Metadata, ontologies and modelling conventions: linguistic and computational use cases and views on lexical 'data' *

***Data are what you find in a database...**

Dafydd Gibbon
Universität Bielefeld

Workshop: Living Sources in Lexical Description
Berlin 2008-09-30 + 2008-10-01

Some practical open issues

- Language typology and (non-) native speaker knowledge:
  - lemmatisation / canonical forms / access strategies in
    - prefixing languages
      - Niger-Congo (inflection)
      - Celtic (mutation)
    - tone languages
- Procedures:
  - lexical resource creation is software development & thus needs
    - well-defined **procedures**
    - well-defined **modelling conventions**
- Merging or mapping lexicons:
  - lexicon A differs from lexicon B in
    - missing information
    - more information
    - contradictory information

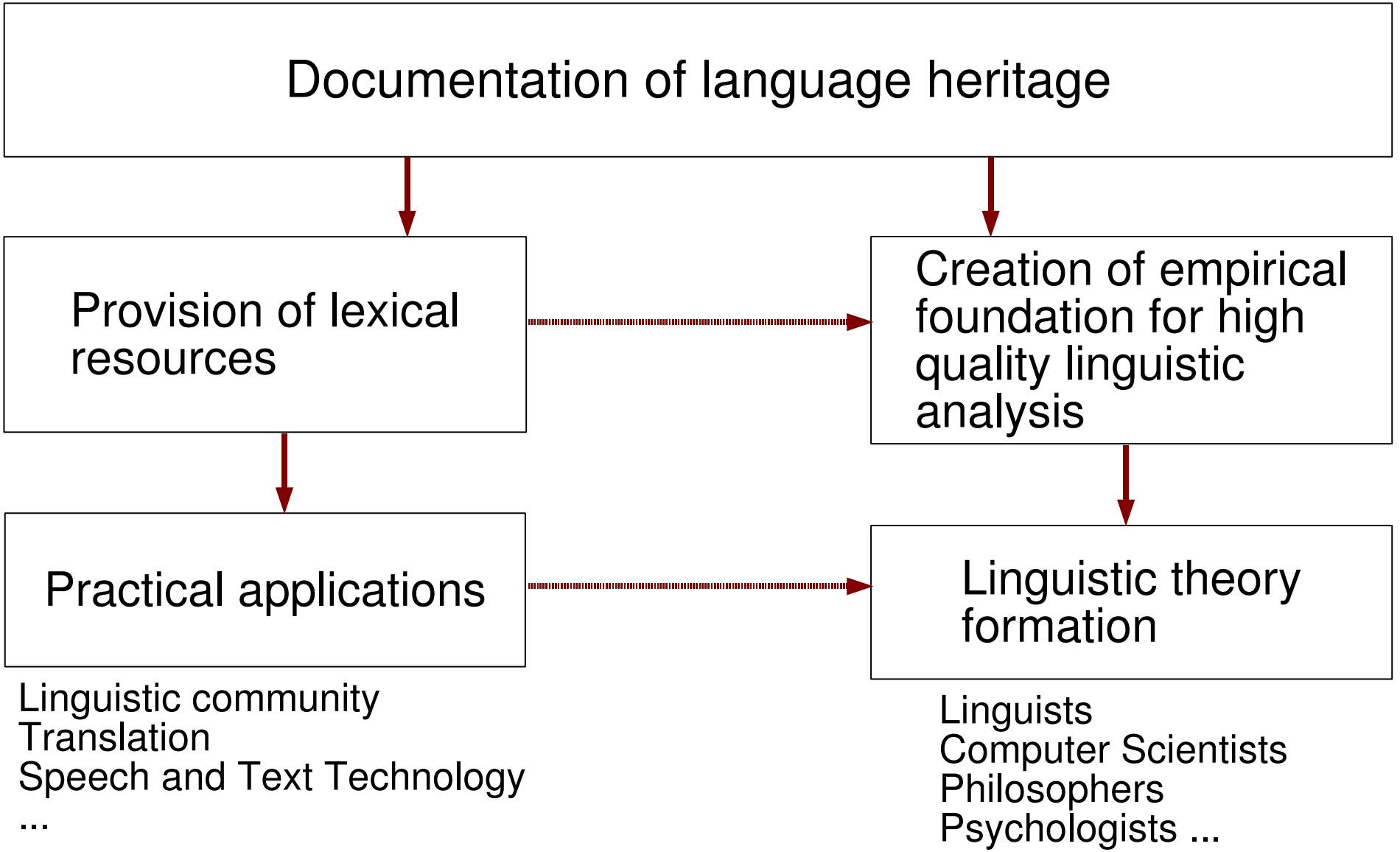
Overview

Background: functionality of lexical resources

Modelling conventions:

1. Lexical resources are software
2. Ontology: objects, meta-objects, metameta-objects
3. The reflexive semiotic ontology
4. Semiotic lexical information
5. Rank levels (idioms &c)
6. Lexicon type hierarchy
7. Lexicon type hierarchy layer 3 – information types
8. Lexicon Type hierarchy layer 3 - mesostructure
9. Procedural issues: lexicographic workflow cycle

Multifunctional use cases for lexical resources



(1) Modelling conventions: resources = software

- Starting point:
  - Creation of lexical resources is software development
  - Why?
    - Requirements level: reusability, interoperability, ...
    - Design level: modelling conventions, data structures, algorithms, interfaces (example: ontologies)
    - Implementation level: tools, databases
    - Evaluation level: quality control wrt design & requirements
- But also:
  - Software development: reality construction
    - input reality: requirements, needs + modelling conventions
    - output reality: transformation of (working) environment
  - And It is a mistake to think that
    - software development has nothing to do with hermeneutics, interpretation, ...
    - lexicon resources have nothing to do with formal modelling

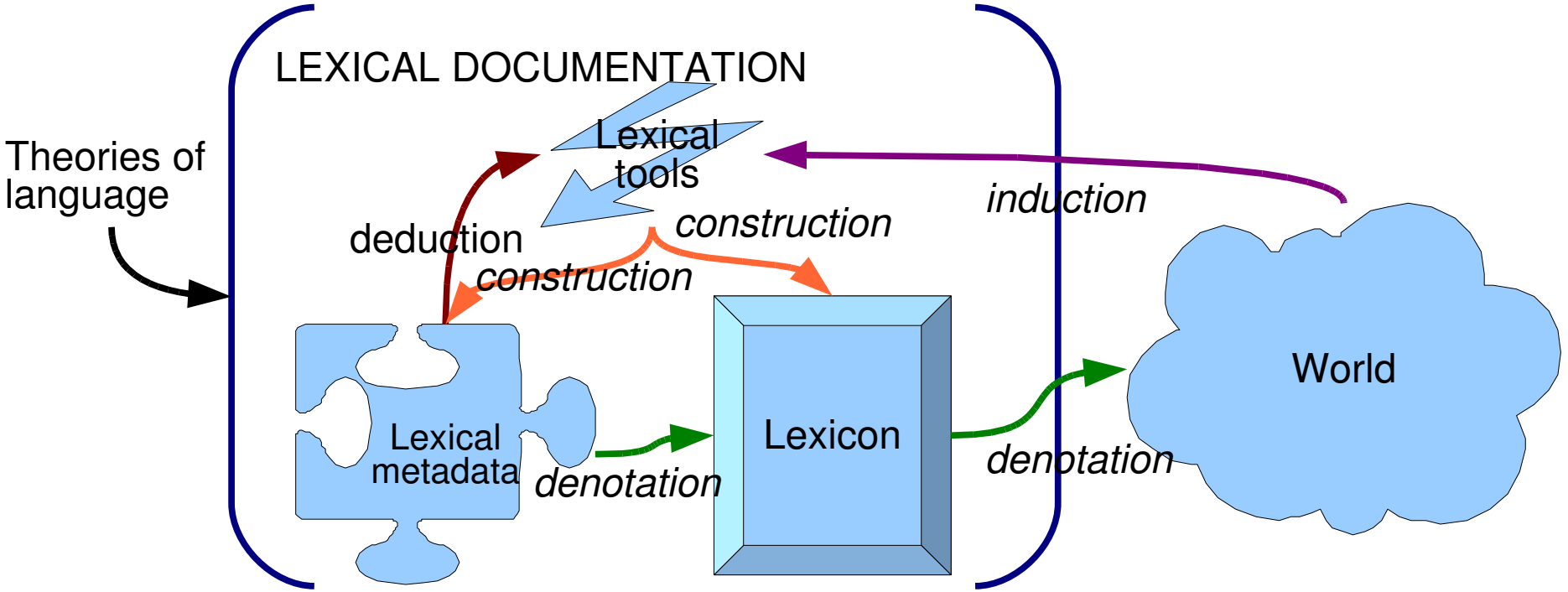
(2) Ontology: objects, meta-(meta-)objects

