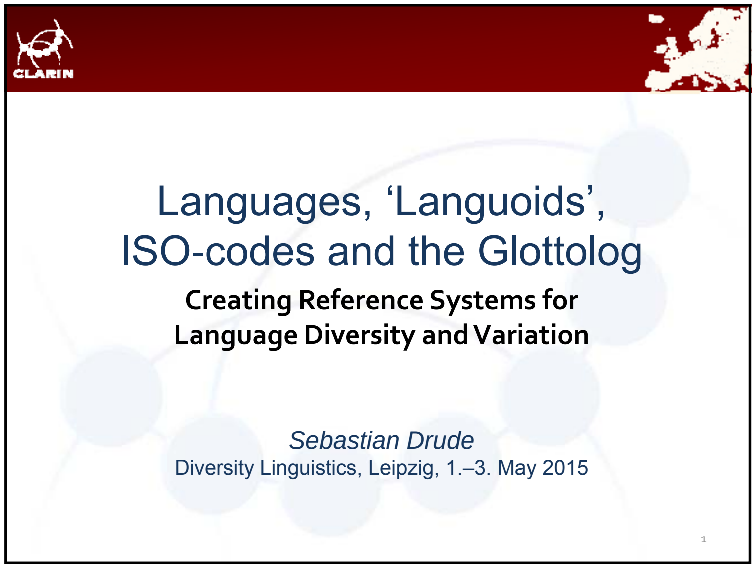
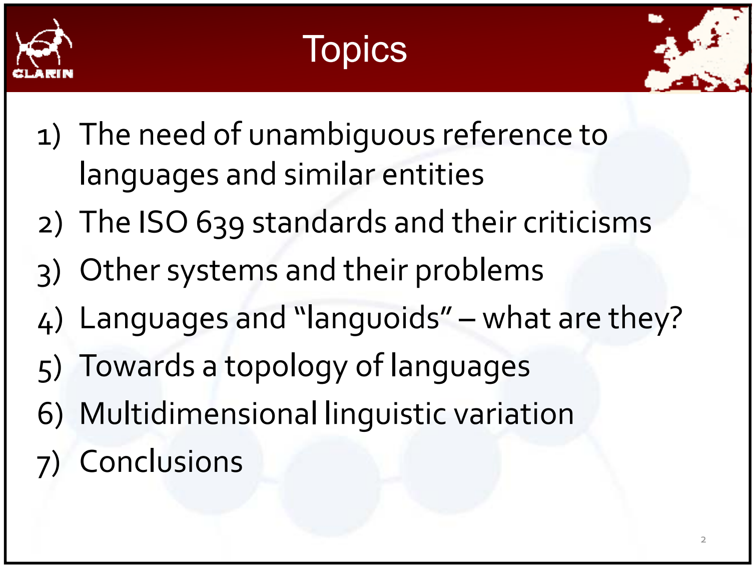


Languages, 'Languoids', ISO-codes and the Glottolog

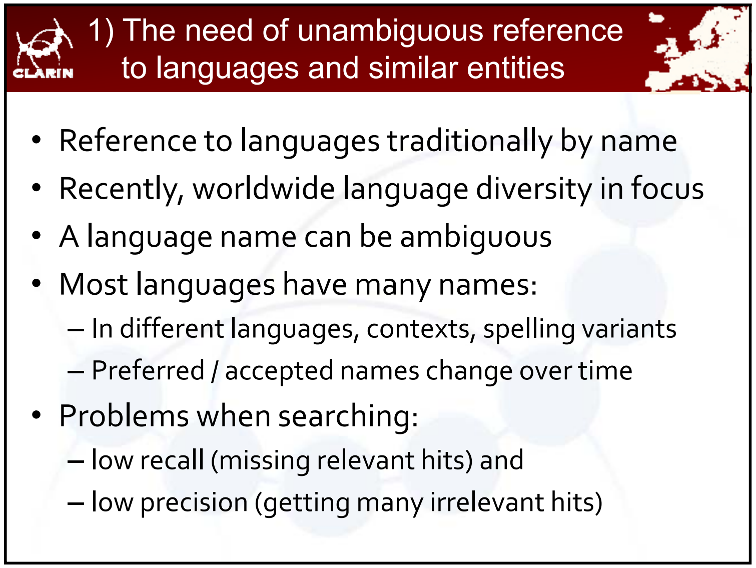
Creating Reference Systems for Language Diversity and Variation

