

Can GOLD “cope” with WALIS?

An examination of some of the problems for ontologies posed by typological databases

Jeff Good

good@eva.mpg.de

Max Planck Institute for Evolutionary Anthropology

University of Washington, April 4, 2005

Outline of talk

- Discuss the current conception of the **GOLD** community
- Discuss the structure of the **WALS** “database”
- Summarize some of the ontological challenges posed by WALS

The GOLD Ontology

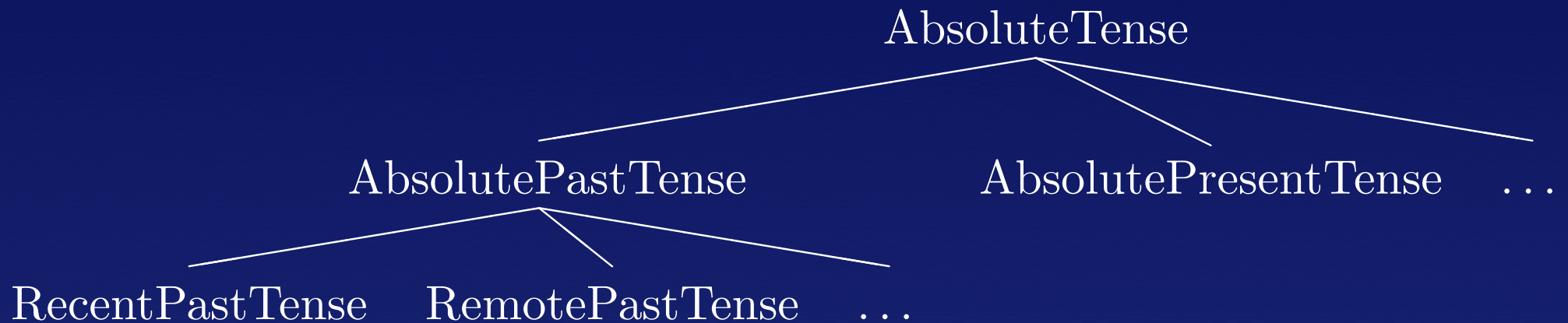
- **GOLD:** The General Ontology for Linguistic Description (hence, “GOLD ontology” is, technically, redundant, but usefully so, as we will see)
- The GOLD project initially viewed its primary aim as the production of an “object”—namely a machine-readable ontology of linguistic terminology

The GOLD Ontology

- **Ontology:** A structured system of categories specified for the relationships which hold among those categories
- The GOLD ontology is intended to apply generally to all descriptive linguistic resources

