#### Languages, 'Languoids', and ISO-codes for Language Diversity and Variation

#### 5 ICLDC, Honolulu, 4<sup>th</sup> March 2017

#### **Sebastian Drude**

CLARIN ERIC / University Utrecht Radboud University Nijmegen Goethe-Universität Frankfurt



- The need of unambiguous reference to languages and similar entities
- The ISO 639 standards and their criticisms
- Other systems and their problems
- Languages and "languoids" what are they?
- Towards a topology of languages
- Multidimensional linguistic variation
- Conclusions

- The need of unambiguous reference to languages and similar entities
- The ISO 639 standards and their criticisms
- Other systems and their problems
- Languages and "languoids" what are they?
- Towards a topology of languages
- Multidimensional linguistic variation
- Conclusions

- Reference to languages traditionally by name
- Recently, worldwide language diversity in focus
- A language name can be ambiguous

Ainu (China) – Ainu (Japan) Aja (Benin) – Aja (Sudan) Ama (Papua New Guinea) – Ama (Sudan) Amba (Solomon Islands) – Amba (Uganda) Aruá (Amazonas State) – Aruá (Rodonia State) Asu (Nigeria) – Asu (Tanzania) Atong (India) – Atong (Cameroon) Awa (China) – Awa (Papua New Guinea)

- Reference to languages traditionally by name
- Recently, worldwide language diversity in focus
- A language name can be ambiguous
- Most languages have many names:
  - In different languages, contexts, spelling variants

German: Deutsch – allemand – Tedesco – niemiecki – Tysk – ...

Niederdeutsch – Plattdeutsch – Platt

Aweti: Awetí, Aueti, Auiti, Auetö, Aueto, Awytyza ti'ingku

- Reference to languages traditionally by name
- Recently, worldwide language diversity in focus
- A language name can be ambiguous
- Most languages have many names:
  - In different languages, contexts, spelling variants
  - Preferred / accepted names change over time
- Problems when searching:
  - low recall (missing relevant hits) and
  - low precision (getting many irrelevant hits)

A standard is clearly needed by / for:

- diversity linguists and linguistic infrastructures: WALS, archives, OLAC, Linguist List, etc.
- Information technology: Unicode, Microsoft, Wikipedia, Apple, Oracle, etc.
- Identifying content, user interfaces, spelling checkers, input methods, and language technology (language recognition, parser, text-to-speech technology and so forth)

- Technology not only for a few major languages
- Industry: "We don't want to be a bottleneck for language communities!" – not limited to the "economically significant" ones
- Where is the line anyways? A moving target...
- Unicode Consortium mission: "This Corporation's specific purpose shall be to enable people around the world to use computers in any language..."

- Soon, oral human-computer-interaction may replace typing and pointing to a large extent
- Technology for coping with oral (and written) variation will be developed
- Therefore a need for identifying and labelling language varieties is arising
- There are objections against standardizing at all, "the situation is just too messy and divers"
- Still, technology will move ahead

- Any standard is necessarily a compromise: compare the time zones, a pragmatic arbitrary cut through a continuum (<u>Simons 2009</u>)
- Our situation is quite comparable to biology
- Clades currently are the best theoretical approach, "species" is a debated notion
- But the "species" concept is a good basis for the Linnaean system for labelling living beings
- The "clades-labeling-system" does not fly

- The need of unambiguous reference to languages and similar entities
- The ISO 639 standards and their criticisms
- Other systems and their problems
- Languages and "languoids" what are they?
- Towards a topology of languages
- Multidimensional linguistic variation
- Conclusions

- ISO/TC37/SC2/WG1: language coding
- 1967: terminologists release ISO 639 (-1), two-letter codes, now 200 entries
- 1998: ISO 639-2, also by librarians, now ~505 three-letter codes for ~410 "major" languages and ~70 collective codes for language groups
- From 2000 on: pressure on ISO to cover all languages (WWW, Unicode, OLAC, diversity linguistics: WALS, language documentation)

- Ethnologue was then identified as best and most comprehensive listing of languages
- SIL agreed to develop and maintain ISO 639-3
- SIL adjusted their three-letter-codes to existing codes in part 2 etc. (~600 changes)
- Part 3 first published in 2007, confirmed 2010
- Yearly updated, with an explicit procedure
- Now 7864 code elements, Ethnologue in sync
- Living and recently extinct individual languages