Sebastian Drude
Diversity Linguistics, Leipzig, 1.–3. May 2015



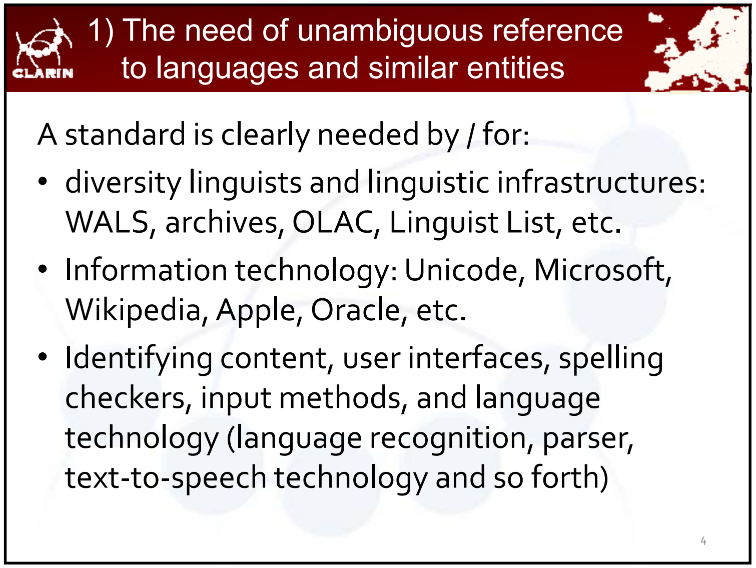
Topics

1) The need of unambiguous reference to languages and similar entities
2) The ISO 639 standards and their criticisms
3) Other systems and their problems
4) Languages and "languoids" – what are they?
5) Towards a topology of languages
6) Multidimensional linguistic variation
7) Conclusions



1) The need of unambiguous reference to languages and similar entities

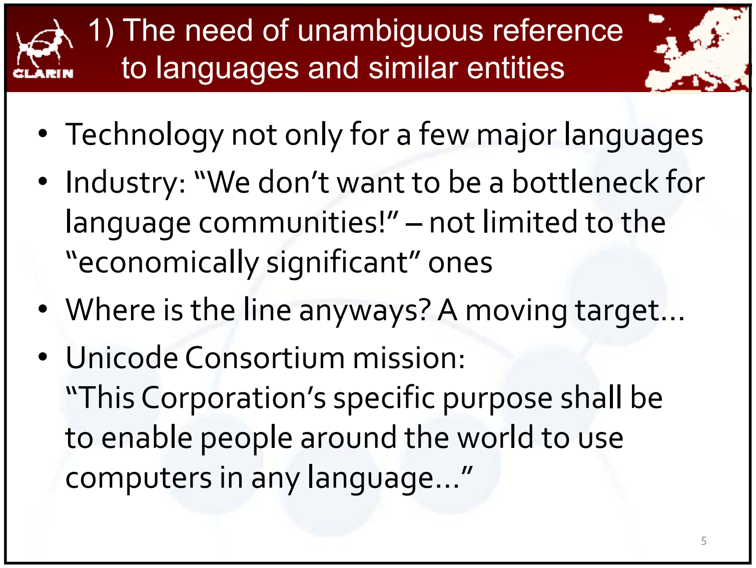
- Reference to languages traditionally by name
- Recently, worldwide language diversity in focus
- A language name can be ambiguous
- Most languages have many names:
  - In different languages, contexts, spelling variants
  - Preferred / accepted names change over time
- Problems when searching:
  -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



1) The need of unambiguous reference to languages and similar entities

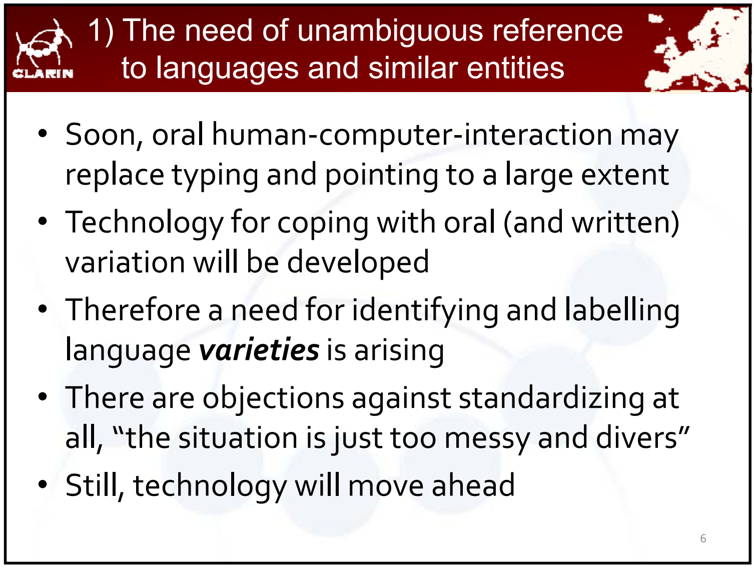
A standard is clearly needed by / for:

