahddkm/klahd.htm? klahd077.htm#Kl.ahd.Dkm._Wurms.1_3_S374__Clm_18524b_203v_47.

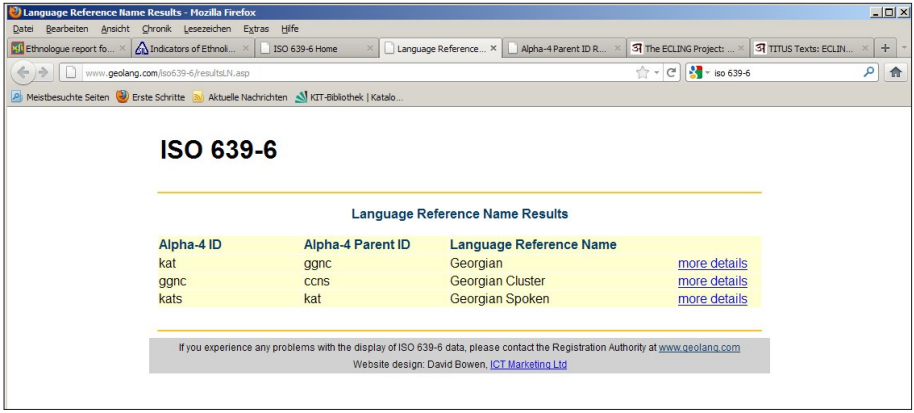

Figure 9a: The ISO 639-6 database (query result for "Georgian")

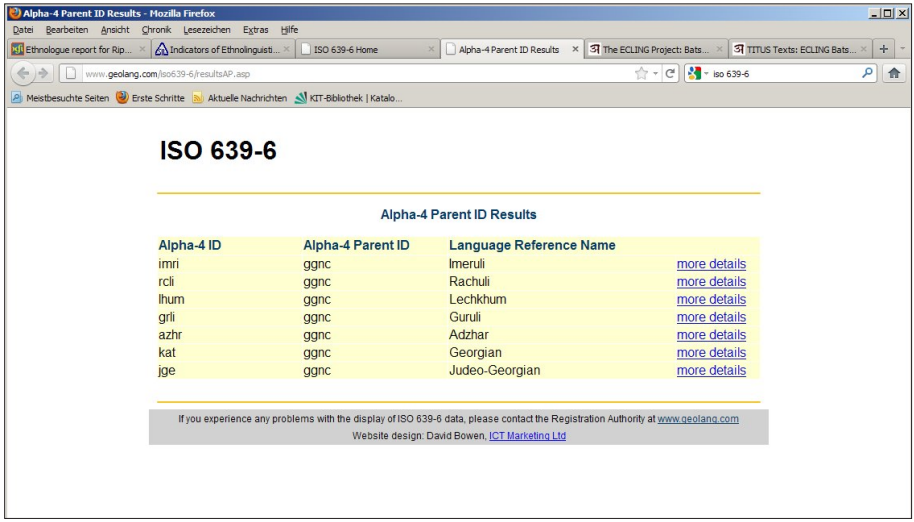

Figure 9b: same (query result for subnode GGNC)

It goes without saying, however, that a three-letter-code of this type is in no way sufficient to cover the diversity of chronological, dialectal, and other layers we have to deal with in the GNC project, all the more since ISO 639-3 distinguishes nothing but "Georgian" (= "kat", ← *kartuli*, the self-designation of the language) and "Old Georgian" (= "oge"). The reduced amount of possible codes in this standard (26^3 = 17,576 possible combinations of three basic letters) has recently led to the foundation of a successor standard, ISO 639-6,