- Part 4: "General principles of coding..." (drafts since 2008, yet to be finalized and confirmed)
- Part 5: three-letter-codes for language groups (70 fr. part 2 and 50 more, maintained by LoC)
- Since 2008: attempts at a part 6, four-lettercodes for "comprehensive coverage of language variants" by GeoLang Ltd., UK
- Rejected by ISO/TC37 in 2014, there is no pt. 6
- Framework for linguistic variation needed

Uptake:

- Parts 1 and 2 are arguably the most often used ISO standards of all (device's user interfaces)
- Part 3 is now largely replacing part 2
- Important: <u>IETF BCP 47</u> is a key industry technology using ISO 639-3 (<u>talk Constable</u>)
- Part 4 is needed to clarify criteria & conception
- Part 5 is apparently hardly used at all

Problems and criticisms of ISO 639(-3):

- Being "authoritative" (funders & archives require it..., government's decisions,...):
   Partly a straw man; – in any case not really ISO's fault, any such standard can and would be misused
- Connection with Ethnologue; missionary organization as registration authority Is problematic, but are there alternatives? Industry needs more stability than a website; revision process is expensive; a long-term commitment is needed

Problems and criticisms of ISO 639(-3):

 The codes look like abbreviations and sometimes are mnemonic of inappropriate labels True, but no good solution seems feasible. 65% of the 17,576 possible combinations are taken. Mnemonic match is now often impossible anyways. ISO will not get into the merits of appropriate labels – who is authorized to complain, and who to decide? Complete replacement impossible, stability needed "Best thought of as three-digit base 26 numbers".

Problems and criticisms of ISO 639(-3):

- Genealogical classification is questionable Not part of ISO standard, Ethnologue is not ISO
- Boundaries between languages / dialect chains
- Language vs. dialect / structural vs. functional General problems, pragmatic solutions are possible
- Change process: involvement of experts is too low; lack of transparency wrt. people involved Possible cure: scientific advisory board from HERE

- The need of unambiguous reference to languages and similar entities
- The ISO 639 standards and their criticisms
- Other systems and their problems
- Languages and "languoids" what are they?
- Towards a topology of languages
- Multidimensional linguistic variation
- Conclusions

#### 3) Other systems and their problems

- UNESCO atlas of languages in danger: only EL
- <u>Multitree</u>: everything side by side; no standardized names; only families, languages & dialects
- <u>Endangered Languages Catalogue</u> (ELcat): EL only, crowd sourcing – review process?
- The <u>Linguasphere Register</u>: Poor PDF-files; last edition: 2000, no sources; idiosyncratic; one flat hierarchy, mixed socio-political, linguistic and geographic criteria. E.g. std. German: "52-ACB-dI"
- All of these face the sustainability problem!

#### 3) Other systems and their problems

• <u>Glottolog</u>: certainly the most promising alternative

- + For languages: best knowledge synthesis around
- + Sources are made explicit
- + Usable unique codes, links to other resources
  Admittedly not reliable for dialects and variants (they are taken from Multitree, no systematic revision)
  Funding? Review process? Sustainability?
  Also only one flat hierarchy, mostly on dialects (not other dimensions whistled 'languages' separate)
  Authoritative for genealogic grouping

- The need of unambiguous reference to languages and similar entities
- The ISO 639 standards and their criticisms
- Other systems and their problems
- Languages and "languoids" what are they?
- Towards a topology of languages
- Multidimensional linguistic variation
- Conclusions

- Glottolog uses "languoids", usually understood as a cover term for languages, language families and dialects, useful for unclear cases
- Sometimes skepticism on feasibility of good definitions for language, lang. family, dialect
- Good/Cysow attempt theoretical underpinning This is not usable, already for ontological reasons Whatever languages are, they are not entities that contain their own names as one of their components

- Linguistics can and should define these terms
- At least a pragmatic framework for a standard
- Sure, one has to recognize different criteria for languages – (1) linguistic (mutual intelligibility) and (2) socio-politico-cultural (group identity)
- The linguistic definition of language is more fundamental ("I-languages" are basic)
- Also "s-languages" (on s/p/c-grounds) may merit a three letter code in ISO 639-3