- Current jargon:
  - classificatory system
  - category system
  - taxonomy
  - mer(e)onomy
  - semantic network

- Ontology - ontologies
  - 'The study of being' (not so far from '*living* sources'?)
- Based on domain-oriented modelling conventions:

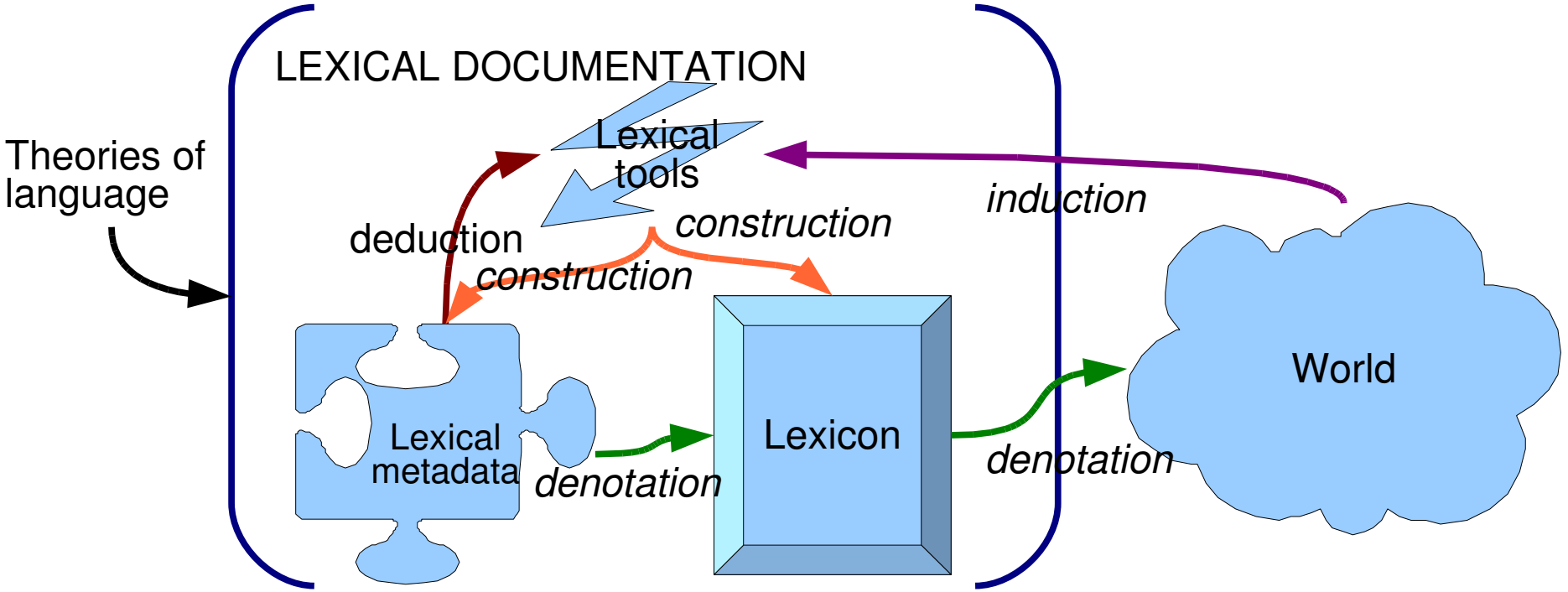
(2) Ontology: objects, meta-(meta-)objects

- Current jargon:
  - classificatory system
  - category system
  - taxonomy
  - mer(e)onomy
  - semantic network
- Ontology - ontologies
  - 'The study of being' (not so far from '*living* sources'?)
- Based on domain-oriented modelling conventions:



(2) Ontology: objects, meta-(meta-)objects

- Language documentation for linguistics:
  - fieldwork, experiments
  - data processing
  - standardisation
  - archiving
- Language documentation for language technologies:
  - fieldwork, experiments
  - data processing
  - standardisation
  - archiving

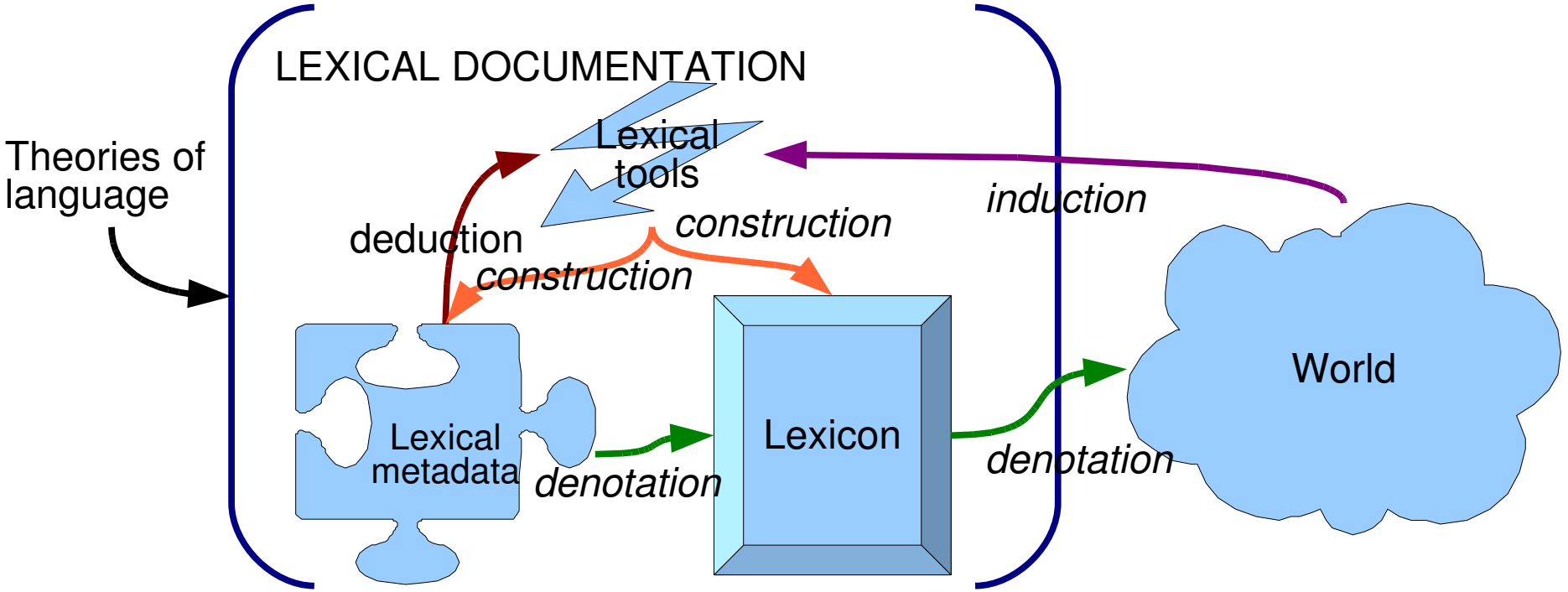


(2) Ontology: objects, meta-(meta-)objects

- Language documentation for linguistics:
  - fieldwork, experiments
  - data processing
  - standardisation
  - archiving

TOOLS

- Language documentation for language technologies:
  - fieldwork, experiments
  - data processing
  - standardisation
  - archiving