The GOLD Ontology

- A “sample” from the first version of the GOLD ontology



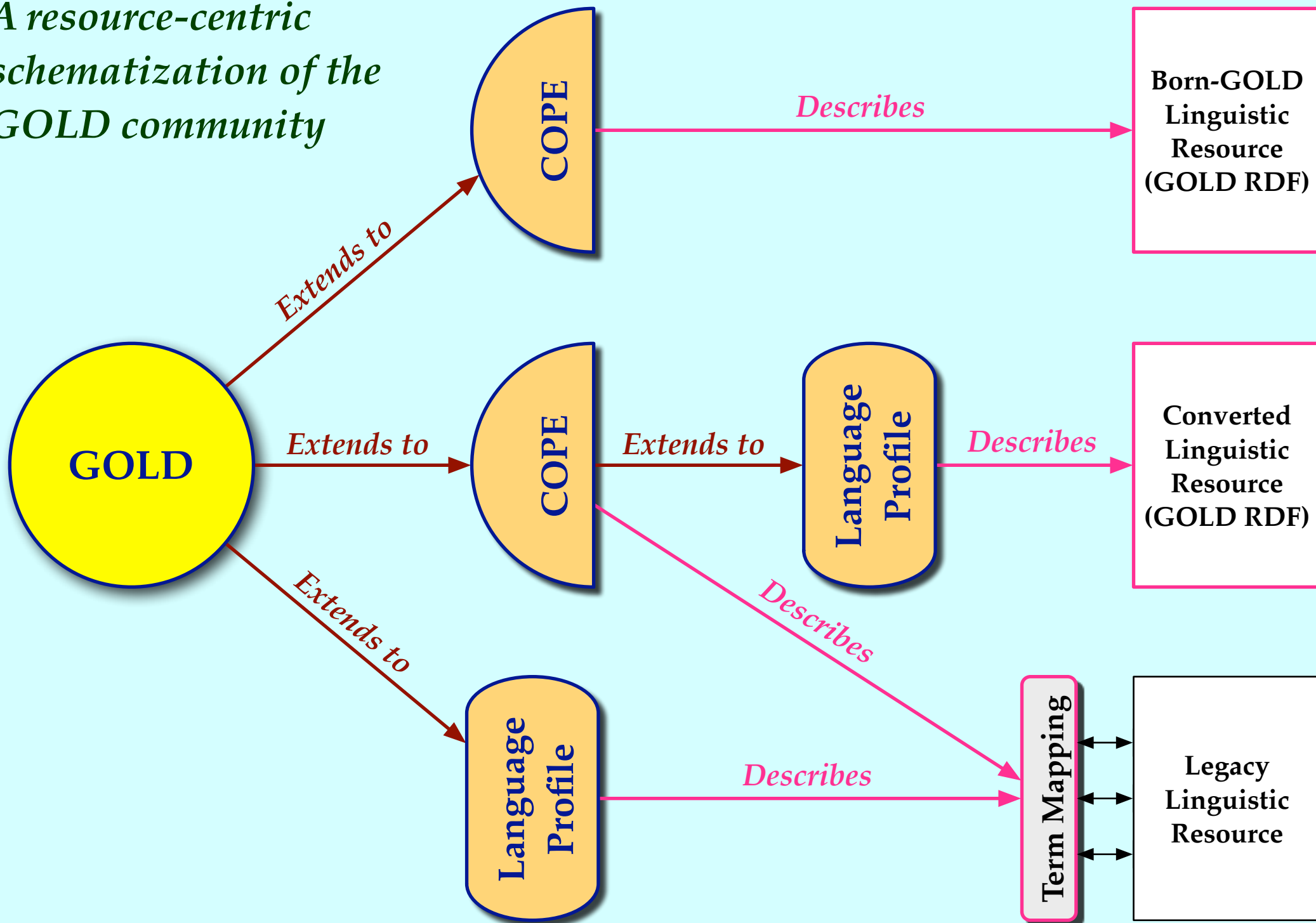
The GOLD Community

- In November of 2004, GOLD was re-envisioned to encompass
 - The production of general linguistic ontology
 - The development of a community of linguists using interlinked linguistic ontologies

The GOLD Community

- The structure of the GOLD community can be schematized with reference to different types of ontological resources
 - GOLD
 - COPEs (Community of Practice Extensions)
 - Language Profiles
- In addition, there is a minor object called a term mapping