- In ISO 639-3 there are "macrolanguage"-codes (cases, e.g.: Arabic, Chinese, Norwegian)
- Pragmatic solutions, but do not fully reflect real the social & linguistic situations
- We need a conceptual framework, starting with answering: what is a language?
- Languages are not systems or similar, they are SETS of individual 'means of communication' ("idiolects", one speaker uses several)

- A feasible framework starts from a definition of <u>mutually intelligibillity</u> (m.i.) between idiolects
- Still, some details need clarification: "understand" (is probably gradual and thus needs to be quantified or tested by a standard test) "without learning" etc. – difficult to test: often passive knowledge of other varieties is pervasive "trans-medial correspondence conventions" are needed for written or whistled, drummed etc. forms

It is useful and possible to define:

- <u>Chain of m.i.</u> between two idiolects
   A sequence with m.i. between adjacent members
- Linguistically defined language at a point in time (<u>*I-s-language*</u>)

Largest set of m.i.-chained idiolects at a point in time

 Linguistically defined <u>language</u> through time Largest set of m.i.-chained idiolects so that no two different l-s-languages are subsets

It is useful and possible to define:

 <u>Variety</u>: a largest subset of a language delineated by both external and structural criteria

External: e.g.: apart medium (e.g. writing); use in a certain type of situation (time, formality etc.); speakers share certain distinctive properties (social, geographical group)

Possibly, the definition needs to include provision for fuzziness, using a prototypical small subset Even without fuzziness, varieties may overlap

It is useful and possible to define:

- L1 <u>descends from</u> L2: m.i.-chain through time
- Language family: largest set of I-s-languages that all descend from an ancestor I-s-language
- A *languoid* at a time t is either
  - a. an I-s-language at t, or
  - b. a variety of a L-s-language at t, or
  - c. a language family at t

Languoids are ontologically heterogeneous

What is the meaning of names of languages?

- English refers to the language named "English"
- ... and that is spoken by the majority of the population of the UK, the USA, Australia, ...
- So at least one name, (location of) speakers are the "defining" (better: identifying) criteria
- Additional properties: number of speakers, other names, belonging to a language family, structural characteristics, historical origin ...

- The need of unambiguous reference to languages and similar entities
- The ISO 639 standards and their criticisms
- Other systems and their problems
- Languages and "languoids" what are they?
- Towards a topology of languages
- Multidimensional linguistic variation
- Conclusions

#### 5) Towards a topology of languages

- We need ore sophisticated terminologies to account for the topology of languages
- For example T. Kaufmann's (1990) proposals:
   *Families languages dialects*: paradigmatic and most common case
   Some languages are *dialect chains* (serial intell.)
   *Language areas and emergent languages*: clear boundaries but high intelligibility
   *Language complexes with virtual languages*: dialect chains with subsets functioning as languages

- The need of unambiguous reference to languages and similar entities
- The ISO 639 standards and their criticisms
- Other systems and their problems
- Languages and "languoids" what are they?
- Towards a topology of languages
- Multidimensional linguistic variation
- Conclusions

### 6) Multidimensional linguistic variation

Dimensions of linguistic variation:

- Space (dialects, over-regional standard varieties)
- Time (epochs, periods, stages)
- Social groups (sociolects of several different types)
- Medium (oral, written, signed, whistled, drummed...)
- Situation (registers of different formality)
- Individual ("personal varieties"~ traditional "idiolects")
- (Possibly) proficiency (for learners varieties of different stages, motherese and similar)

- The need of unambiguous reference to languages and similar entities
- The ISO 639 standards and their criticisms
- Other systems and their problems
- Languages and "languoids" what are they?
- Towards a topology of languages
- Multidimensional linguistic variation
- Conclusions

### 7) Conclusions

- A pragmatic labelling system is essential
- Currently no way around ISO 639 for languages
- Glottolog could complement / supersede it
- Experts panel for sound review process is needed
- The topology of languages is more complex than "family" – "language" – "dialect"
- A sound pragmatic conceptual framework for "languages" and other types of "langoids" is possible
- Language internal variation is multidimensional

### Languages, 'Languoids', and ISO-codes for Language Diversity and Variation

### 5 ICLDC, Honolulu, 4<sup>th</sup> March 2017

### **Sebastian Drude**

CLARIN ERIC / University Utrecht Radboud University Nijmegen Goethe-Universität Frankfurt







# 2) The ISO 639 standards and their criticisms

View by:

○ 639-3 ○	639-2/639-5	0 639-1	Name	O Scope	🔾 Туре	Starting with A	~	Submit
-----------	-------------	---------	------	---------	--------	-----------------	---	--------

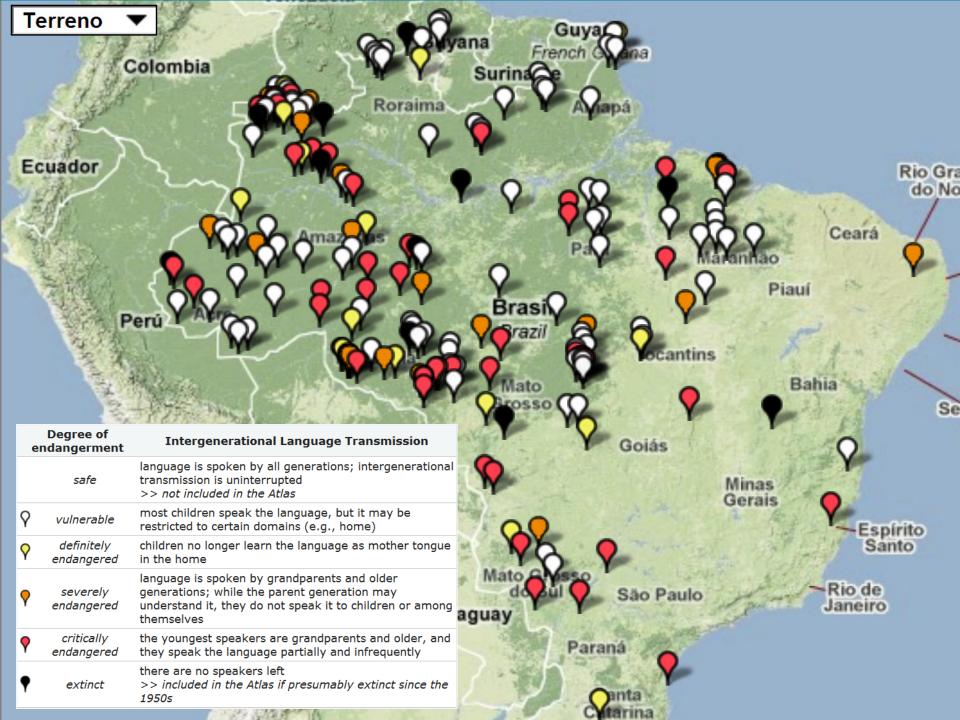
See also: Library of Congress <u>ISO 639-2 Registration Authority</u> (which is the primary authority for Parts 1 and 2 of the standard) and <u>ISO 639-5 Registration Authority</u>. These sites are the definitive sources in the case of a discrepancy with this site.