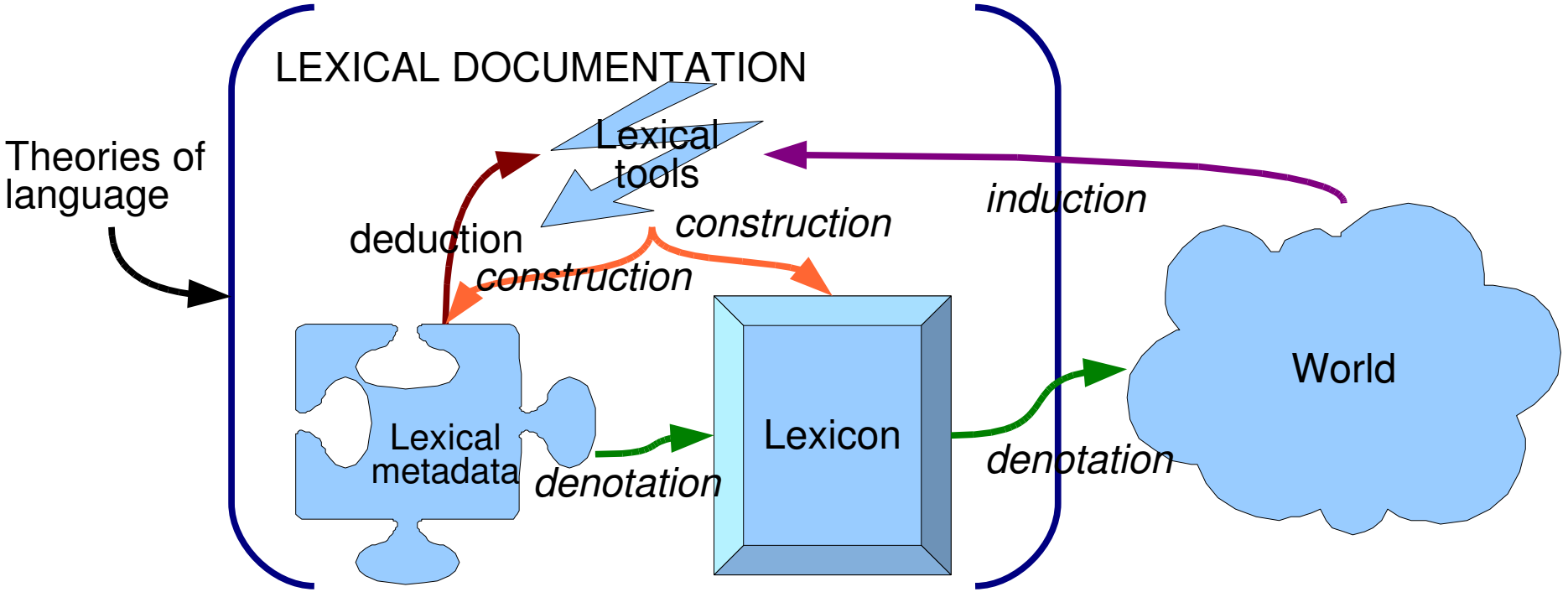
(2) Ontology: objects, meta-(meta-)objects

- Language documentation for linguistics:
  - fieldwork, experiments
  - data processing
  - standardisation
  - archiving

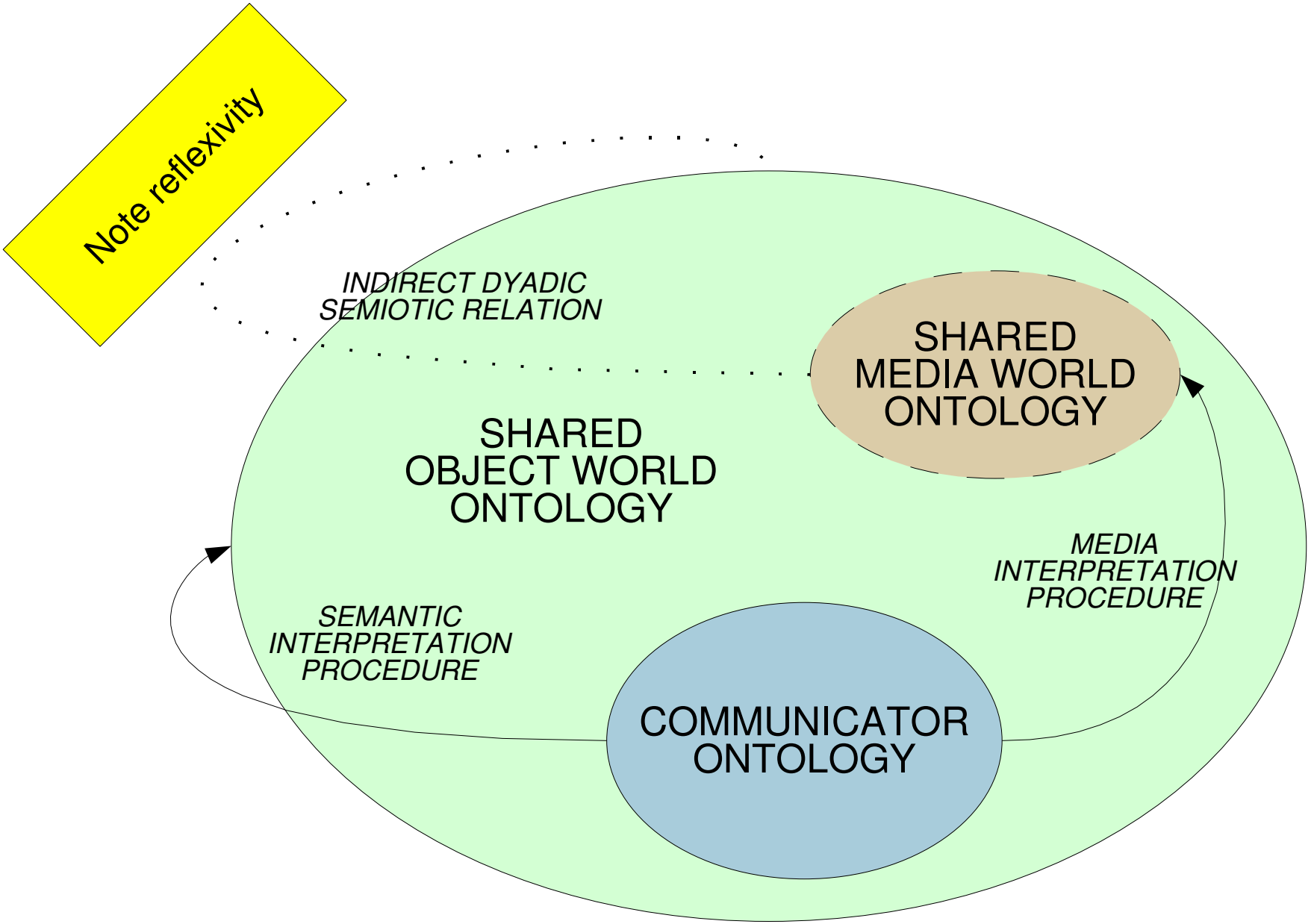
TOOLS

DATA

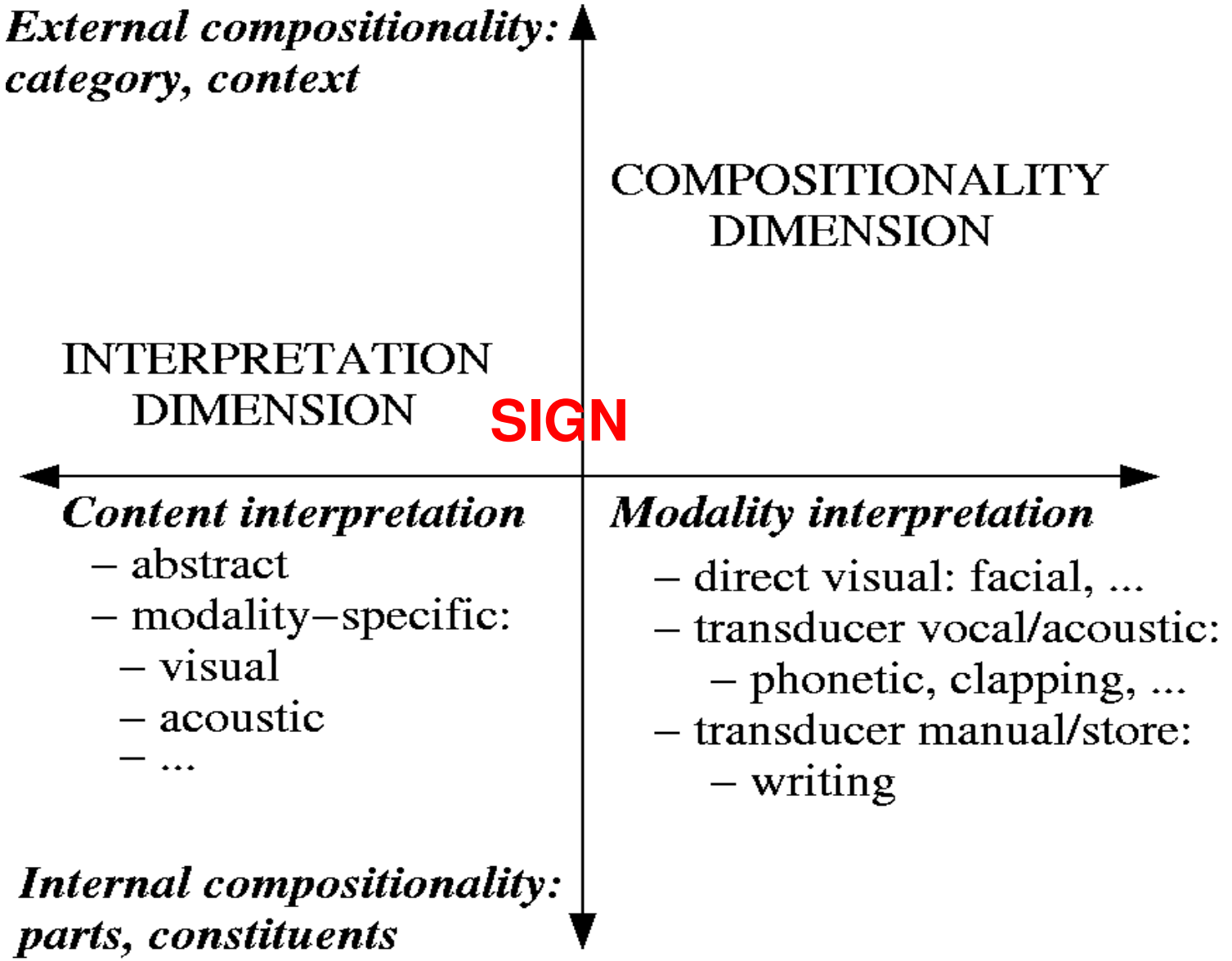
- Language documentation for language technologies:
  - fieldwork, experiments
  - data processing
  - standardisation
  - archiving