- 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)



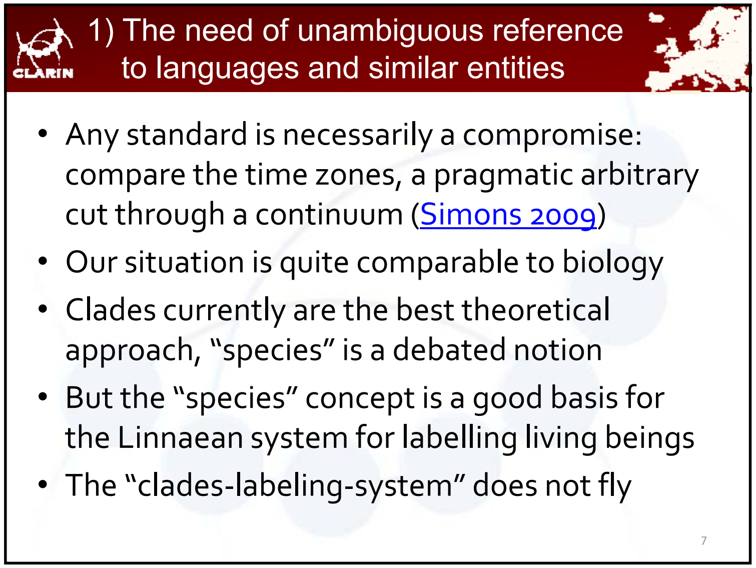
1) The need of unambiguous reference to languages and similar entities

- 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…"



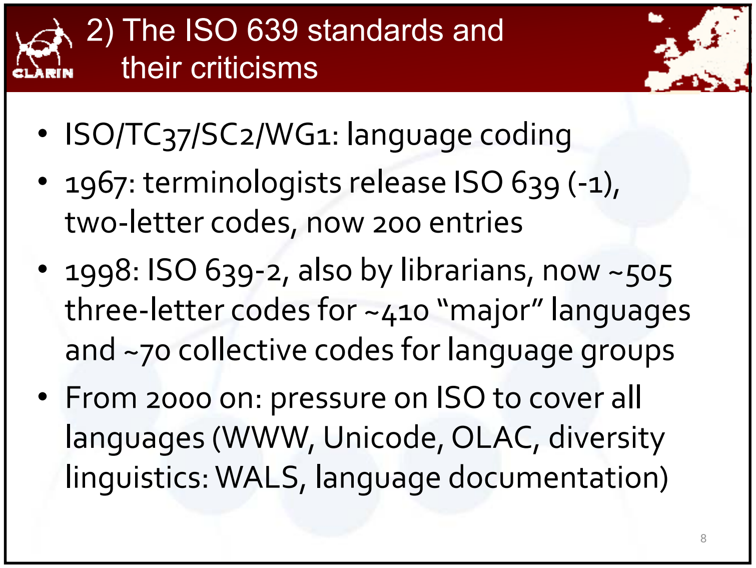
1) The need of unambiguous reference to languages and similar entities

- 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



1) The need of unambiguous reference to languages and similar entities

- Any standard is necessarily a compromise: compare the time zones, a pragmatic arbitrary cut through a continuum (Simons 2009)
- 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



2) The ISO 639 standards and their criticisms

- 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)

2) The ISO 639 standards and their criticisms

- 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

2) The ISO 639 standards and their criticisms

- 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-letter-codes 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

2) The ISO 639 standards and their criticisms

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: IETF BCP 47 is a key industry technology using ISO 639-3 (talk Constable)
- Part 4 is needed to clarify criteria & conception
- Part 5 is apparently hardly used at all

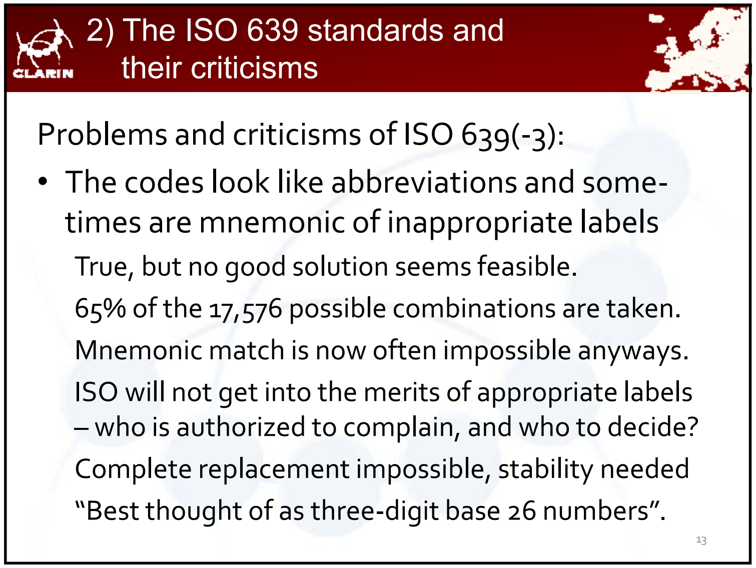
2) The ISO 639 standards and their criticisms

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



2) The ISO 639 standards and their criticisms

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.