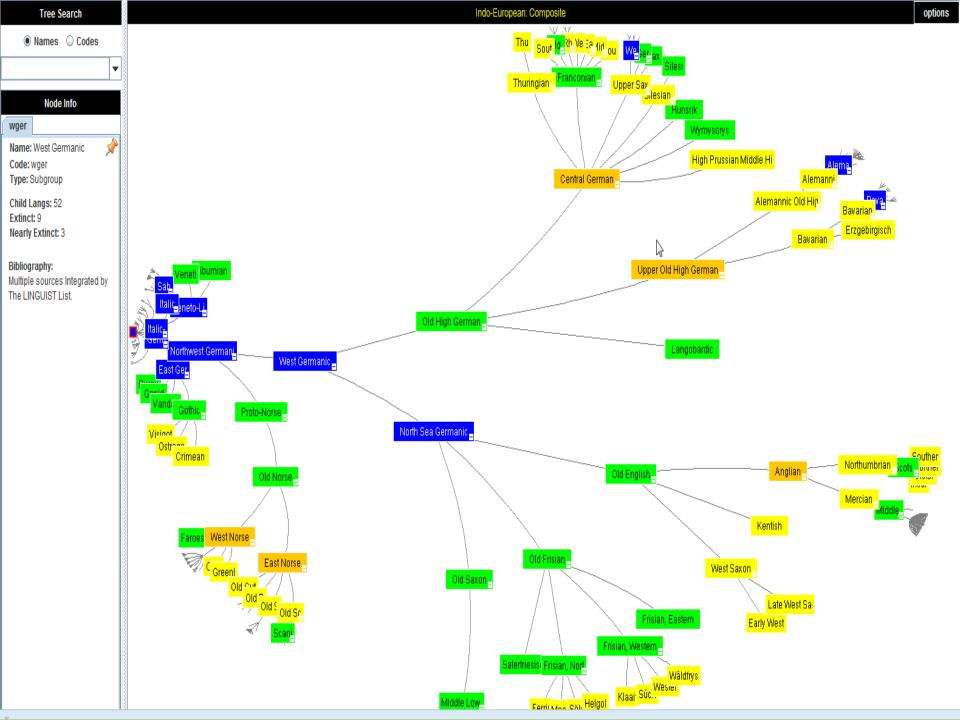
See also Retired Code Elements Index

639-3	639-2/ 639-5	639-1	Language Name	Scope	Туре	
dgc			Casiguran Dumagat Agta	Individual	Living	more
spa	spa	es	Castilian	Individual	Living	more
asc			Casuarina Coast Asmat	Individual	Living	more
cat	cat	са	Catalan	Individual	Living	more
CSC			Catalan Sign Language	Individual	Living	more
chc			Catawba	Individual	Extinct	more
cca			Cauca	Individual	Extinct	more
	cau		Caucasian languages	<u>Collective</u>		more
cav			Cavineña	Individual	Living	more
cyb			Cayubaba	Individual	Extinct	more
cay			Cayuga	Individual	Living	more
хсу			Cayuse	Individual	Extinct	more

# 2) The ISO 639 standards and their criticisms

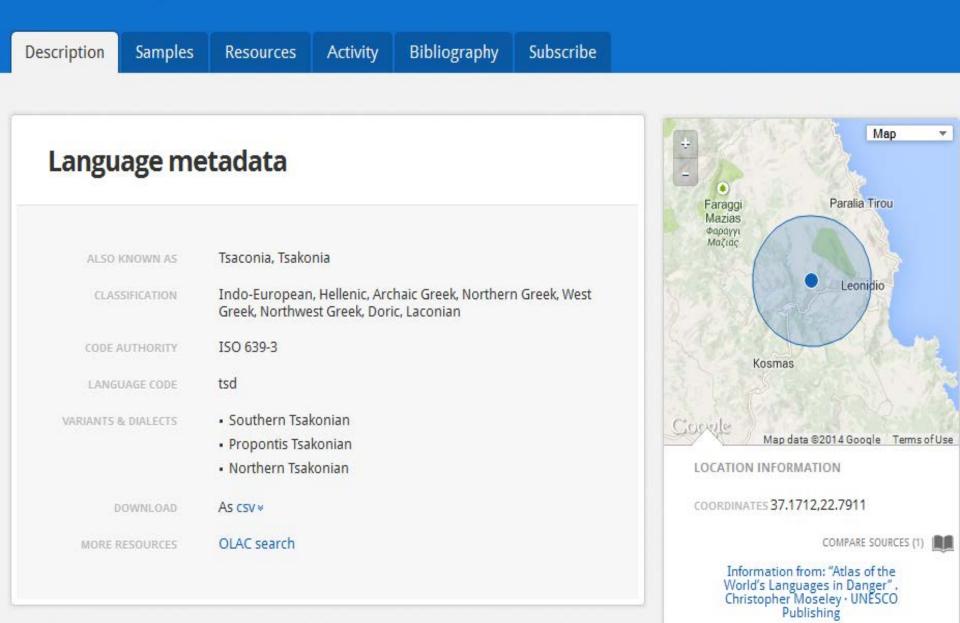
ISO 639-3 Home	Chang	e request d	ocumentation for: 2014-042
Contact	Change	2014-042.pdf	
Code Tables	request:		
Language codes	Requested change:	1. Affected language identifier:	aam (Aramanik)
Macrolanguage		Change type:	Merge
mappings		Retirement remedy:	Merged into Aasax [aas]
Download		Old value:	Aramanik
Code Channes		Outcome:	Adopted
Code Changes Change management		2. Affected language identifier:	aas (Aasáx)
Submitting change		Change type:	Update
requests		Change attribute:	The proposed change is a change (expansion) in Denotation, from the merging of another code element into this code element
Change request index		Old value:	Aasax
		Outcome:	Adopted
Related Pages			
Relationship to other parts of ISO 639		Que	estions/Comments: iso639-3@sil.org