*A resource-centric
schematization of the
GOLD community*



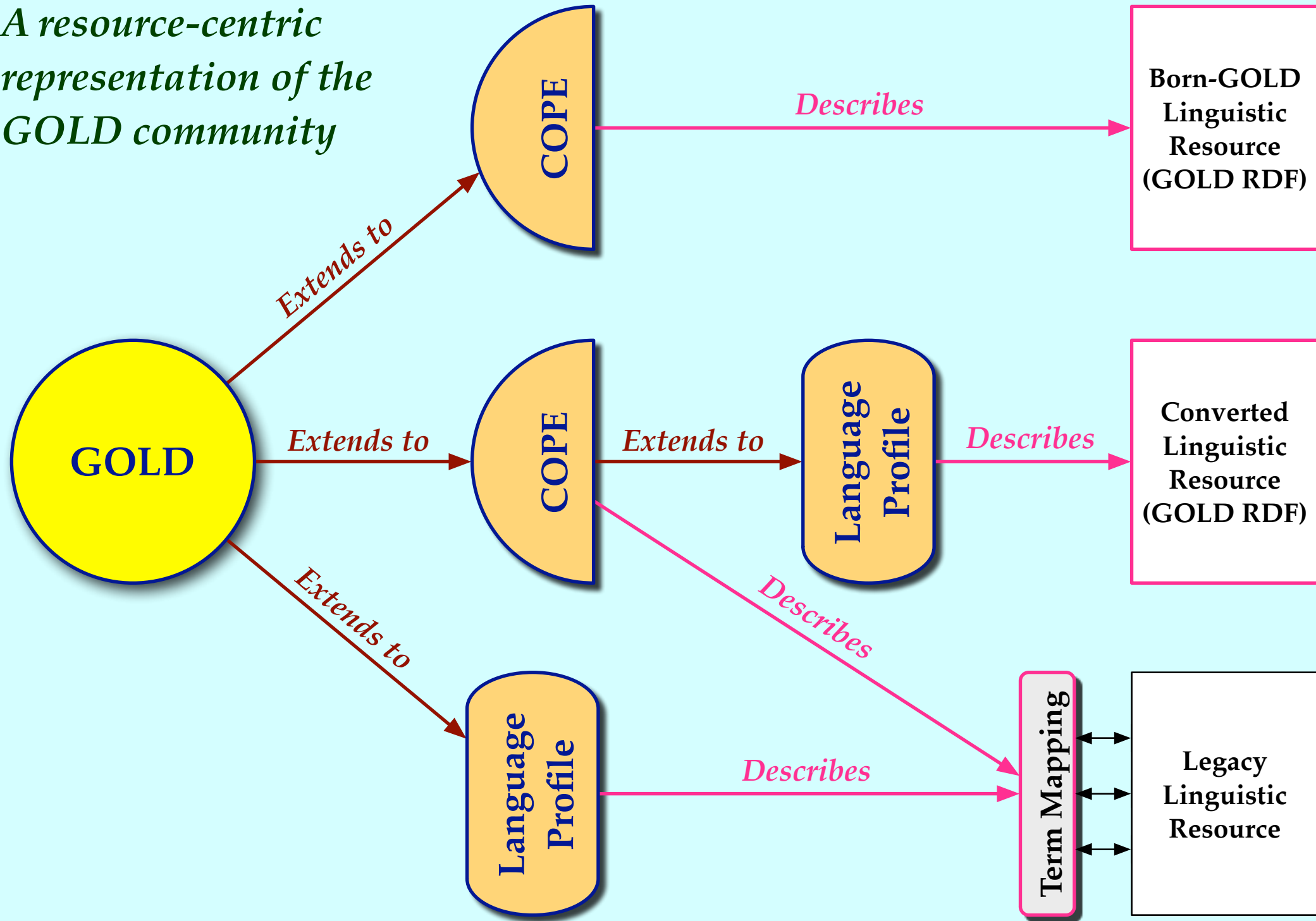
The GOLD Community

- **COPE:** A Community of Practice Extension to GOLD which defines an ontology of terms adopted by a subcommunity of linguists and relates it to GOLD
- **Example:** A Bantu COPE would define the set of noun classes needed for Bantu description and relate those noun classes to GOLD concepts
- But, what will such relationships look like?

The GOLD Community

- **Language Profile:** Conceptually similar to a COPE, except the term set is understood to apply to a particular language
- Language Profiles may be related to directly to GOLD or related to GOLD via a COPE
- Also, a COPE could be related to GOLD via another COPE (or both...)

A resource-centric representation of the GOLD community



The GOLD Community

- Notes on implementation
 - The GOLD ontology, COPEs, and Language Profiles will all be constructed using a common ontological language
 - From an implementation standpoint, therefore, the resources of the GOLD community should seamlessly form a single super-ontology

WALS Overview

- **WALS:** The World Atlas of Language Structures
- WALS is a large-scale effort to create to database of typological databases to facilitate the production of typological maps
- The primary “deliverables” of WALS are a forthcoming physical atlas and a program for drawing maps based on WALS

WALS Overview

- The structure of WALS can be usefully understood as follows
 - 141 individual typological databases by more than forty authors
 - A common metadata standard used by each database to allow for partial interoperability among them
- The WALS editors, though, tend to conceptualize it as one database...

