Micropublication[1]

Michael Cysouw
MPI-EVA Leipzig

[1] Footnotes for the 21st Century

Keep it simple, keep it local

- Instead of doing everything in one system, split up large databases into independent modules
- Cooperation can take place by sharing some parts, while other aspects remain proprietary
- The challenge is to define small tasks that can be shared independently of other parts

Proposals

- Languoids & Doculects
- Micropublication
- How can this become part of your database ?!

What is a Languoid?

Languoids are *language-like entities*, including

- All kinds of lects:
  - language, dialect, sociolect, idiolect, stylistic register, etc.
- Genealogical groupings:
  - all levels, from dialect cluster to stock
- Geographic groupings:
  - sprachbund, spread zone, macro area, climate zone, continent, etc.

Why languoid?

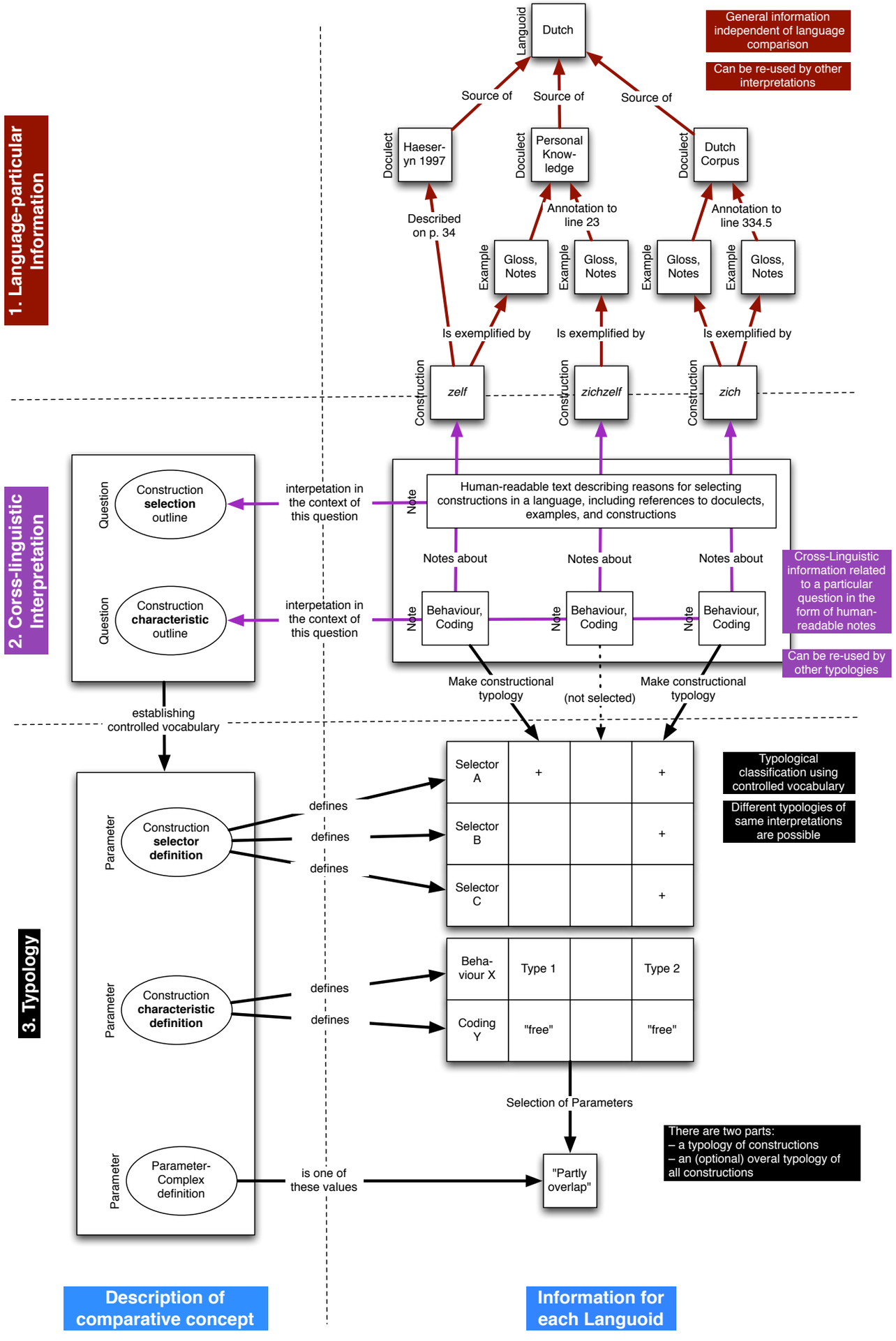
- To allow us to move forward to investigate "languages" while avoiding the insoluble problem of deciding what a "language" is
- A languoid can be catalogued separately from the specification as to what kind of languoid it is
  - Dialect or language?
  - Language or small family?
  - Genealogical or areal group?
  - Different register or different language?