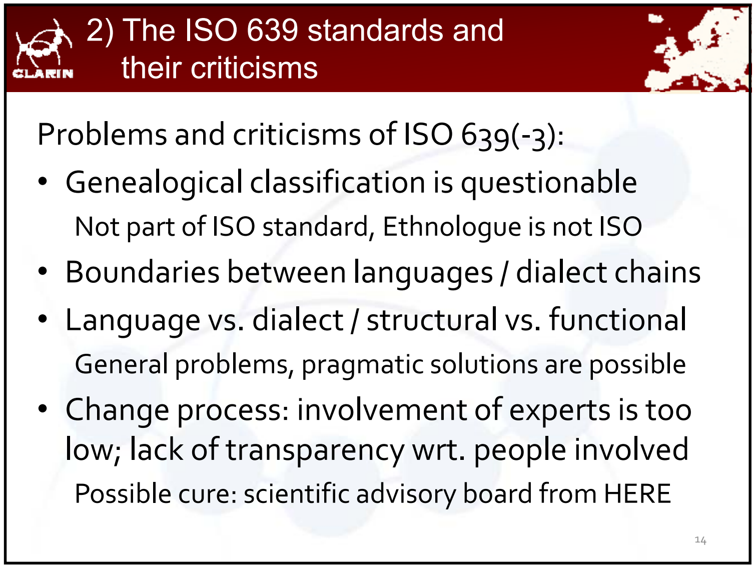
  Mnemonic match is now often impossible anyways.

  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”.

13



2) The ISO 639 standards and their criticisms

Problems and criticisms of ISO 639(-3):

- Genealogical classification is questionable

  Not part of ISO standard, Ethnologue is not ISO

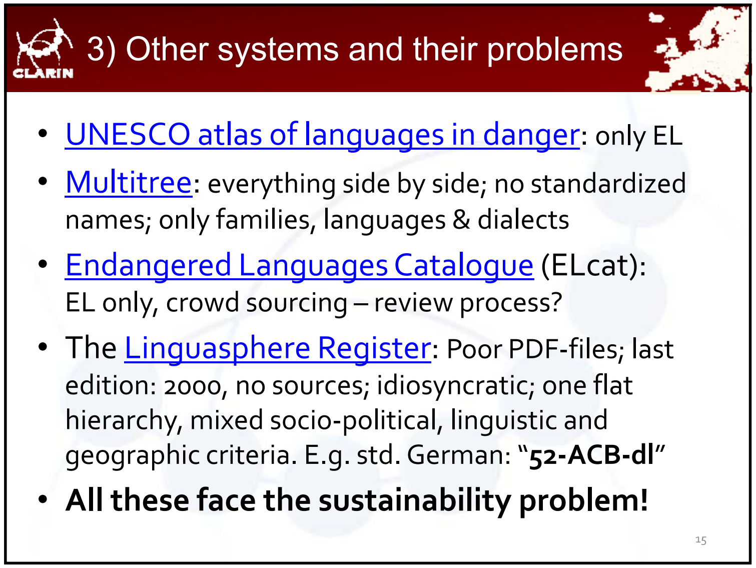
- Boundaries between languages / dialect chains
- Language vs. dialect / structural vs. functional