WALS Overview

- The metadata standard includes such things as
 - A series of language codes
 - Geographic locations for the languages
 - Genealogical affiliations for each language

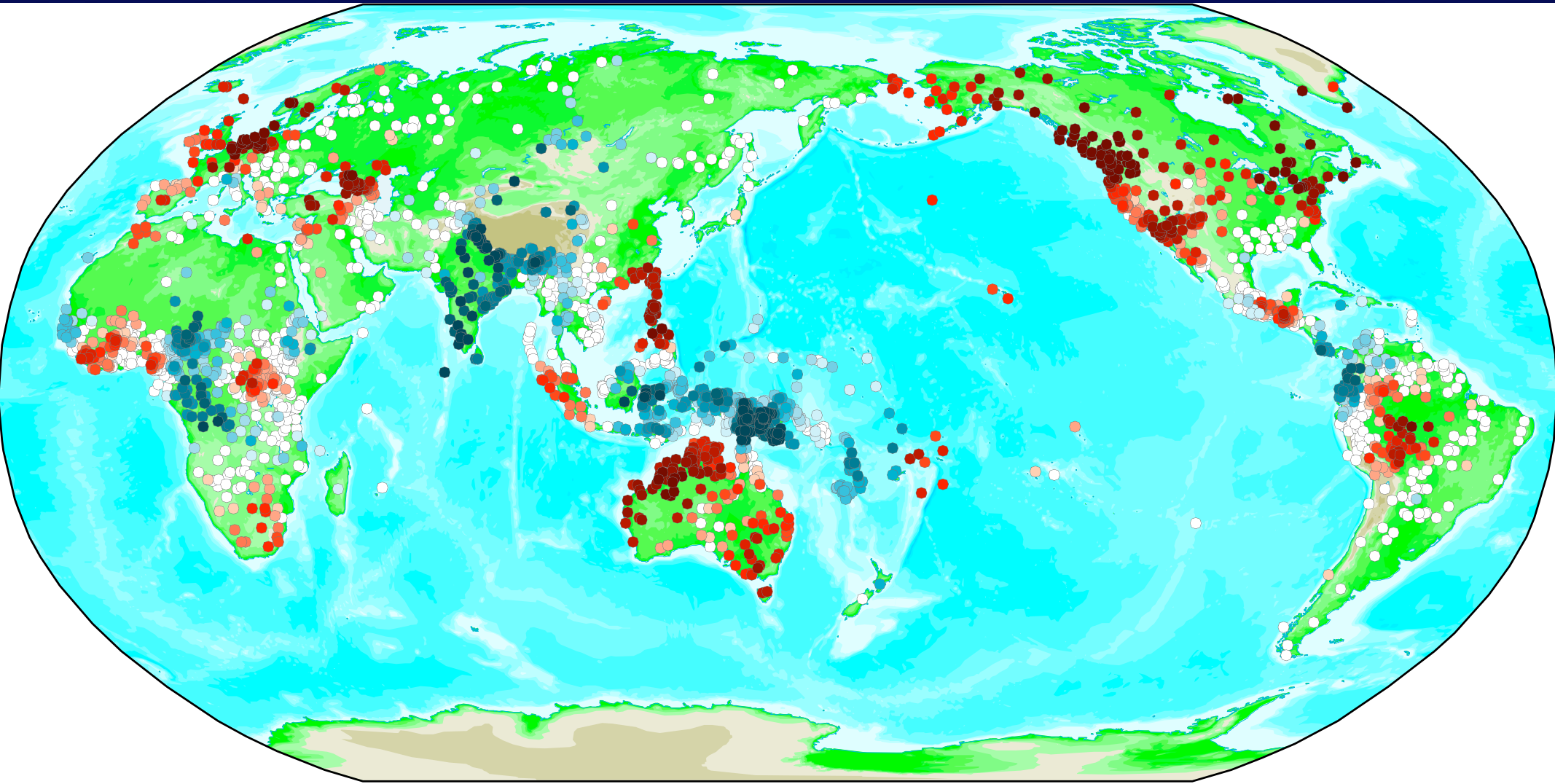
WALS Overview

- Some uses of WALS
 - A reference for 141 typological feature systems
 - Discovering and refining linguistic areas
 - Discovering correlations among different types of typological features
 - Assisting in long-range genealogical analysis

