3

Language Documentation & Conservation Special Publication No. 3 (August 2012):
Potentials of Language Documentation: Methods, Analyses, and Utilization,
ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann,
Dagmar Jung, Anna Margetts, and Paul Trilsbeek, pp. 17–24
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4512

Language-specific encoding in endangered language corpora

Jost Gippert
Goethe University of Frankfurt/Main

The paper addresses problems of corpus building and retrieval resulting from code-switching, which is a characteristic feature of endangered language recordings. The typical appearance of code-switching phenomena is first outlined on the basis of data collected in the DoBeS 'ECLinG' project, which dealt with three endangered Caucasian languages spoken in Georgia: Tsova-Tush (Batsbi), Udi, and Svan. The problem of language-specific retrieval is illustrated with examples showing the usage of the word *da* in Tsova-Tush contexts, which represents, as a homonym, either a native copula form ('it is') or the Georgian conjunction 'and'. The subsequent section discusses the annotation requirements that are necessary to automatically distinguish the languages involved in code-switching, with a focus on the emerging ISO standard 639-6. It is argued that the fine-grained distinction of varieties and subvarieties and their interrelationship – as aimed at in this standard – requires a thorough reconsideration if it is to be applied in the markup of corpus data.

1. INTRODUCTION. It is well known that recorded texts of natural speech in endangered languages abound in code-switching, mostly between the endangered vernacular and dominant languages, but also other languages involved in the bi- and multilingual settings that are typical for language endangerment. This multilingual data is crucial for all kinds of language-specific or cross-linguistic research into endangered languages, as well as for the theory of language endangerment in general (see also Gullberg, this volume) .[1] However, at present, annotation schemes such as those developed in the DoBeS framework do not admit of an easy differentiation of linguistic units pertaining to different linguistic layers, and language-specific search functions are still wanting. The present paper first illustrates the presence of multiple languages in the documentation of Caucasian languages (section 2) and then discusses ways to cope with this, considering, among other things, the advantages of the emerging ISO standard 639-6 'Language Names' (section 3).

[1] Cf. Gippert 2008: esp. 174–188, for a case study based on the three Caucasian languages Svan, Tsova-Tush and Udi. Cf. Gullberg (this volume) for a more general view on the impact of bi- and multilingualism in endangered language communities for linguistic theory.

2. GEORGIAN ELEMENTS IN TSOVA-TUSH (BATSBI) AND UDI. The effect of a missing distinction between the languages involved in bi- or multilingual settings can easily be demonstrated with the materials that were collected between 2003 and 2007 by the DoBeS 'ECLinG' project, which addressed three endangered Caucasian languages spoken in Georgia, viz. Svan, Tsova-Tush (Batsbi), and Udi. In the text recordings that were provided by the project to the DoBeS Archive[2], we can clearly see that Georgian as the dominant language of the area in question has left its traces everywhere in both monologic and dialogic speech of speakers of all generations.

In the case of Tsova-Tush, an East-Caucasian ('Nakh') language closely related to Chechen and Ingush but unrelated to (South-Caucasian) Georgian, this has brought about a peculiar homonymy, given that one of its most frequent verb forms, the copula form *da* '(it) is', is indistinguishable from the most frequent particle of Georgian, the conjunction *da* 'and'. Executing an 'annotation content search' for the word form *da* in the DoBeS Tsova-Tush materials with the TROVA tool[3] yields both Tsova-Tush and Georgian contexts from the annotated text recordings,[4] with the latter representing ca. 10%. Among them we find Georgian *da* 'and' in the following contexts:

(a) utterances not pertaining to a given narrative but addressing people present in the recording session as in *is sḳami, gadmodit, švilo,* ***da*** *axlos daǯekit, ḳaco* 'That chair, come over, boy, **and** sit down close by, man!';[5]

(b) sentences of reported Georgian speech inserted into a Tsova-Tush narrative as in the case of *ṗeṭresac avaġebine švebulebao* ***da*** *čamovedit alvanšio…* "'I will make Peter take a vacation (too), **and** let's go down to Alvani"…(he said)', introduced by Tsova-Tush *ḳoxiv var – ḳoxiv…* 'He was a Georgian, a Georgian';[6]