  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

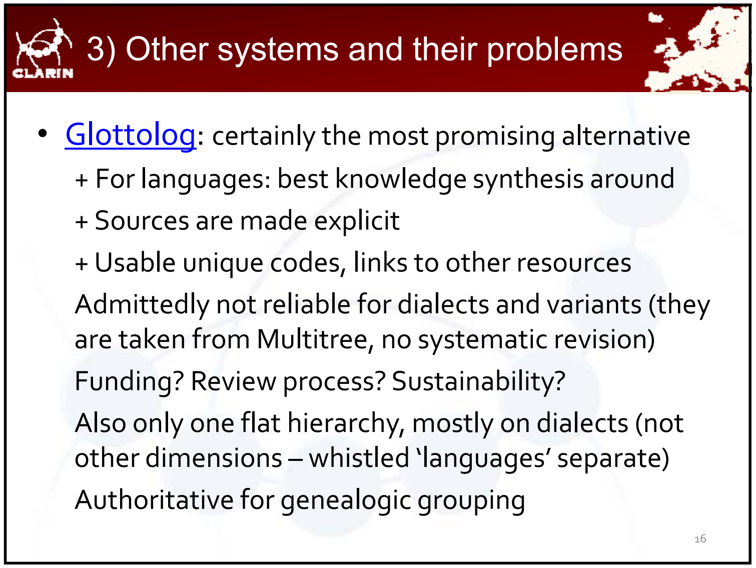
14



3) Other systems and their problems

- UNESCO atlas of languages in danger: only EL
- Multitree: everything side by side; no standardized names; only families, languages & dialects
- Endangered Languages Catalogue (ELcat): EL only, crowd sourcing – review process?
- The Linguasphere Register: 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-dl**”
- **All these face the sustainability problem!**

15



3) Other systems and their problems

- Glottolog: certainly the most promising alternative

  + For languages: best knowledge synthesis around

  + Sources are made explicit

  + Usable unique codes, links to other resources

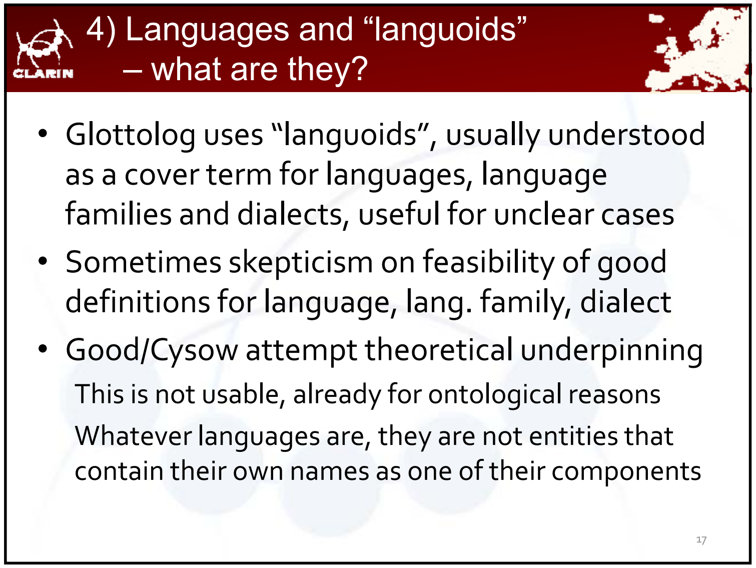
  Admittedly not reliable for dialects and variants (they are taken from Multitree, no systematic revision)

  Funding? Review process? Sustainability?

  Also only one flat hierarchy, mostly on dialects (not other dimensions – whistled ‘languages’ separate)

  Authoritative for genealogic grouping

16

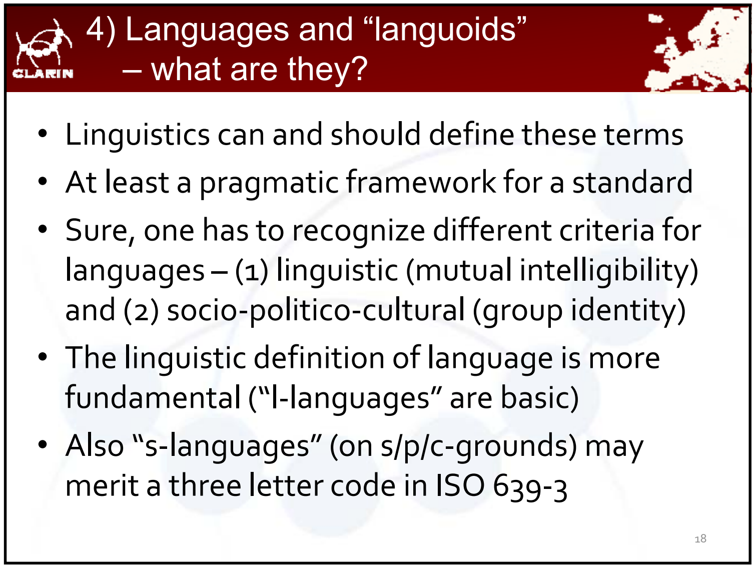


4) Languages and "languoids" – what are they?