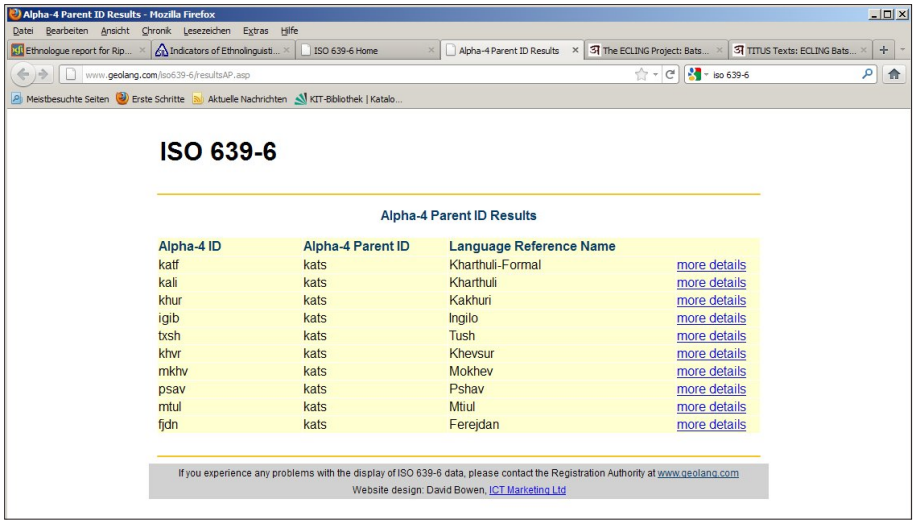

Figure 9c: same (query result for subnode KATS)

which operates with four letters, yielding a total of (26^4 =) 456,976 new codes.[28] In its first stage of development, this standard comprised[29] a set of 20 codes related to Georgian and its varieties, arranged as parent-child relations in a tree-like structure. The picture thus achieved was anything but convincing, however, let alone sufficient for our purposes. First, there were no codes available concerning older stages of Georgian, not even the "oge" code of ISO 639-3, albeit the codes of this standard were declared to be maintained in the new proposal and "kat" for "Georgian" was still present (cf. Figure 9a showing the output of a query for "Georgian" in the database of the site that has been responsible for the registration since 2009).[30] Second, there was no differentiation in the codes as to dialectal and sociolectal layers; thus, "jge" for "Judeo-Georgian" (again taken over from ISO 639-3) was registered on the same level as, e.g., the Rachian dialect of western Georgia ("rcli" = "Rachuli"; cf. Figure 9b). Third, the tree structure remained enigmatic, given that nine Georgian dialects (plus "katf" = "Georgian formal") were subsumed as children under

[28] Cf. www.iso.org/iso/catalogue_detail?csnumber=43380 for a rough outline and Gippert (2012: 21-23) for a preliminary account of the standard.

[29] The official Registration Authority for the standard was until 2014 the British company *GeoLang* (now *Ascema*; http://www.geolang.com/).

[30] The URL in question (www.geolang.com/iso639-6/) was still available on January 28, 2015, but not working properly. It seems that the process of further developing ISO 639-6 has been interrupted.

"kats" = "Georgian spoken" (cf. Figure 9c), whereas six other dialects (plus "jge" = "Judeo-Georgian") were children of "ggnc" = "Georgian cluster", in its turn the parent of "kat" and the grand-parent of "kats" (cf. the schematic illustration in Figure 10). The very fact that the nine first-mentioned dialects pertain to the eastern group and the six other ones, to the western group, is in no way a satisfactory explanation why only the former depended on "kats" = "Georgian spoken" (and, further up, "kat" = "Georgian").