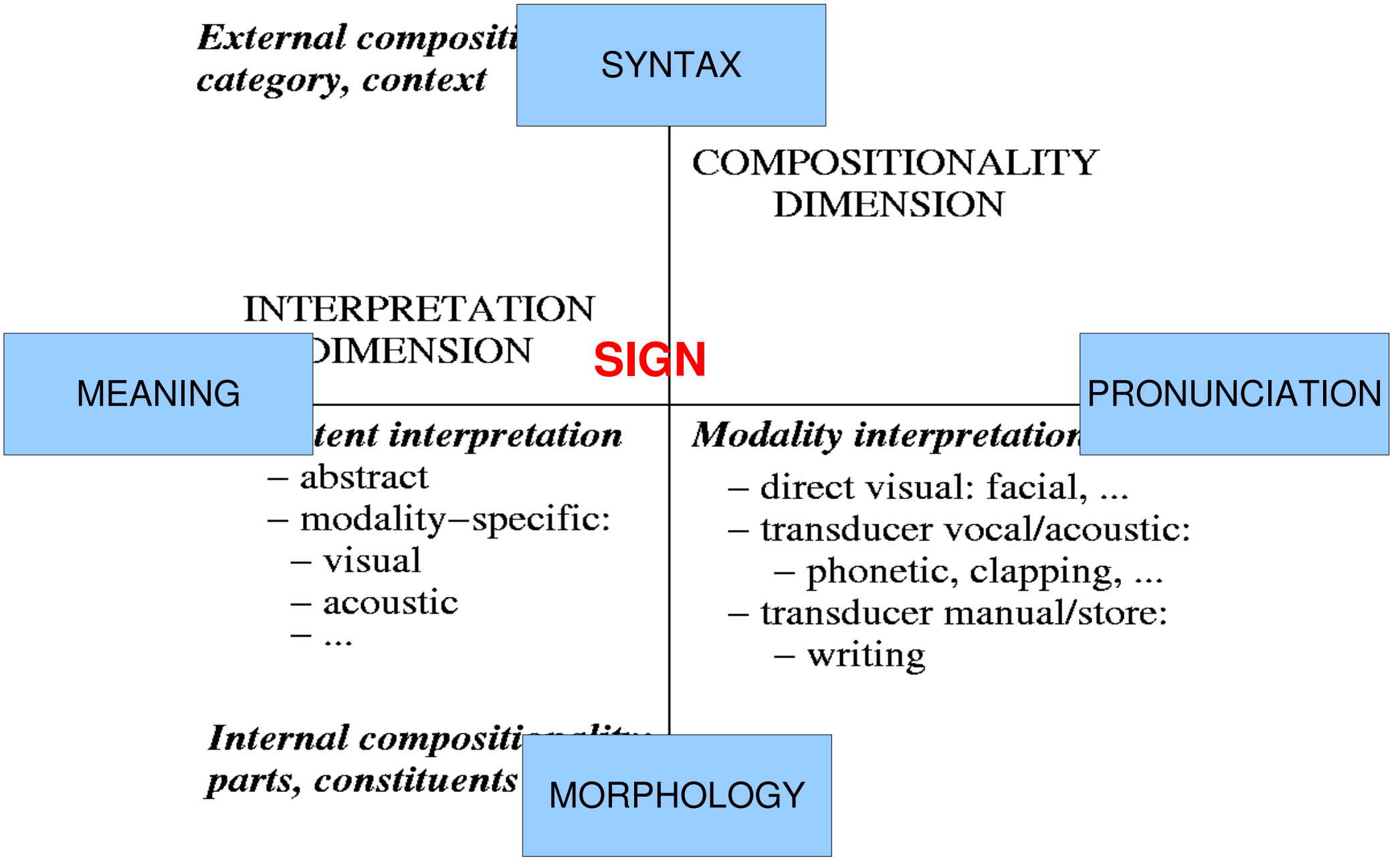
(3) Modelling conventions: semiotic ontology



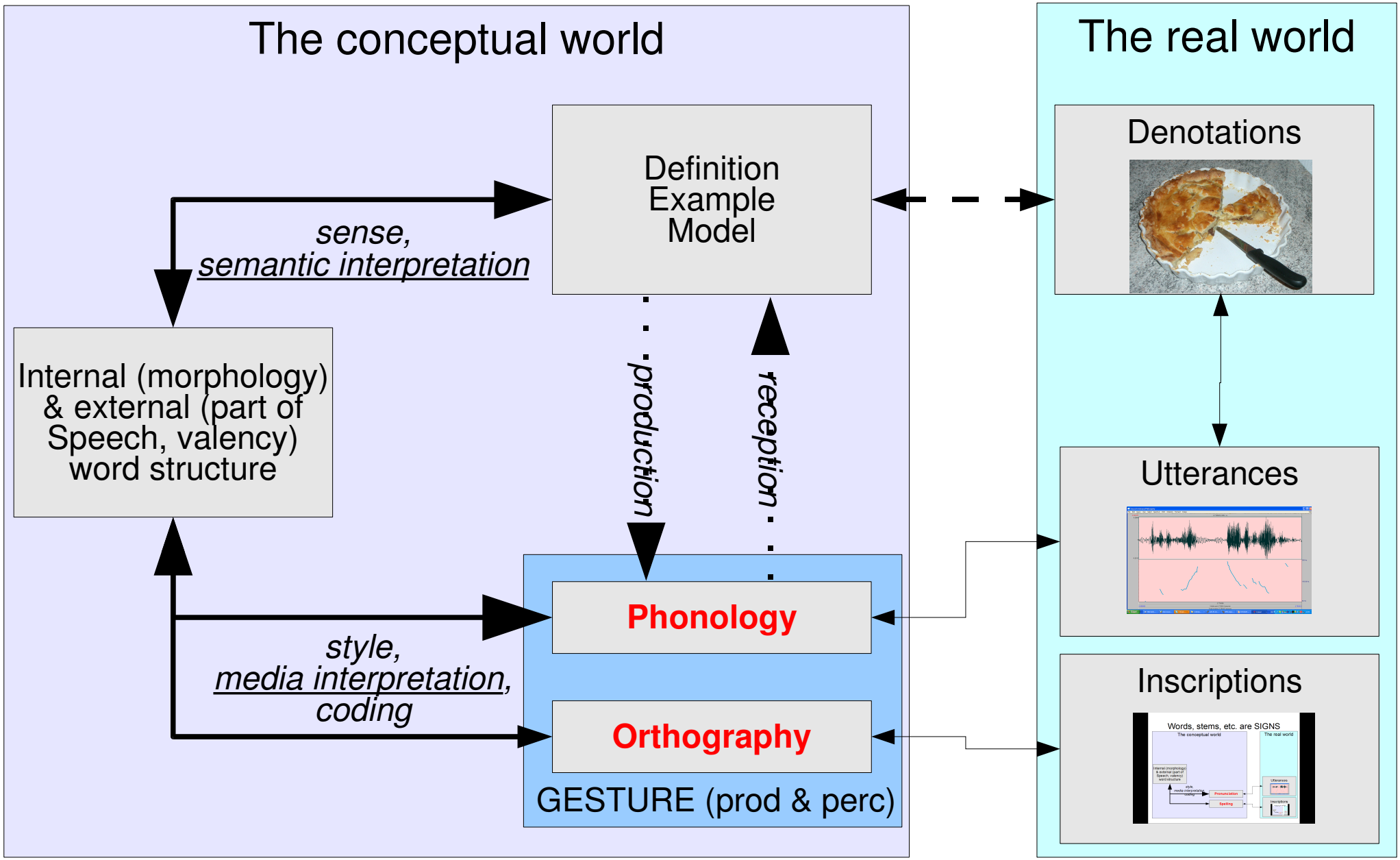
(3) Modelling conventions: semiotic ontology