- 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

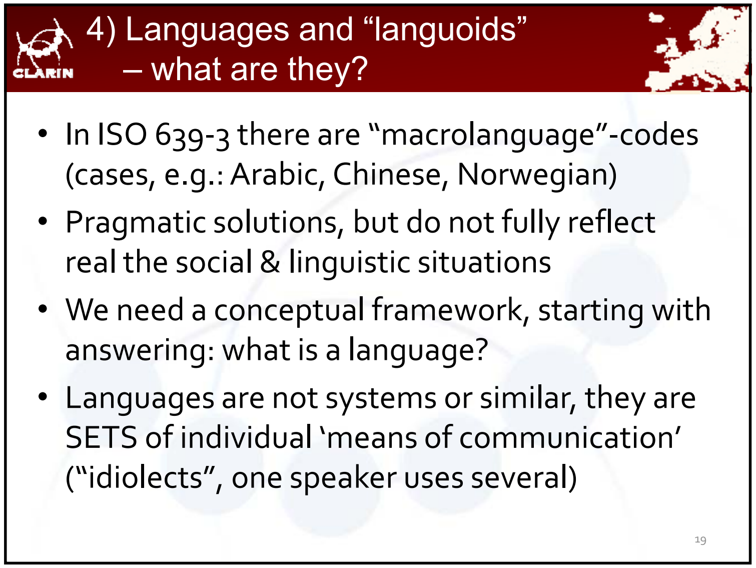
  This is not usable, already for ontological reasons

  Whatever languages are, they are not entities that contain their own names as one of their components



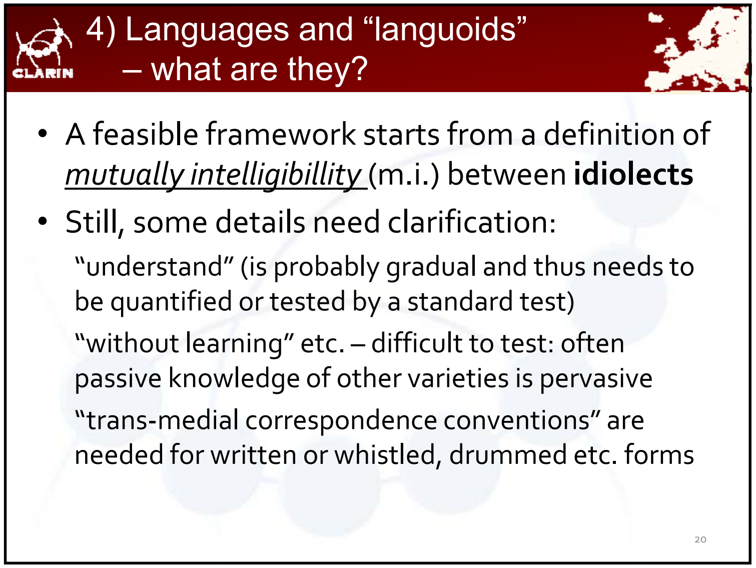
4) Languages and "languoids" – what are they?

- 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 ("l-languages" are basic)
- Also "s-languages" (on s/p/c-grounds) may merit a three letter code in ISO 639-3



4) Languages and "languoids" – what are they?

- 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)



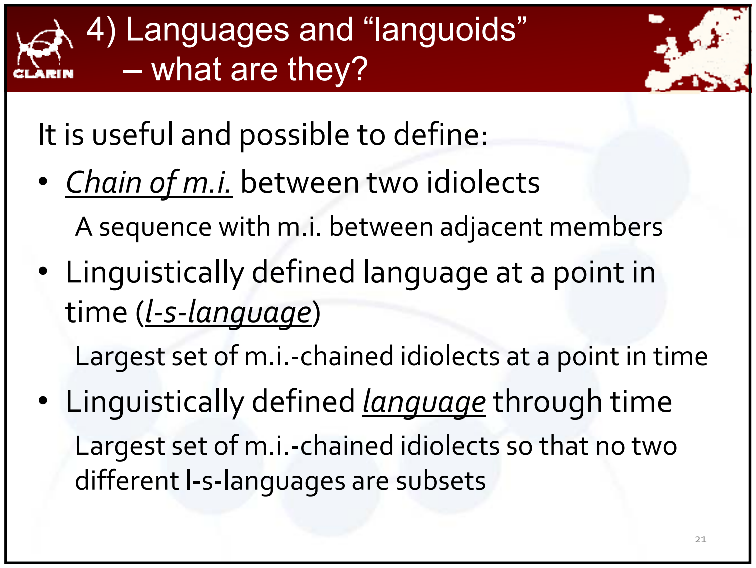
4) Languages and "languoids" – what are they?

- A feasible framework starts from a definition of *mutually intelligibillity* (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



4) Languages and “languoids” – what are they?