### Tsakonian [aka Tsaconia, Tsakonia,]

Classification: Indo-European • threatened



page **432** 

1

×

### The Linguasphere Register 🚇 1999/2000 edition

L					
	52-ACB-dkk	meissnisch	<i>including</i> nord-meissnisch west-meissnisch süd-meissnisch ost- meissnisch, meissen+ dresden ≽a key area in the historical development of formal and generalised German (see Hochdeutsch-F. and Hochdeutsch-G. below). ; as settlers speaking northern, central and southern idioms of German (see Deutsch-N., Deutsch-C., Deutsch- S.) converged in Saxony during the eastward colonisation of Slav- speaking areas, the Elbe valley around Meissen and Dresden became a focal-point for the development of a compromise written and spoken form of the German language in the late Middle Ages. ⊕Mulde valley (Grimma Lunzenau) Elbe valley (Meissen Dresden Pirna)	(Sachsen)	
	52-ACB-dkl	böhmisch-N.	"bohemian german"-N. ⊕Erzgebirge-E. <i>mountains</i> : Chomutov, Komotau… Dĕčín, Tetschen ¶ <i>most speakers dispersed to Germany in</i> or after 1945	Czech Rep. (Severočeský)	3*
	52-ACB-dl	hochdeutsch-F.	<b>high-german</b> , "formal german", "standard german", gepflegtes hochdeutsch (=« <i>cultivated high german</i> »), deutsche schriftsprache (=« <i>german written language</i> »), deutsche hochsprache (=« <i>german high language</i> »), deutsche literatursprache (=« <i>german literary language</i> »), bühnendeutsch (=« <i>stage german</i> »), deutsche standardsprache (=« <i>german standard language</i> »); einheits-sprache (= « <i>unity language</i> »), part of neu-hochdeutsch, "new high german" ; the development of a compromise speech-form (Ausgleichssprache) based on converging forms of German in Thüringen and Sachsen culminated in the 14 <sup>th</sup> cent. written language of the imperial chancellery in Prague, which was in turn consolidated in the written language of the Meissen, Dresden and other Saxon and Thüringian chancelleries, and of the universities of Leipzig and Wittenberg. The determining event for the establishment of modern standard German was the publication of Luther's translation of the Bible in the early 16 <sup>th</sup> cent. To the "national" idioms listed below should be added the written idioms of Hochdeutsch (often archaic or influence < other German or non-German idioms) which correspond to many of idioms of Auswanderungsdeutsch (Émigré German). Idioms are listed below for the countries where German is currently an official language	A Germany; A Austria; formerly also (A) Austro-Hungarian Empire; Italy (A Südtirol; Alto-Adige); A Switzerland; A Liechtenstein; A Luxemburg; Belgium (A Liège-E.); formerly also France (A)Alsace; (A ) Lorraine-N.)	7
	52-ACB-dla	bundesdeutsch+ reichsdeutsch-F.	"formal german of Germany", <i>including present-day</i> bundesdeutsch, "federal german", "formal german of German Federal Republic" and recent d.d.rdeutsch, "g.d.r. german"; "formal german of East Cormony" and preceded biotorically by p.c. doutsch, "patienal socialist	Germany	

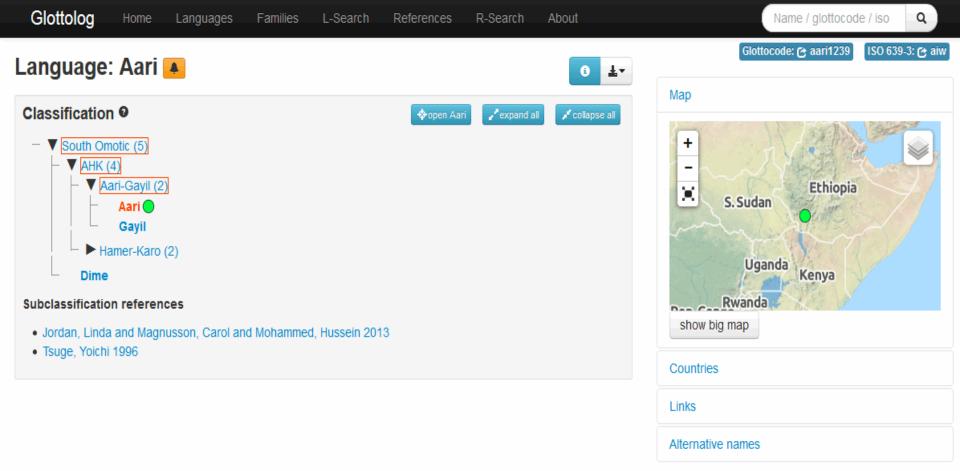
### Languages

Showing 1 to 100 of 7,816 entries

 $\leftarrow \text{Previous} \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad \text{Next} \rightarrow \qquad 3$ 

±×.