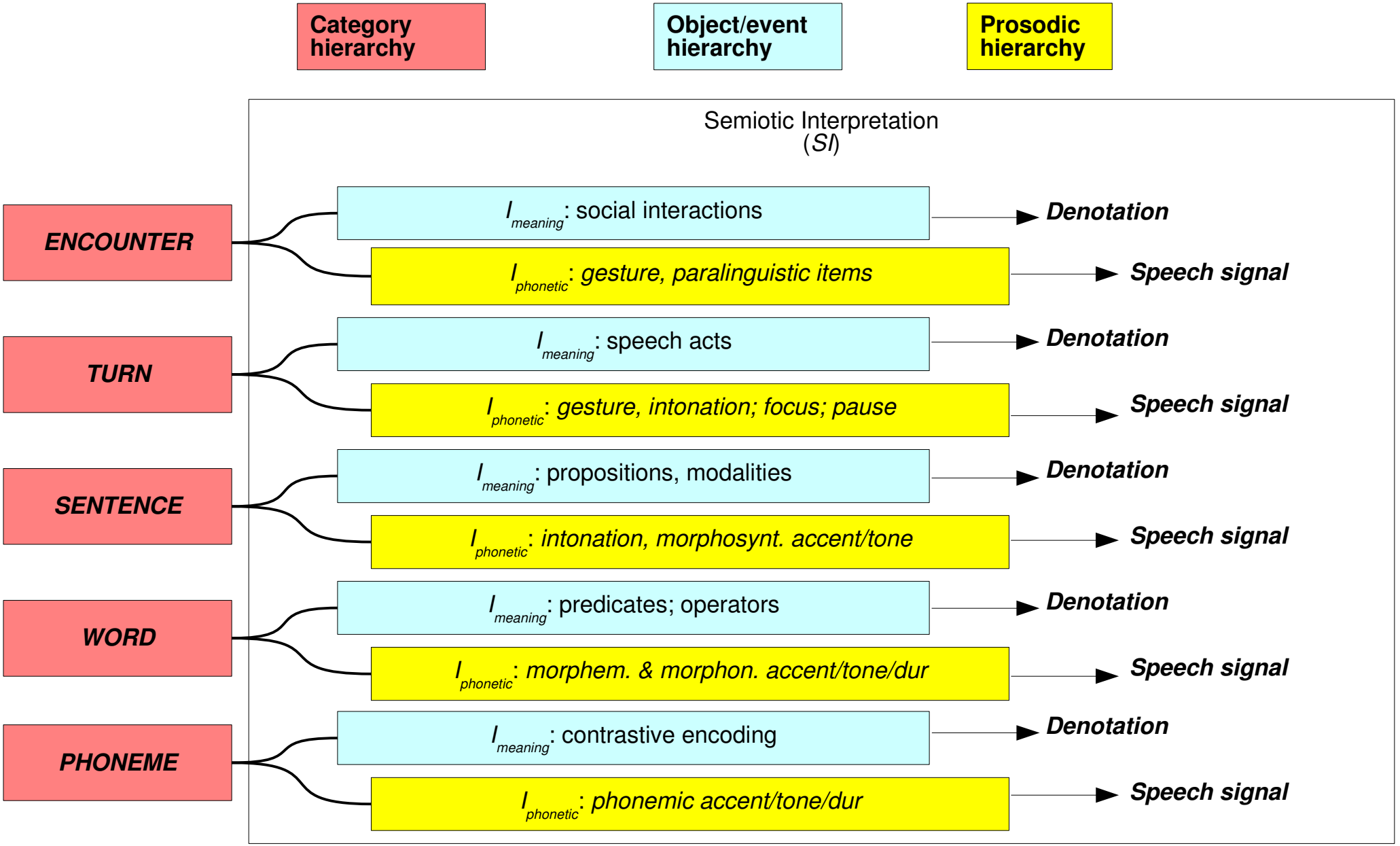
(3) Modelling conventions: semiotic ontology



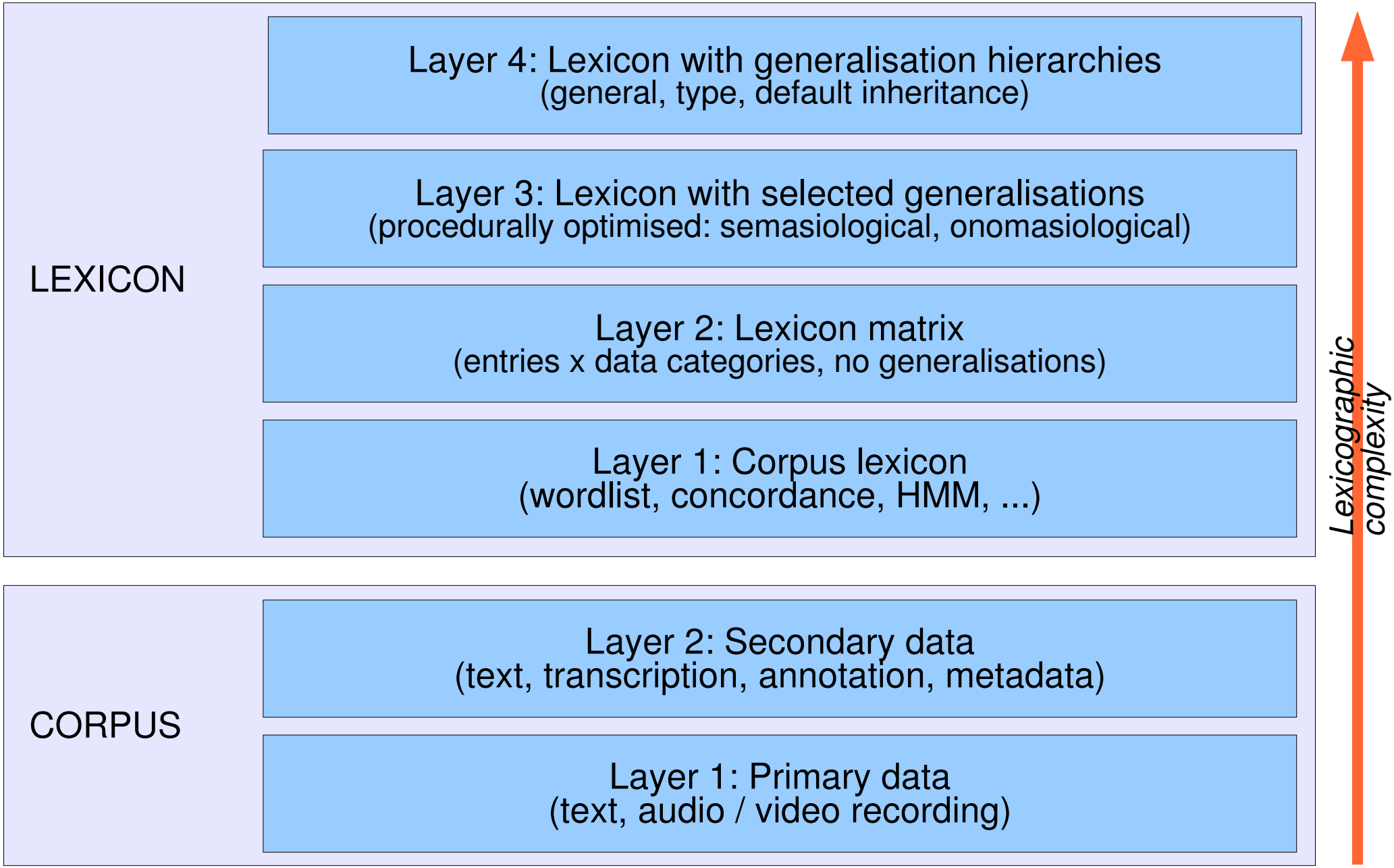
(4) Modelling conventions: semiotic lexical information