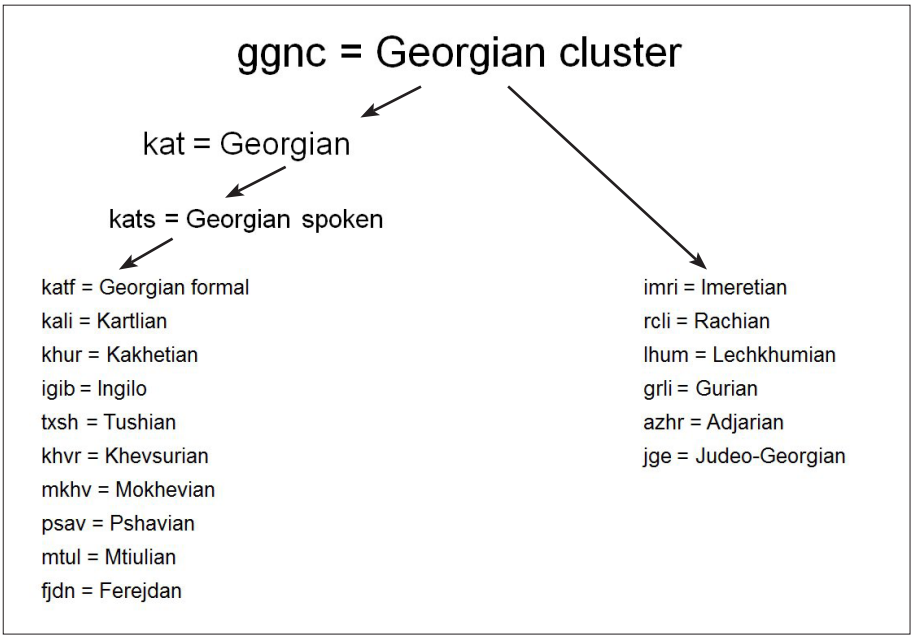

Figure 10: Dependencies of Georgian varieties in ISO 639-6

As a matter of fact, it is more than doubtful that the complex interrelationship between the chronological, dialectal, and other "lectal" layers of Georgian can at all be depicted adequately in a flat tree structure of the given sort. Instead, we should rather conceive this as a set of matrixes, among them one of spoken dialects interacting with sociolects, one of chronological layers interacting with genre-based variants (as illustrated in Table 1), and others.

	Old Georgian						Middle	Modern Georgian		
	Khanmeti	Haemeti	Standard	Sinai	Athonite	Late	Middle	Early	Soviet	Contemporary
Biblical	×	×	×	×	×	×		×		×
Hagiographical	×		×	×	×	×		×		
Homiletic	×	×	×	×	×	×		×		×
Theological			×	×	×	×		×		×
Historical			×			×	×	×	×	×
Philosophical			×			×	×	×	×	×
Documentary			×				×	×	×	×
Epigraphic	×	×	×	×		×	×	×	×	×
Scientific			×	×		×	×	×	×	×
Journalistic								×	×	×
Belletristic							×	×	×	×
Poetical			×	×		×	×	×	×	×

Table 1: Matrix of chronological layers of written Georgian and text genres attested in them

In such a system of matrixes, the information that a given word belongs to a biographical text in a record of spoken Judeo-Georgian from Kutaisi in West Georgia would not be covered by a mere three- or four-letter-code such as "jge" but by a set of specifications (structured, e.g., as a sequence of codes for "language – chronological layer – mode – dialectal area – sociolect – genre", in the given example quasi "Georgian – Modern – Spoken – Imeretian – Jewish – Biographical").[31] This concept would not only allow us to keep different types of "lects" apart but also to assign common "layers" (modes, genres, and, to a certain extent, sociolects) cross-linguistically. The development of a repertoire of annotation codes for these purposes, ideally to be standardized, is a

[31] A similar approach has been outlined by the WordWideWeb consortium (www.w3.org/International/articles/language-tags/Overview.en.php), which proposes *language-extlang-script-region-variant-extension-privateuse* as a sequence of "types of subtag" (cf. also the discussion in www.rfc-editor.org/rfc/rfc5646.txt). This sequence has the shortcoming that there is no clear distinction between chronological, dialectal, sociolectal, and genre-specific layers, all to be covered by the "extended language" (*extlang*), *region*, and *variant* subtags.

task of high priority indeed. The GNC project will contribute to this in elaborating and schematizing the distinctions surfacing in the text materials it covers. This process will also be the basis for determining the necessary extensions of the corpus with a view to an optimal balancing between the text genres and "lects" reflected in them.

References

Gippert, Jost (1992): Zum Status des Mittelpersischen im südlichen Kaukasus. Paper read on the conference "Bilingualism in Iranian Cultures", Bamberg, July 1992. http://titus.uni-frankfurt.de/personal/jg/pdf/jg1992b.pdf.

Gippert, Jost (2012): Language-specific encoding in endangered language corpora. In: Seifart, Frank/Haig, Geoffrey/Himmelmann, Nikolaus P./Jung, Dagmar/Margetts, Anna/Trilsbeek, Paul (eds.): Potentials of language documentation: methods, analyses, and utilization. Honolulu: University of Hawai'i Press, 17-24. http://hdl.handle.net/10125/4512.

Gippert, Jost (2013): The Gospel Manuscript of Kurashi. A preliminary account. In: Le Muséon 126: 83-160.