(c) idiomatic formulae such as *me magisi ase* ***da*** *ise* 'I …his …this **and** that way', i.e., 'I could do his mother this and that', embedded between the Tsova-Tush sentences *oḳuyxvas āl'i*^{n} *sog mē aḥ bʿivnaḥēn* 'That Kakhetian (man) said to me, "you killed

---

[2] Cf. http://corpus1.mpi.nl/ds/imdi_browser?openpath=MPI533677%23. The recordings stored in the archive consist of about 500 texts (ca. 70 hrs.) for Svan, 312 texts (ca. 37 hrs.) for Tsova-Tush, and 43 texts (ca. 6 hrs.) for Udi; all recordings are fully transcribed and provided with a Georgian and English translation, ca. 10% additionally with a multilevel grammatical analysis.

[3] Cf. http://corpus1.mpi.nl/ds/annex/search.jsp?transferuid=1&nodeid=MPI534223%23&row=29. The search yields 774 hits (27.11.2011, 20:01h) in 'Single Layer' mode set to 'exact match'. A similar result (768 hits) is achieved searching for *da* in Georgian script because the annotations were mostly provided in both Georgian and Latin scripts.

[4] As an example of Tsova-Tush *da* 'it is' we may quote the sentence *vir ma aṭṭan da, ḳaċḳon daʾoḥe, davina da* 'However, the donkey is light, it is small, it is light' (from a monologue on donkey breeding, http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI793920%23&time=170662&duration=666&tiername=trs@AS; the sentence in question starts at 00:02:46 in the recording).

[5] In a monologic account on Tushian house-building, cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI793880%23&time=107000&duration=500&tiername=tl@EC, sentence starting at 00:01:45.

[6] In a biographical narrative, cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI793894%23&time=215816&duration=727&tiername=tl@EC, sentence starting at 00:03:33. Note that the interviewer admonishes the narrator to return to the Tsova-Tush language by interjecting *vēġeš…vēġeš…vēġeš…*, i.e. 'in our (language), in our (language), in our (language)' after the quoted sentence. – Note also that Tsova-Tush *var* 'he was' and Georgian *var* 'I am' form another remarkable pair of homonyms.

it"', and *as ōqpinivḥē, beḳxētlex co vas^{o}, moḥ āl'i^{n}* 'I had released him, (so) I am really surprised how he could say (so)';[7]

(d) in insertions of Georgian geographical denominations such as *zemo* ***da*** *kvemo alvani* 'Upper **and** Lower Alvani' (in the given case dependent as a quasi-genitive on *amgēgmav* 'planner', which is in turn an integrated loan from Georgian *(da)mgegmavi* 'id.'),[8] but also in

(e) Georgian phrases mixed without any obvious motivation with Tsova-Tush contexts as in the case of *isev ċava* ***da*** *mova, ċaiq̇vans* 'she (the cat) will go **and** come back (and) bring it away', linked as an apodosis to a Tsova-Tush protasis, *me qe eyl'čeḥatx me dolix o qenaʾā doʾdebēn* 'afterwards, when we tell her "come on, bring that other one away, too"'.[9]

In one instance, we even find the inner-Georgian homonym of *da* 'and', viz. the noun *da* 'sister', in a Georgian sentence embedded in a Tsova-Tush context: with *beladiant enċeras… beladiant elane ro ari, ḳutxeši rom cxovrobs moxuci kali, ai imis* ***da*** *iq̇o* 'Beladianti Entsera…(she) who is Beladianti Elane, the old woman who lives at the corner, look, her **sister** it was', the speaker replies to the question *veċer ḥan?* 'Who was in love with him?'.[10]

In the case of Udi, things are different in that we have no homonymous equivalent of Georgian *da* in this language. Nevertheless we arrive at 63 hits of *da* searching through the Udi ECLinG corpus,[11] all representing the Georgian conjunction. Different from the Tsova-Tush examples illustrated above, we here even find cases where *da* is not used in a longer Georgian phrase or chunk but isolated, in a plain Udi context, as if being a loan; cf., e.g., *zu qayzupe meṭġo tärämišbaki garxox* ***da*** *evaxṭe qayzupe…* 'I discovered the places of their emergence, **and** when I had discovered them…',[12] or *mya buyanq̇e q̇eiri, q̇eiri ʒüra ḳinigiux serbayan, manote bakale oxari äʿyluġo baxṭink,* ***da*** *manote äʿyluġo sṭudenṭġo baxṭink, aspiranṭur baxṭink…* 'Here we want to make other, other types of books, which

---

[7] In a narrative on a shepherd's life, cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI689484%23&time=296926&duration=642&tiername=tl@EC, sentence starting at 00:04:53. The fixed Georgian formula *mainc da mainc* 'nevertheless, however' occurring several times in Tsova-Tush contexts can already be taken to be a loan.