WALS Overview

- Demo of beta version of WALS program
 - Reduplication
 - Numeral bases
 - Basic color terms
 - Composition of features
- After that, something fancy using the whole database but not the WALS program

An aggregate WALS map (Cysouw 2004)



Red = rare; Blue = common

WALS Overview

- Importantly, only the WALS metadata was standardized
- No attempt was made to standardize cross-database usage of categories
- In order to obtain reliable cross-database typological results, the interpretation of the categories needs to be standardized
- Sounds like a job for an ontology...

WALS Ontology

- Three broad problems
 - Non-canonical linguistic categories
 - Implicit logical dependencies
 - Interdatabase category relationships
- The first and the third problem offer challenges for building a WALS COPE
- The second is straightforwardly handled

WALS Ontology

- **Canonical linguistic category:** A linguistic categorization based on the *presence* of a well-definable grammatical feature
- For example
 - “Has accusative case”
 - “Has labial fricative”

WALS Ontology

- **Non-canonical linguistic category:** A linguistic categorization which is not canonical, for example one based on the absence of a feature or by a feature not straightforwardly definable
- For example
 - “Missing /p/” (*absence* category)
 - “Complex syllable structure” (*fuzzy* category)

WALS Ontology

- From an ontological perspective canonical categories can be linked to an ontology using an “instance of” relation
- For example
 - English /p/ is an instance of the GOLD class *voiceless bilabial stop*
 - Latin Nominative Case is an instance of the GOLD class *nominative case*

WALS Ontology

- Non-canonical categories cannot be expressed using an “instance of” relation
- For example
 - “Missing /p/” means that some language *lacks* an instance of some GOLD category
 - “Complex syllable structure” relates a language’s syllable structure to the syllable structures of all other languages

WALS Ontology

- Part of the solution to this problem would seem to lie in enumerating a sufficiently rich set of linguistic ontological “predicates”
- In addition to “instance of”, perhaps
 - “absence of”
 - “no information on”
 - “similar to”

WALS Ontology

- There would seem to be at least two possible solutions to the problem of fuzzy categories
- As part of the standard, require they have non-fuzzy definitions
- For example, “Complex syllable structure” could be defined as allowing syllables of form “CCVC”

WALS Ontology

- Or, perhaps, important fuzzy concepts could be defined as modifiers in GOLD
- **Complex:** For a given gradable linguistic feature X , an instantiation of X is complex if that instantiation is in the upper quartile of complexity of attested instances of X using some specified complexity metric
- Or whatever...