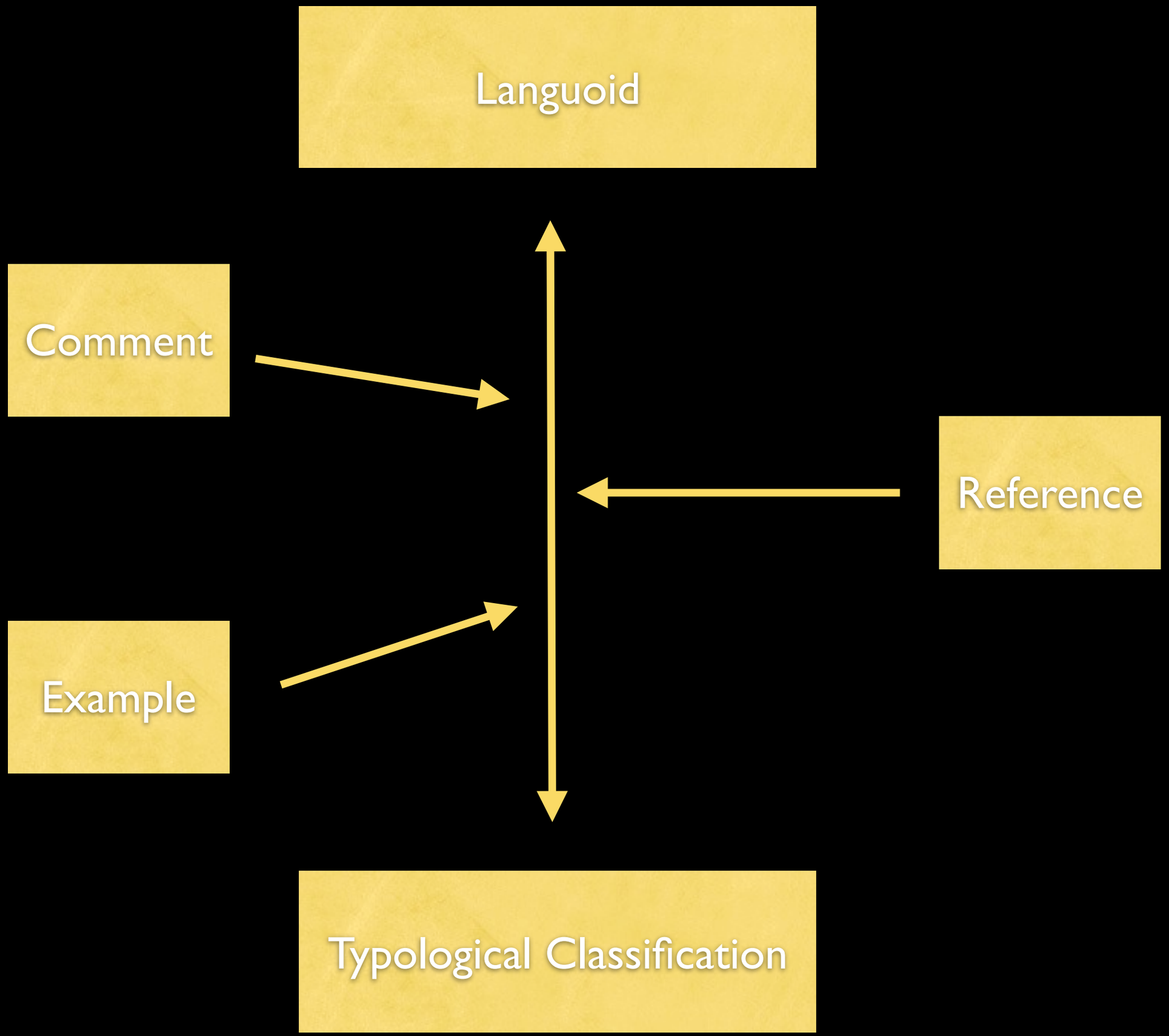
Defining Languoids

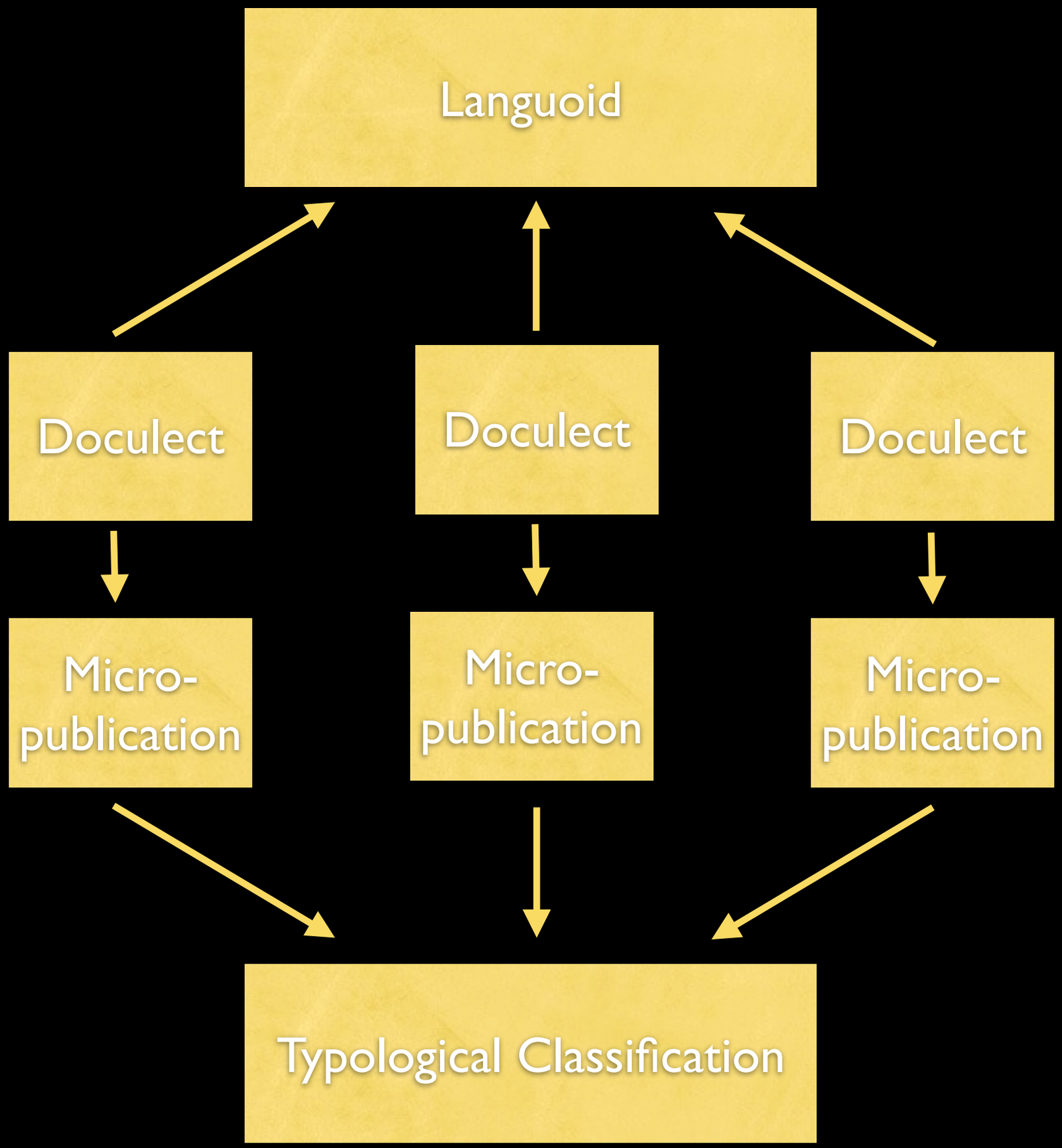
- Languoids are defined as a set of languoids
  - Recursion ends at doculects
- *Doculect:* variety as instantiated by any available documentation
  - Grammatical description (grammar, article)
  - Dictionary, wordlist
  - Inscription, transcription, recording, questionnaire
  - Description of personal knowledge
  - Language notes in a traveller's diary
  - Name given in an ancient text, or in a census