[8] In the biographical narrative mentioned above, cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI793894%23&time=296800&duration=600&tiername=tl@EC, phrase starting at 00:04:55.

[9] In a dialogue on cat breeding, cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI793914%23&time=145800&duration=400&tiername=tl@CD, sentence starting at 00:02:21.

[10] In a dialogue discussing the contents of a folk song, cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI793871%23&time=30929&duration=533&tiername=tl@EA, sentence starting at 00:00:24.

[11] With the TROVA tool; cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI1360405%23&time=247290&duration=1458&tiername=tl@MN.

[12] In a monologue on the origin of the Udi people, cf. http://corpus1.mpi.nl/ds/annex/runLoader?jsessionid=BE2633512018044F706DBF19BD1CDDAF&nodeid=MPI1360405%23&time=247290&duration=1458&tiername=tl@MN, sentence starting at 00:04:00.

will be easy (to read) for the children, **and** which (will be) for the children, the students, for becoming aspirants. . . '[13]

3. DEMARCATION OF LANGUAGES. It is clear that any automatic retrieval mechanism aiming at a distinction of the Georgian elements in Tsova-Tush or Udi contexts presupposes an adequate demarcation of the languages in question. In the ECLinG project, we have, for lack of more suitable means, started by inserting curly braces to mark the beginning and end of Georgian insertions.[14] This is not an expedient method, however, as braces may easily be neglected by retrieval engines (and the TROVA search function of the DoBeS Archive does neglect them). Instead, a consistent language-specific retrieval would require the linguistic affinity to be marked for every single word form,[15] a task that can easily be achieved using a semi-automatic annotation software such as the Summer Institute of Linguistics (SIL) Toolbox[16] where the information in question can be stored in a lexicon and transferred to annotation tiers in the text files (see Figure 2 below). This, however, presupposes a thorough grammatical analysis of the texts which would require the morphology of the 'mixed-in' language to be accounted for alongside that of the 'basic' vernacular (cf. the case of Georgian *da* sister in the Tsova-Tush example above which would have to be defined as a Georgian nominative or absolutive singular). This task, too, could be fulfilled in connection with an additional lexicon-based markup, but 'complete' grammatical annotations of this type cannot always be provided in the course of a given documentation project. As a matter of fact, only ca. 10% of the ECLinG data could be prepared in this way so that the searches are mostly restricted to the sentence level, which does not allow for a markup of individual words.

A peculiar problem arises if a language-specific search is to be executed not within a given corpus (with, maybe, an idiosyncratic demarcation of languages) but across resources of different origins. In this case it is inevitable to provide the information as to the linguistic affinity of word-forms in a standardized way. As a matter of fact, unique codes denoting languages have been the object of standardization endeavors for many years,[17] and computer users have for long been acquainted with two-letter codes such as EN for English or DE for German indicating the keyboards they use or other language-relevant information.[18] Dealing with endangered vernaculars, two-letter codes of this type are of little help, however, given that it is a maximum of (26^2 =) 676 languages that can be assigned by a pair of characters, and languages such as Tsova-Tush/Batsbi, Udi, or Svan are not among those

[13] In a monologue on the foundation of an Udi school, cf. http://corpus1.mpi.nl/ds/annex/runLoader?nodeid=MPI1360403%23&time=56676&duration=473&tiername=tl@MN, sentence starting at 00:00:51. – A notable Georgian-Udi homonym occurring in the texts is *xe*, which means 'water' in Udi and 'tree' in Georgian.

[14] In a similar way, square brackets have been used to denote Russian passages. The same denotations were also used in the materials of Caucasian languages recorded in the 'SSGG' project ('The sociolinguistic situation of present-day Georgia', project funded by the Volkswagen Foundation from 2005 to 2009) which are as well stored in the Archive of the MPI Nijmegen (cf. http://corpus1.mpi.nl/ds/imdi_browser?openpath=MPI663243%23).