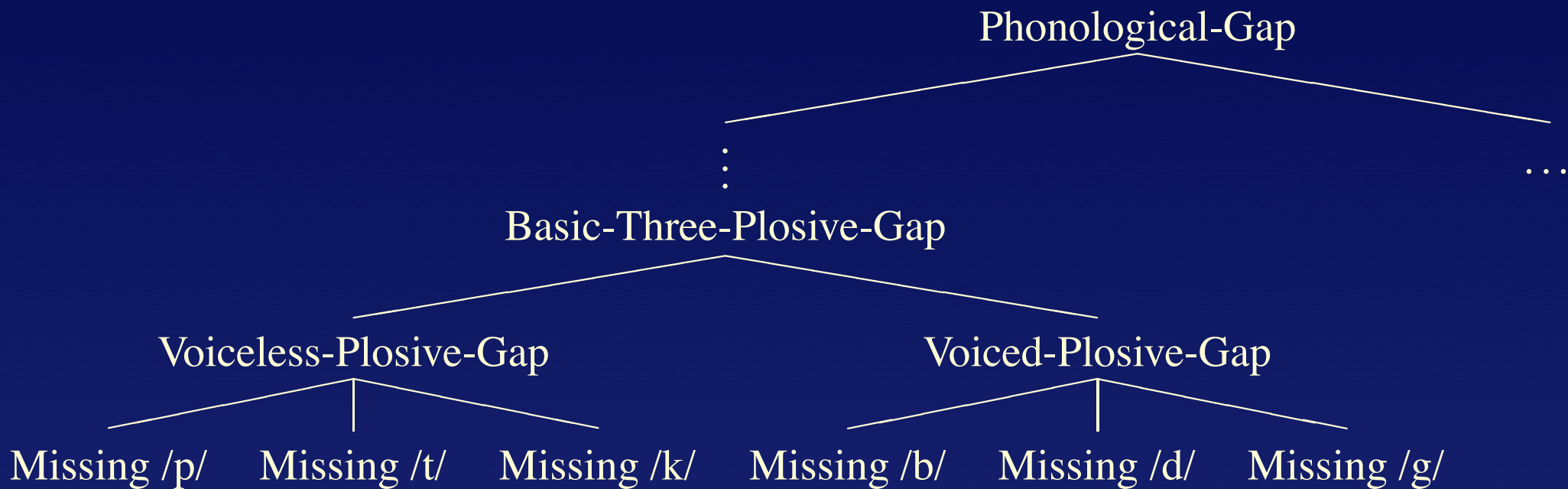
WALS Ontology

- The first solution would seemingly be easier to implement
- It would simply require a COPE defining a fuzzy category concretely in terms of GOLD concepts
- It is already in line with some cases of typological practice (e.g., Nichols (1992))
- Perhaps there are possible solutions other than these two worth exploring

WALS Ontology

- Implicit logical dependencies are found throughout WALS
- “Voicing gaps in plosive systems”
 - Missing /p/
 - Missing /g/
 - Missing both
 - None missing in /p t k b d g/
 - Other