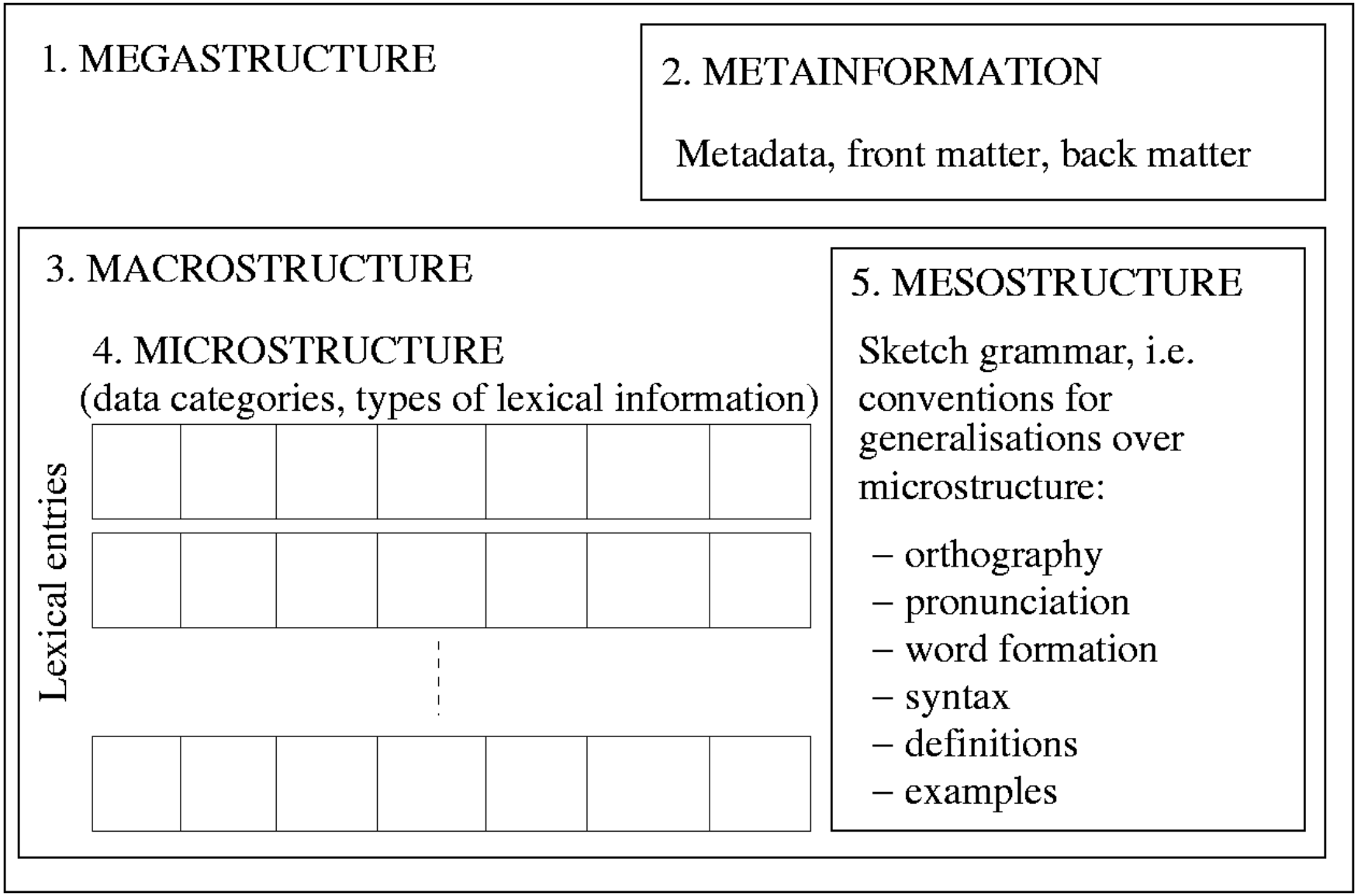
(5) Modelling conventions: rank levels (idioms &c)



(6) Modelling conventions: lexicon type hierarchy



(7) Modelling conventions: layer 3 – information types



(8) Modelling conventions: layer 3 - mesostructure

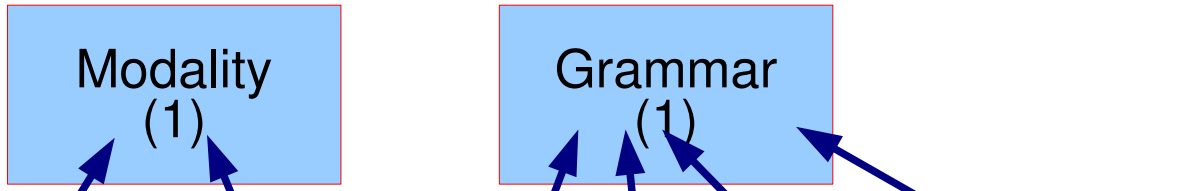
- Linguistically motivated class hierarchy of DATCAT subvectors e.g. modality, grammar, object semantics
- Linguistic description references, e.g. use of abbreviations for parts of speech, characterisations of spelling
- Cross-references between related entries, e.g. co-hyponyms (synonyms, antonyms, ...)
- Corpus references (concordance)

ID	Orthography	Pronunciation	POS	Gender	Morpohology	Inflection_class	Definition	Instance

ID	Orthography	Pronunciation	POS	Gender	Morpohology	Inflection_class	Definition	Instance

(8) Modelling conventions: layer 3 - mesostructure

- Linguistically motivated class hierarchy of DATCAT subvectors e.g. modality, grammar, object semantics
- Linguistic description references, e.g. use of abbreviations for parts of speech, characterisations of spelling
- Cross-references between related entries, e.g. co-hyponyms (synonyms, antonyms, ...)
- Corpus references (concordance)

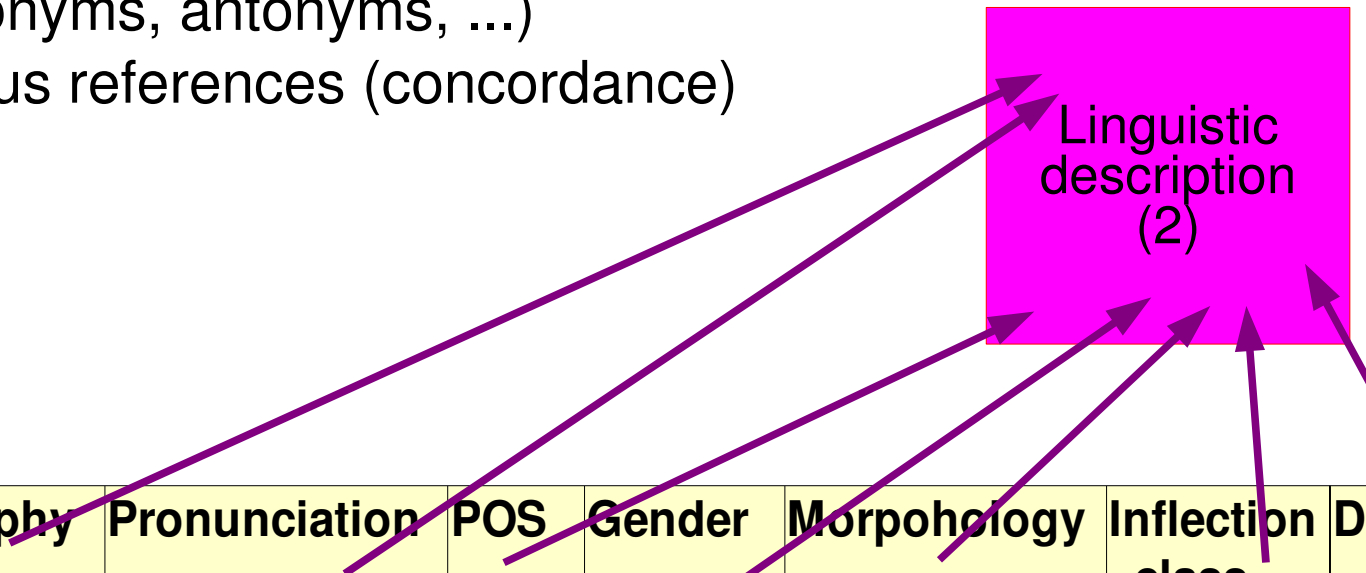


ID	Orthography	Pronunciation	POS	Gender	Morpohology	Inflection_class	Definition	Instance