[15] A prototypical distinction of linguistic affinities as represented in the ECLinG and SSGG recordings has been developed for the TITUS search engine which covers the texts of the recordings, too (cf. http://titus.fkidg1.uni-frankfurt.de/database/titusinx/titusinx.htm).

[16] Cf. http://www.sil.org/computing/toolbox/ for the software in question.

[17] They are the objects of the ISO standard 639 ('language names').

[18] The two-letter codes are standardized in ISO 639-1, a sub-standard of ISO 639.

registered in the standard. To overcome this, a standard consisting of three-letter codes (ISO 639-3) was conceived a few years ago,[19] under the aegis of the SIL as a 'registration authority'.[20] Albeit this standard would theoretically comprise (26^3 =) 17,576 entries, only about 6,900 codes have been assigned so far, obviously in accordance with SIL's 'Ethnologue' and the '6,909 languages' identified in it.[21] It is true that among these, we do find codes for Tsova-Tush ('Bats', BBL), Udi (UDI), and Svan (SVA), but there is no distinction possible yet of dialectal variants such as Upper Bal, Lower Bal, Lashkh, and Lentekh in the case of Svan or Vartashen (Oghuz) and Nidzh (Nij) Udi.[22] We must further consider that the elements 'mixed in' in code-switching are not necessarily representative of a given 'standard language' but usually dialectally or sociolectally biased.[23] Therefore it is clear that a much more fine-grained reference system is needed to adequately represent the diversity we are dealing with in the contexts of endangered languages.

Such a reference system has recently been initiated, with the four-letter code inventory of ISO 639-6, which is meant to cover all human language varieties including dialects, sociolects, historical stages, and the like. Different from the former sub-standards of ISO 639, the new standard, which implies a maximum of (26^4 =) 456,976 individual assignments,[24] is not restricted to a mere list of entries but comprises information as to the mutual interdependency of entries in terms of parent-child-relations; a system that would help a lot indeed if, e.g., a given search is not to be restricted to a given variety but to be expanded to a larger scope. Unfortunately, a first analysis of the standardization work undertaken by the responsible Technical Committee of the International Standardization Organization (ISO/TC 37/SC 2) and the institution authorized for the registration of the codes[25] reveals remarkable inconsistencies in the varieties accounted for and their hierarchical arrangement. E.g., we do find 'Spoken Bats' with the code BBLS as a 'child' of Bats, i.e. Tsova-Tush (BBL), and the latter is correctly subordinated to NXAX, i.e., the 'Nakh' subfamily of (North-)East-Caucasian languages (CCNE). Similarly, we find the Tushian ('Tush') dialect of Georgian (TXSH) as a child of KATS, i.e. 'Georgian spoken', in its turn depending on KAT = 'Georgian', which is a child of GGNC = 'Georgian cluster' and a grand-child of CCNS = 'South Caucasian'. On the other hand, Georgian dialects such as Imeretian ('Imeruli', IMRI), Rachian ('Rachuli', RCLI), Gurian ('Guruli', GRLI) or sociolects such

---

[19] In 2007; cf. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39534.

[20] Cf. http://www.sil.org/iso639-3/.

[21] Cf. Lewis 2009. - How dubious the calculation of languages in 'Ethnologue' is, becomes obvious immediately if we consider that it contains 21 entries (with appertaining three-letter-codes) under 'High German' (including 2 varieties of Yiddish) plus 10 entries under 'Low Saxon', but only 2 entries under 'English' (viz. 'English' and 'Scots'). As the criteria and standards applied for counting vary between different countries, regions, or investigators, the number of 6,500 languages world-wide, consistently repeated in both scientific and popular publications since the 10th edition of 'Ethnologue' (ed. by Barbara F. Grimes) was published in 1984 (with 6,519 languages counted), is nothing but a popular myth.

[22] Cf. Gippert 2008: 162–163 and 187–188 on the importance of these dialectal varieties.

[23] Cf. ib.: 175 as to an example.

[24] Given that the four-letter codes include the existing three-letter codes, the number of possible codes must be increased by the 6,900 entries of ISO 639-3.