Micropublications
(in the context of linguistic typology)

- Annotations of doculects
- Interpretation of individual doculect in the context of a research question
- Small notes normally not important enough to be published
- Traditionally, such content would at most appear in a footnote or an appendix


1. Language-particular Information
Languoid
Dutch
General information independent of language comparison
Can be re-used by other interpretations
Source of
Source of
Source of
Doculect
Haeser-yn 1997
Doculect
Personal Know-ledge
Doculect
Dutch Corpus
Described on p. 34
Annotation to line 23
Annotation to line 334.5
Example
Gloss, Notes
Example
Gloss, Notes
Example
Gloss, Notes
Example
Gloss, Notes
Is exemplified by
Is exemplified by
Is exemplified by
Construction
zelf
Construction
zichzelf
Construction
zich
2. Corss-linguistic Interpretation
Question
Construction selection outline
Question
Construction characteristic outline
interpetation in the context of this question
interpetation in the context of this question
Note
Human-readable text describing reasons for selecting constructions in a language, including references to doculects, examples, and constructions
Notes about
Notes about
Notes about
Note
Behaviour, Coding
Note
Behaviour, Coding
Note
Behaviour, Coding
Cross-Linguistic information related to a particular question in the form of human-readable notes
Can be re-used by other typologies
Make constructional typology
(not selected)
Make constructional typology
establishing controlled vocabulary
3. Typology
Parameter
Construction selector definition
Parameter
Construction characteristic definition
Parameter
Parameter-Complex definition
defines
defines
defines
defines
defines
is one of these values
Selector A
+
+
Selector B
+
Selector C
+
Beha-viour X
Type 1
Type 2
Coding Y
"free"
"free"
Typological classification using controlled vocabulary
Different typologies of same interpretations are possible
Selection of Parameters
"Partly overlap"
There are two parts:
– a typology of constructions
– an (optional) overal typology of all constructions
Description of comparative concept
Information for each Languoid


Languoid
Comment
Reference
Example
Typological Classification


Languoid
Doculect
Doculect
Doculect
Micro-publication
Micro-publication
Micro-publication
Typological Classification

Typological databases

- The core of typological databases are:
  - Doculects
  - Micropublications (i.e. annotations of doculects)
- This information is reusable !
- Other levels are more debatable:
  - Languoids
  - Typological classifications

Format of Micropublication

- Standard-compliant
- Authored
- Fixed
- Stable
- Referenced

Structure of Micropublication

- Written in natural language
- Annotation to a restricted number of doculects (preferably only one)
- Include snippets !

Micropublication

Footnotes for the 21st Century

« Tagalog content interrogatives

Tagalog 'how' questions

Subject of Annotation

Schachter, Paul, and Fe T. Otanes. *Tagalog Reference Grammar*. Berkeley: University of California Press, 1972. (COinS)

The various questions that are in English constructed using 'how' show an interesting lexicalisation structure in Tagalog (Schachter & Otanes, 1972:514-516) (COinS). There are five Tagalog lexemes that appear to be relevant:

- *kumusta* (from Spanish *comó está*)
- *paano*
- *gaano*
- *ilan*
- *magkano*

kumusta is the interrogative that asks for the quality of something (514). Note that it does not seem to be alike to English *what kind of* questions. A more suitable translation seems to be something like "what kind of word would you