The structure of the “Voicing gaps in plosive systems” categories



“Missing both” = Missing /p/ and Missing /g/

“Other” = Missing /t/ or Missing /k/ or Missing /b/ or Missing /d/

“Missing none” = Absence of Basic-Three-Plosive-Gap

WALS Ontology

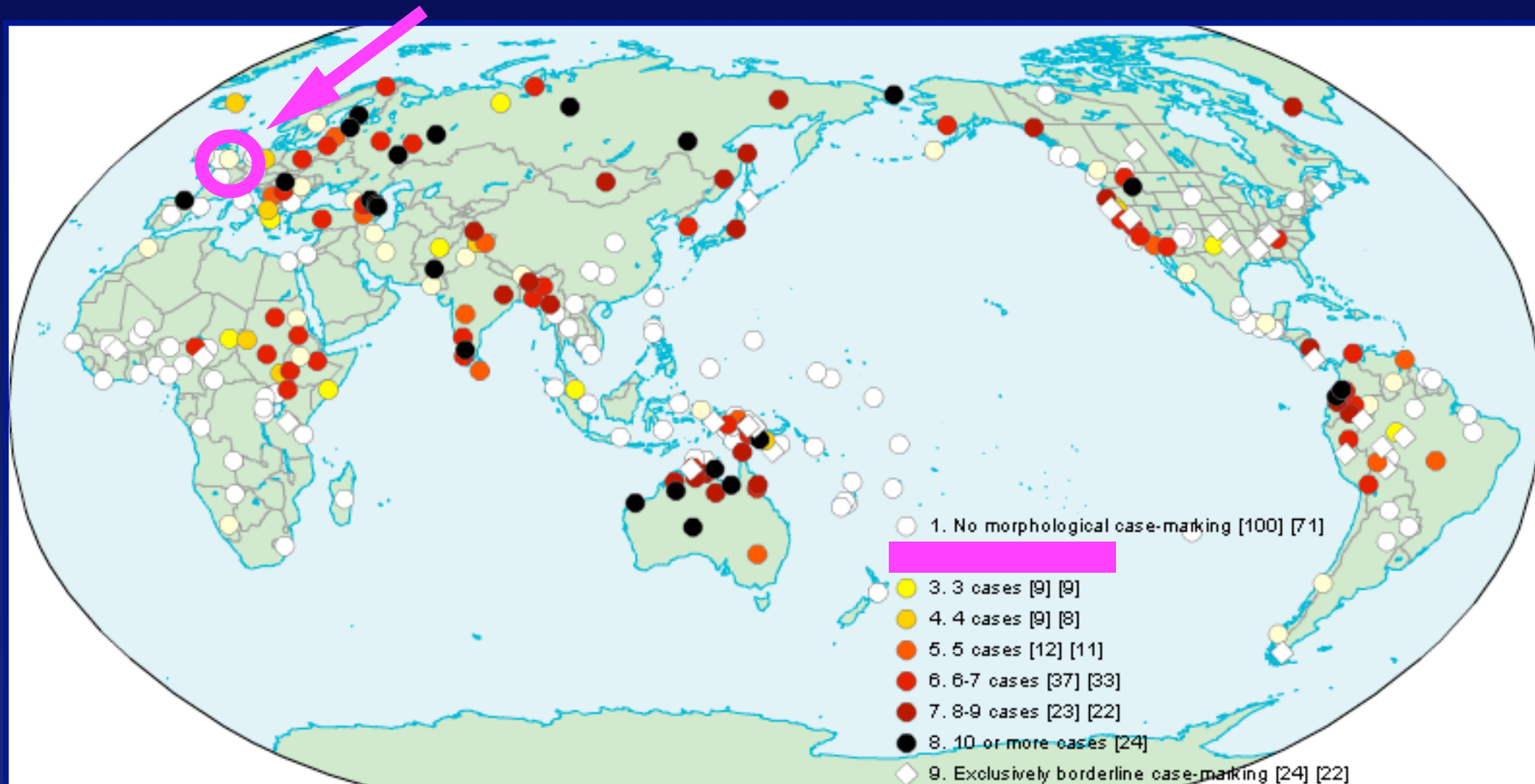
- From a human-labor perspective, implicit logical dependencies present a major problem
- Each set of categories must have its logical dependencies worked out “by hand”
- From an ontological perspective, fortunately, such dependencies do not seem to present a problem at all

WALS Ontology

- Interdatabase category relationships are a particularly thorny problem in WALS
- For example, five of the databases make use of a category name that should correspond to “no case marking”
- However, the same language may be categorized as having no case in one map but as having case in another map