ID	Orthography	Pronunciation	POS	Gender	Morpohology	Inflection_class	Definition	Instance

(8) Modelling conventions: layer 3 - mesostructure

- Linguistically motivated class hierarchy of DATCAT subvectors e.g. modality, grammar, object semantics
- Linguistic description references, e.g. use of abbreviations for parts of speech, characterisations of spelling
- Cross-references between related entries, e.g. co-hyponyms (synonyms, antonyms, ...)
- Corpus references (concordance)

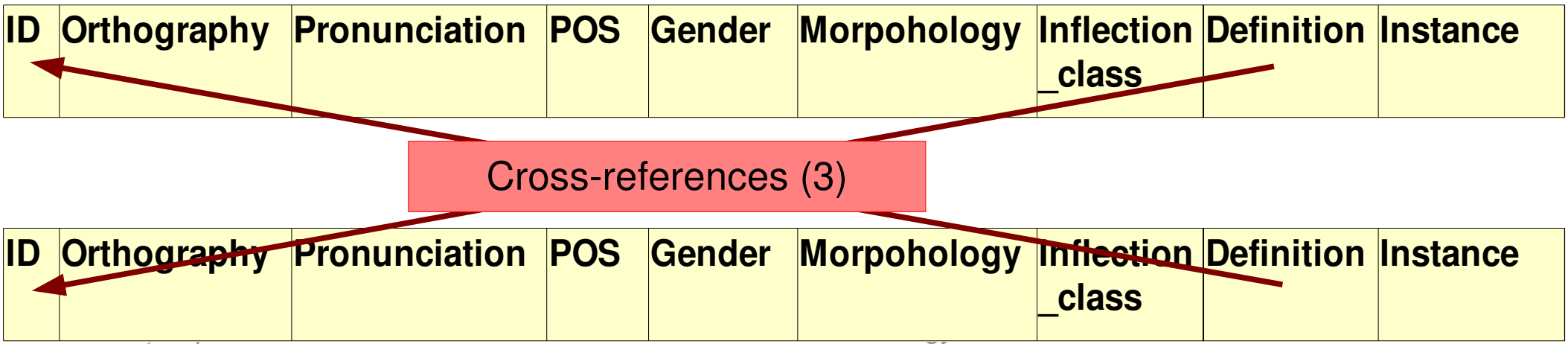


ID	Orthography	Pronunciation	POS	Gender	Morpohology	Inflection_class	Definition	Instance

ID	Orthography	Pronunciation	POS	Gender	Morpohology	Inflection_class	Definition	Instance

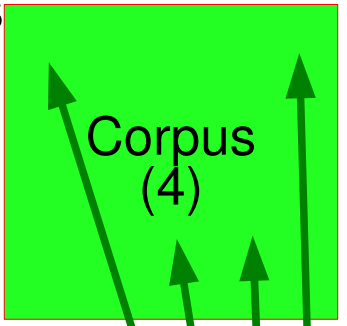
(8) Modelling conventions: layer 3 - mesostructure

- Linguistically motivated class hierarchy of DATCAT subvectors e.g. modality, grammar, object semantics
- Linguistic description references, e.g. use of abbreviations for parts of speech, characterisations of spelling
- Cross-references between related entries, e.g. co-hyponyms (synonyms, antonyms, ...)
- Corpus references (concordance)



(8) Modelling conventions: layer 3 - mesostructure

- Linguistically motivated class hierarchy of DATCAT subvectors e.g. modality, grammar, object semantics
- Linguistic description references, e.g. use of abbreviations for parts of speech, characterisations of spelling
- Cross-references between related entries, e.g. co-hyponyms (synonyms, antonyms, ...)
- Corpus references (concordance)

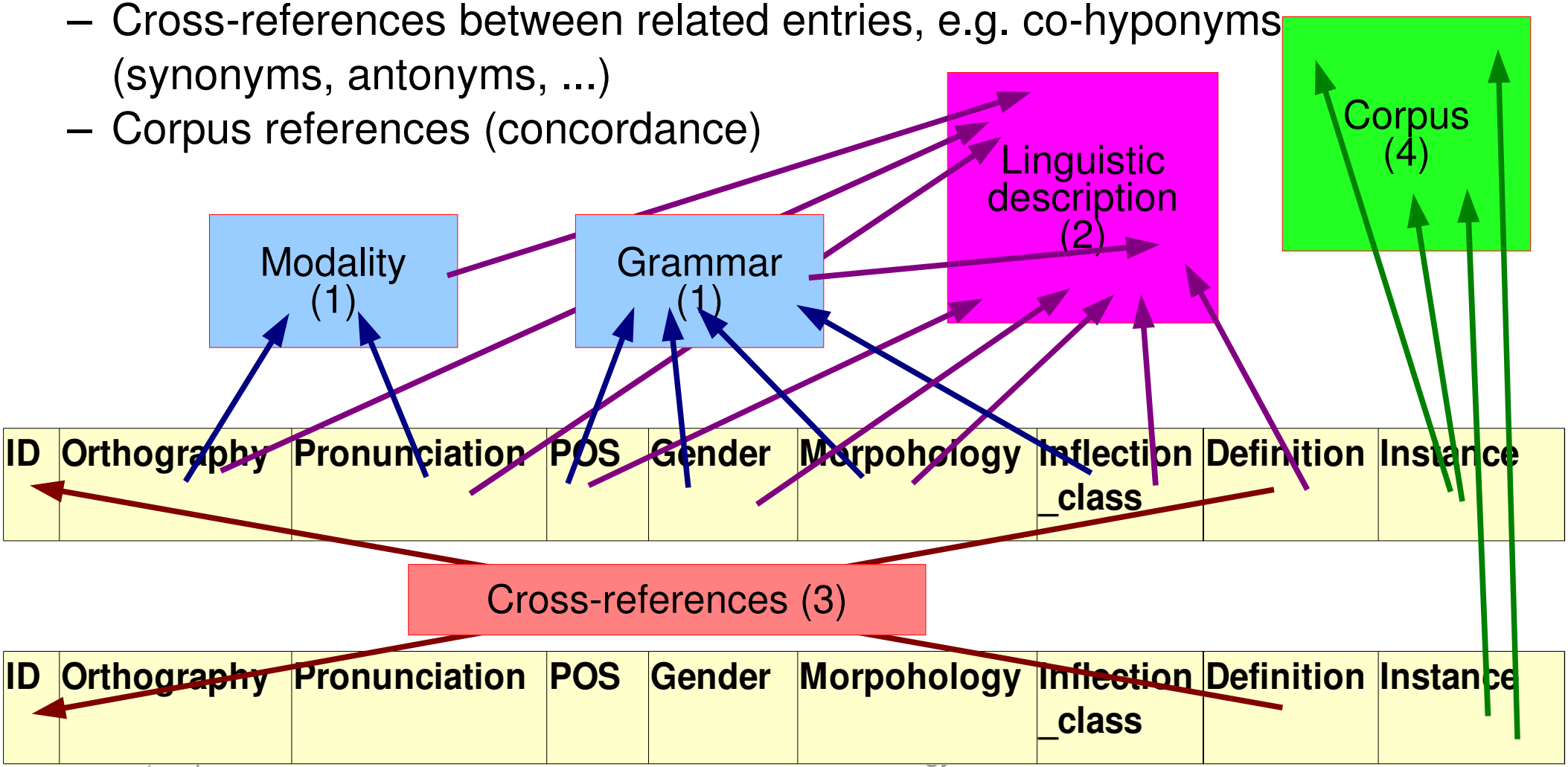


ID	Orthography	Pronunciation	POS	Gender	Morpohology	Inflection_class	Definition	Instance