It is useful and possible to define:

- *Chain of m.i.* between two idiolects

  A sequence with m.i. between adjacent members

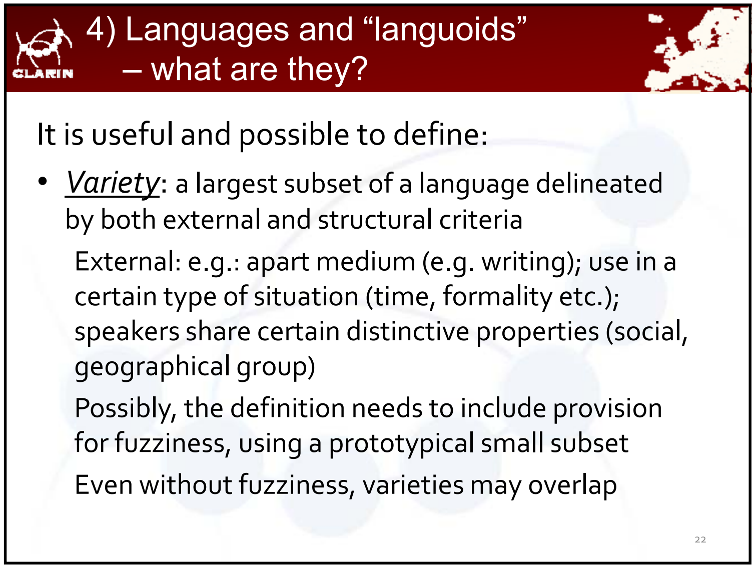
- Linguistically defined language at a point in time (*l-s-language*)

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

- Linguistically defined *language* through time

  Largest set of m.i.-chained idiolects so that no two different l-s-languages are subsets

21



4) Languages and “languoids” – what are they?

It is useful and possible to define:

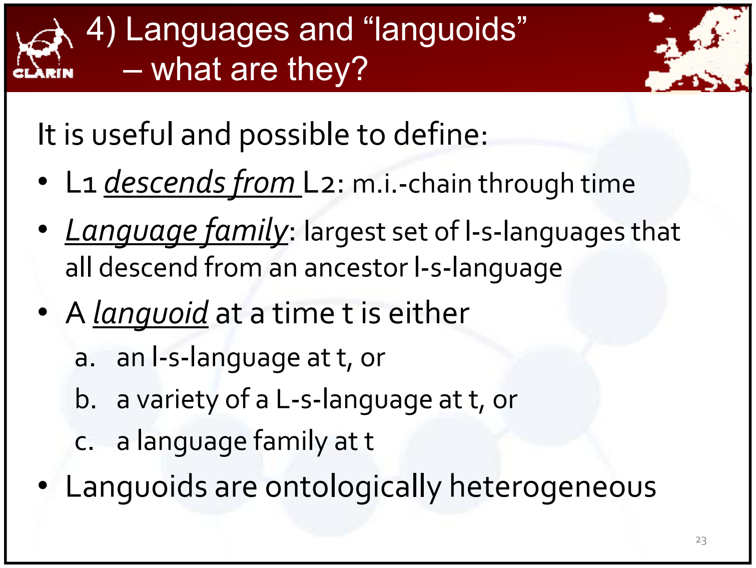
- *Variety*: a largest subset of a language delineated by both external and structural criteria

  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

22

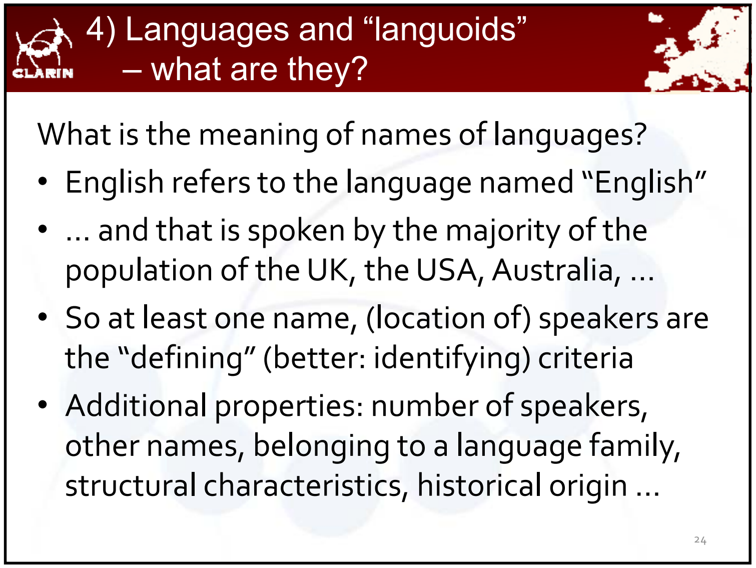


4) Languages and “languoids” – what are they?