[25] This is the World Language Documentation Centre, Wales (cf. http://www.thewldc.org/); cf. the website in http://www.geolang.com/, which makes queries about the standard available in http://www.geolang.com/iso639-6/.

as ‘Judeo-Georgian’ (JGE) are direct children of GGNC (and accordingly, siblings of the ‘Georgian’ standard language, KAT). As a matter of fact, the arrangement of varieties of Georgian in the dependency tree (cf. Figure 1) is enigmatic, and all linguists interested in providing data for cross-corpus retrieval should try to influence this on-going standardization process before its results have been accepted. This is all the more true as the standard is also meant to encompass sociolectal and historical varieties, which renders the application of one simple tree-like structure with parent-child-relations rather problematical.

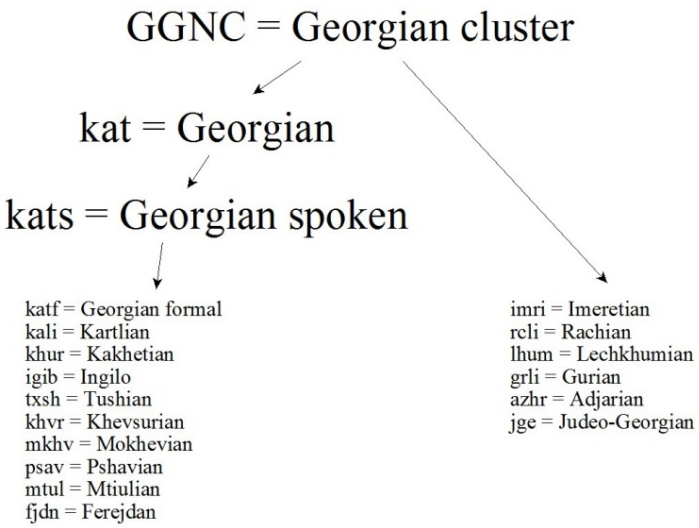

FIGURE 1: Varieties of Georgian in ISO 639–6

At the same time, we should prepare for applying fine-grained language codes in our linguistic analyses, given that they provide a means of clearly distinguishing the different layers we usually have to deal with in recordings of endangered languages. Figure 2 shows a first example of four-letter language codes applied to indicate the languages involved in the setting of Khinalug, an East-Caucasian language spoken in Azerbaijan, which has been the object of a DoBeS project since July, 2011.[26] The task of developing means to use these codes in language-specific corpus retrieval remains still to be solved. One solution might consist in assigning the language as a property of a given annotation, rather than storing the

[26] My thanks are due to Monika Rind who provided the given example from her fieldwork in Khinalug. In the language-related tiers, KJJS stands for ‘Spoken Khinalug’, AZJS for ‘Spoken Azeri’, RUSS for ‘Spoken Russian’, and ARA for ‘Arabic’. The tier \lan indicates whether a given word (form) is part of Khinalug speech (i.e., with grammatical properties such as case endings of this language) or of code switched to Azeri (with Azeri grammar), while \src indicates the immediate source of a word (form) in question (usually Azeri, as this is the main contact language of the Khinalug speakers). In addition, \etylan indicates the etymological origin of a word (e.g., Arabic) wherever applicable. Thus, e.g., the language (\lan) of *turistin* is styled as being Azeri (AZJS) because the word bears the Azeri genitive ending, *-in*, while *turizmi* is styled as being Khinalug because it bears the Khinalug genitive ending, *-i*. The source (\scr) is Azeri in both cases, while the etymological origin (\etylan) is Russian (further derivation from French etc. notwithstanding).

information in separate markup tiers.[27] This, however, would not help for annotations on the sentence level as instances of code-switching would not be coverable in this case.