Glottocode Top-level family 👙 ISO-639-3 Child dialects Latitude Longitude Name Macro-area ÷ Ŧ Search Search Search Search Search Search Search --any-aari1239 Aari South Omotic Africa 5.95 36.57 aiw 0 aasa1238 36.86 Aasax Afro-Asiatic Africa -4.01 aas 0 -9.03 146.99 abad1241 Abadi Austronesian kbt Papunesia 0 Nuclear Trans New -6.12 145.66 abag1245 Abaga abg Papunesia 0 Guinea abai1240 Abai Sungai abf Papunesia 0 5.55 118.31 Austronesian abai1241 Abai Tubu-Abai Austronesian Papunesia 0 Sembuak aban1242 6.31 8.63 Abanyom Atlantic-Congo abm Africa 0 abar1238 Abar Atlantic-Congo mij Africa 0 6.58 10.23 141.32 abau1245 Abau Sepik Papunesia -3.97 0 aau 42.73 abaz1241 Abaza Abkhaz-Adyge Eurasia 3 41.12 abq abee1242 Abe Atlantic-Congo aba Africa 4 5.60 -4.38 120.20 aben1249 Abenlen Ayta Austronesian Papunesia 15.41 abp 0 abid1235 Abidji Atlantic-Congo abi Africa 2 5.66 -4.58 abin1243 Abinomn bsa Papunesia 0 -2.92 138.89 abip1241 South 0 -29.00-61.00 Abipon Guaicuruan axb America



#### References

Showing 1	to 30 of 30 entries	← Previ	Previous 1 Next → 3 ▲▼						
Details	Name 🍦	Title 🛓	ca	Year 🕴	Pages 🔶	Doctype	ca	Provider	da
	Search	Search		Search	Search	any 🗸		any 🗸	
more	Naty, Alexander 1992	The culture of powerlessness and the spirit of rebellion among the Aari people of southwest Ethiopia		1992	700	ethnographic		eball, hh	•
more	Tucker, Archibald N. and	Linguistic Analyses: The Non-Bantu		1966	634	minimal, overview		eball, guldemann,	•

- A feasible framework starts from a definition of <u>mutually intelligibillity</u> (m.i.) between 2 idiolects
- Two idiolects I1 used by Speaker Sp1 and I2 used by Sp2 are M.I. iff the two speakers Sp1 using I1 and Sp2 using I2 are both able to understand one another only on the basis of their own respective knowledge of their own idiolect (Sp1 of I1 and Sp2 of I2) and possibly some trans-medial correspondence conventions.

### <u>Chain of m.i.</u> between two idiolects

A set of idiolects is a chain of M.I. between I<sub>a</sub> and I<sub>z</sub> iff it can be exhaustively ordered into a sequence so that between all two adjacent members I<sub>i</sub> and I<sub>j</sub> of the chain exists M.I. and I<sub>a</sub> is the first member and I<sub>z</sub> the last member of that sequence.

- Linguistically defined language at a point in time (<u>*I-s-language*</u>)
- A non-empty set of idiolects L is an L-s-Language at a point in time t iff it is a largest set so that all elements of L are used at t and between any two I<sub>i</sub> and I<sub>j</sub> elements of L, there exists a chain of M.I. of elements of L

Linguistically defined <u>language</u> through time

 A non-empty set of idiolects L is an L-t-Language iff it is a largest set so that between any two li and lj elements of L, there exists a chain of idiolects, and all idiolects are used between two points in time t1 and t2, and there neither at t1 nor at t2 nor at any point in time between there are two different L-s-Languages that are both a subset of L.

### • <u>Variety</u>

 A non-empty set of idiolects L1 is a variety of a L-t-Language L2 iff L1 is a subset of L2 that can be set apart from other elements of the language by both (a) and (b):

a) there is a set of shared external properties of the elements of L1 that sets them apart from other elements of L2, and

b) the idiolects share a significant amount of distinctive structural properties