It is useful and possible to define:

- L1 *descends from* L2: m.i.-chain through time
- *Language family*: largest set of l-s-languages that all descend from an ancestor l-s-language
- A *languoid* at a time t is either
  a. an l-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

23

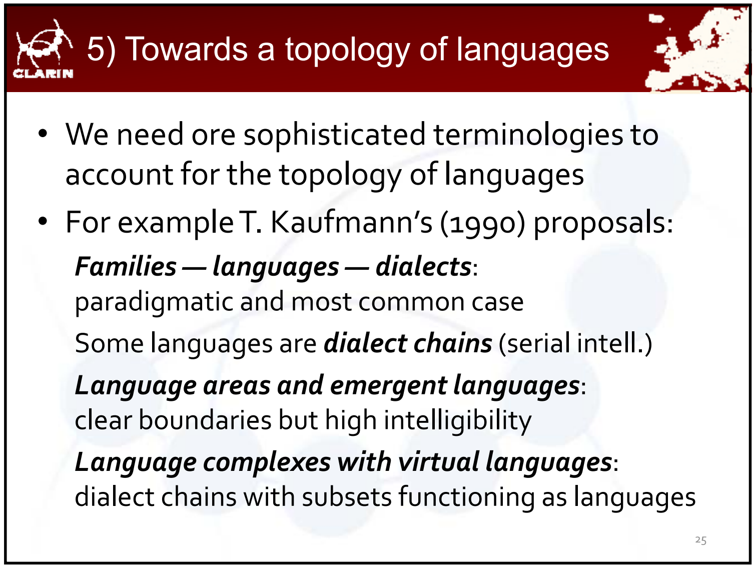


4) Languages and “languoids” – what are they?

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 …

24



5) Towards a topology of languages

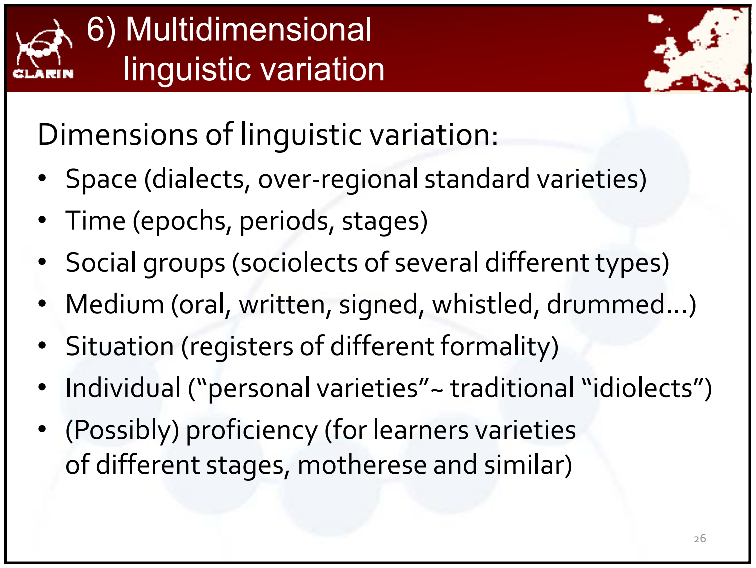
- We need ore sophisticated terminologies to account for the topology of languages
- For example T. Kaufmann’s (1990) proposals:

  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

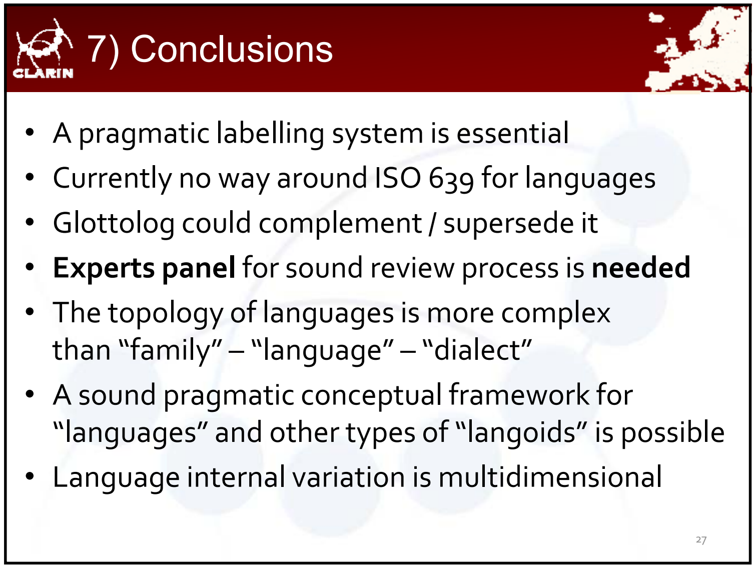
  Language complexes with virtual languages: dialect chains with subsets functioning as languages



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)



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