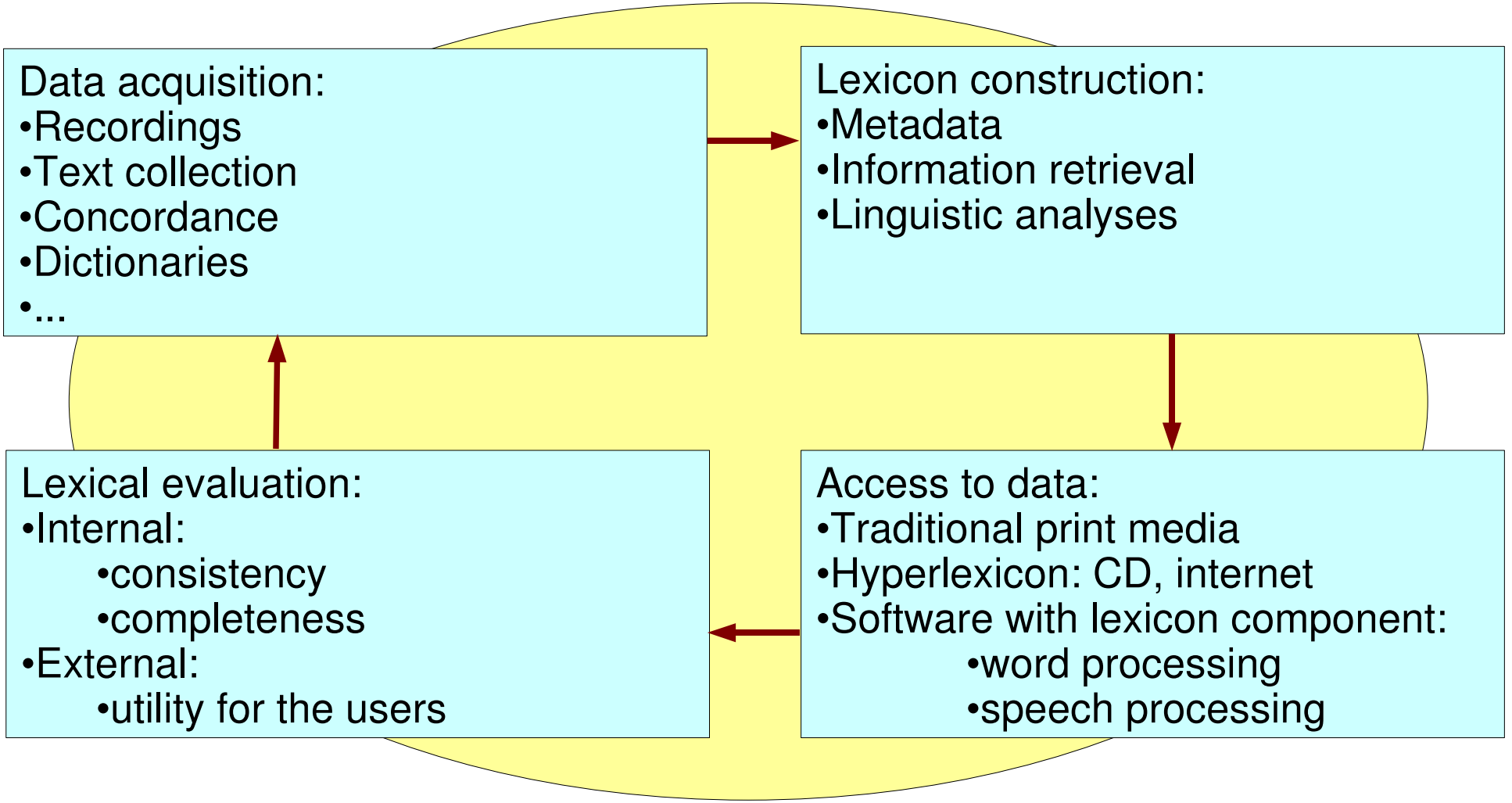
ID	Orthography	Pronunciation	POS	Gender	Morpohology	Inflection_class	Definition	Instance

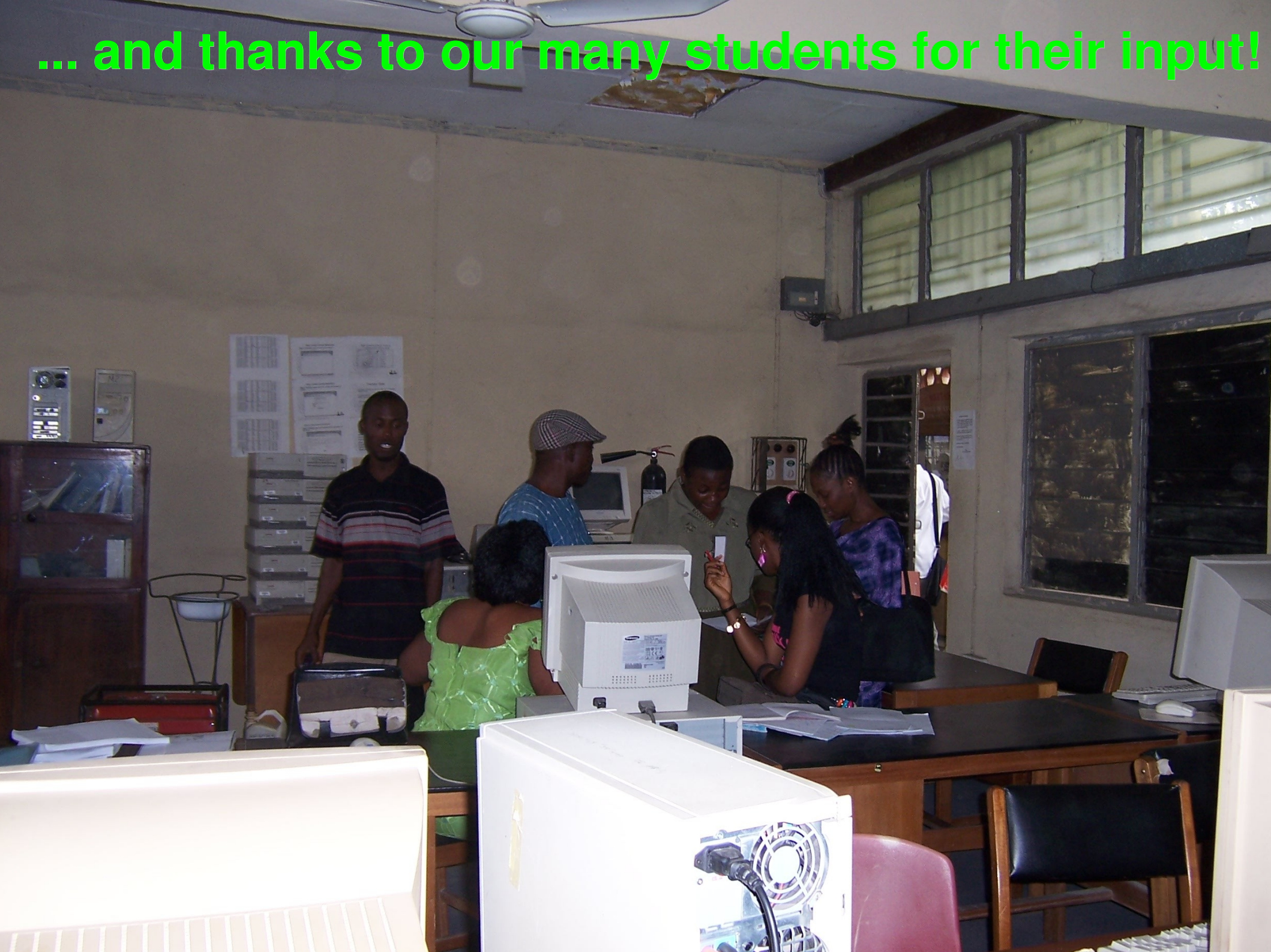
(8) Modelling conventions: layer 3 - mesostructure

- Linguistically motivated class hierarchy of DATCAT subvectors e.g. modality, grammar, object semantics
- Linguistic description references, e.g. use of abbreviations for parts of speech, characterisations of spelling
- Cross-references between related entries, e.g. co-hyponyms (synonyms, antonyms, ...)
- Corpus references (concordance)



(9) Modelling conventions: lexicographic workflow




... and thanks to our many students for their input!