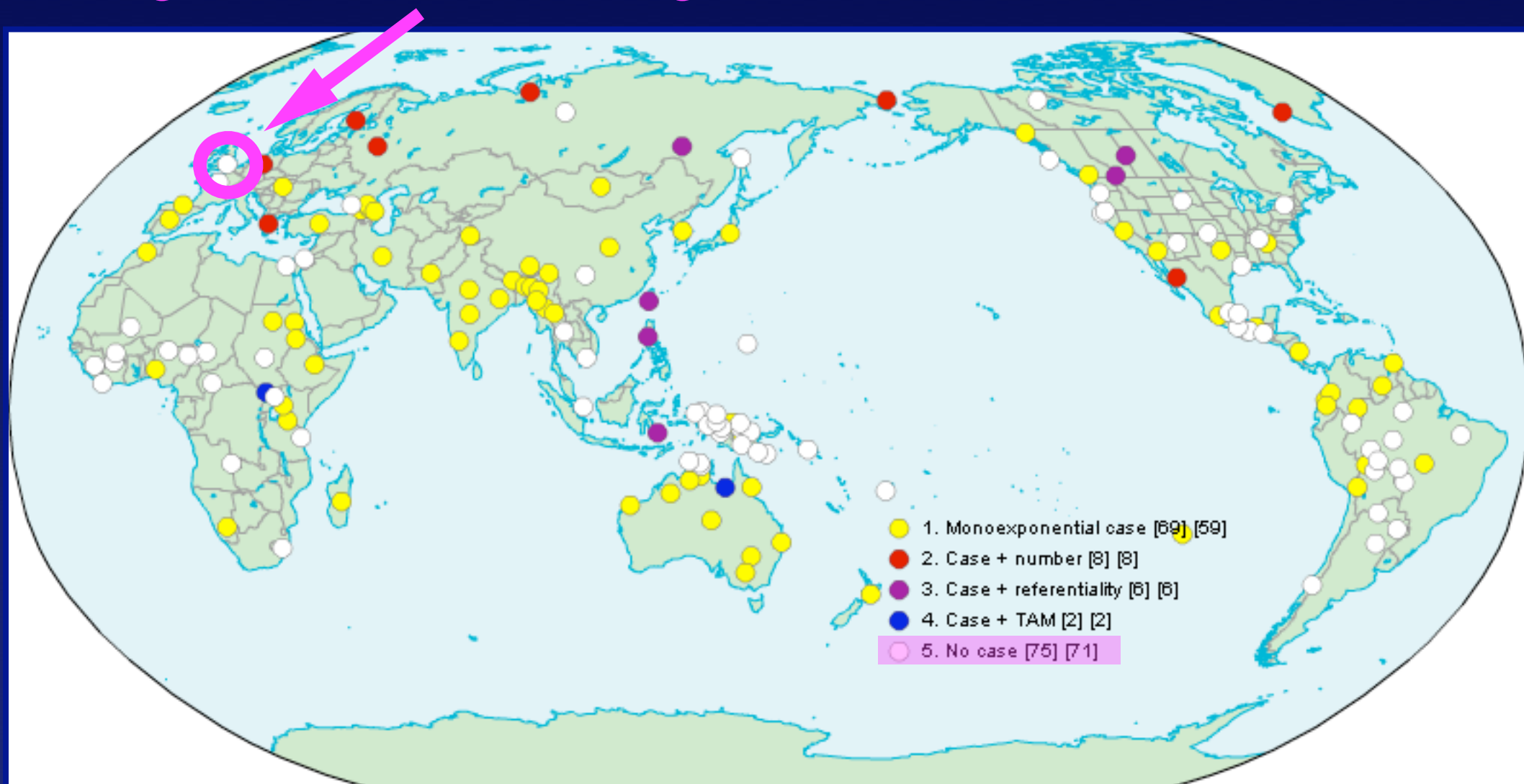
The “Number of cases” map

English coded as having two cases



The “Exponence of selected grammatical formatives” map

English coded as having no case



WALS Ontology

- Such “disagreements” could be caused by various things
 - Genuine disagreement about how to categorize a language
 - Consultation of different sources
 - **Typological “false friends”**: One “term” with different senses depending on context—an ontological problem

WALS Ontology

- In principle, such false friends could be handled by making sure they are ultimately associated with different GOLD concepts
- However, until the perfect ontology is created, it seems useful to be able to indicate “lateral” connections among the categories within different resources

WALS Ontology

- Some possible lateral relationships
 - “Informally equivalent category labels” (e.g., “No case” and “No case marking”)
 - “Same concept”
 - “Same label, same author”
 - “Same label, different author”
 - “Similar label, very different category”

WALS Ontology

- Some possible lateral relationships (contd.)
 - “Theoretically equivalent, empirically distinct” (e.g., the “no case” category discussed above)
 - “Theoretically equivalent, empirically equivalent” (i.e., no conflicting categorizations of languages)

WALS Ontology

- Ultimately, lateral relationships should only be necessary for annotating legacy resources
- But, right now, there are only legacy resources
- And we'll need strategies, like encoding lateral relationships, that facilitate their transformation

WALS Ontology

- One can imagine building a tool that would
 - Facilitate marking up resources for both ontological and lateral relationships
 - Use encoded lateral relationships to inform and facilitate the development of a COPE

Conclusions

- GOLD probably can cope with WALS
- But, it will require the development of
 - An appropriate set of linguistic ontological predicates
 - Supported strategies (like encoding lateral relationships) for facilitating the migration of legacy resources to the GOLD-standard

Acknowledgments

Thanks to Mihai Albu, Michael Cysouw, and Hans-Joerg Bibiko for their discussions on the structure of WALS.