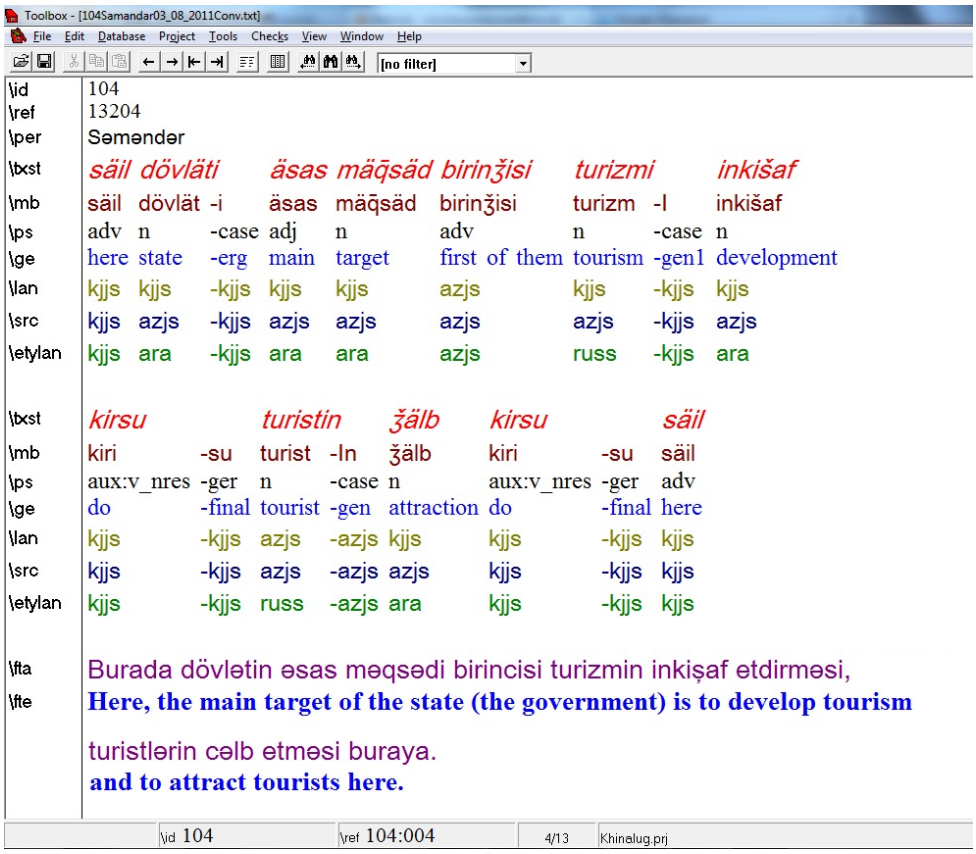

FIGURE 2: Toolbox example sentence with indication of languages

4. CONCLUSION AND OUTLOOK. The multilingual nature of many endangered languages corpora makes them especially interesting for a number of research questions. At the same time, the demarcation of material from different languages in these corpora poses severe problems for the annotation of data and for automatic retrieval mechanisms. The standardization of fine-grained language codes is one important prerequisite for coping with these problems but the ongoing endeavor towards this requires input from specialists to avoid misleading solutions. Scholars working on endangered languages are especially encouraged to bring in their expertise in this respect.

[27] It goes without saying that this would require a major addition to the functionality of the ELAN tool and the XML structure it relies upon.

REFERENCES

ECLING Corpus. ECLING Project (Endangered Caucasian Languages in Georgia). DoBeS Language Resource Archive. http://corpus1.mpi.nl/ds/imdi_browser?openpath=MPI533677%23.

Gippert, Jost. 2000–2012. TITUS Datenbank (Thesaurus Indogermanischer Text- und Sprachmaterialien). http://titus.fkidg1.uni-frankfurt.de/database/titusinx/titusinx.htm.

Gippert, Jost. 2008. Endangered Caucasian languages in Georgia: Linguistic parameters of language endangerment. In K. David Harrison, David S. Rood & Arienne Dwyer (eds.), *Lessons from Documented Endangered Languages*, 159–194. Amsterdam: Benjamins.

Grimes, Barbara F. (ed.). 1984. *Ethnologue: Languages of the World, Tenth edn*. Dallas, Tex: SIL International.

Gullberg, Marianne. this volume. Bilingual multimodality in language documentation data.

International Organization for Standardization (ISO). 2007. ISO 639–3:2007. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39534.

Lewis, M. Paul (ed.). 2009. *Ethnologue: Languages of the World, Sixteenth edn*. Dallas, Tex: SIL International. Online version: http://www.ethnologue.com/.

SIL International (Summer Institute of Linguistics). 2007. ISO 639–3. http://www.sil.org/iso639-3.

The World Language Documentation Center. Wales. http://www.thewldc.org/, http://www.geolang.com/, http://www.geolang.com/iso639-6/.

Toolbox. SIL International (Summer Institute of Linguistics). http://www.sil.org/computing/toolbox/.

Jost Gippert
gippert@em.uni-frankfurt.de