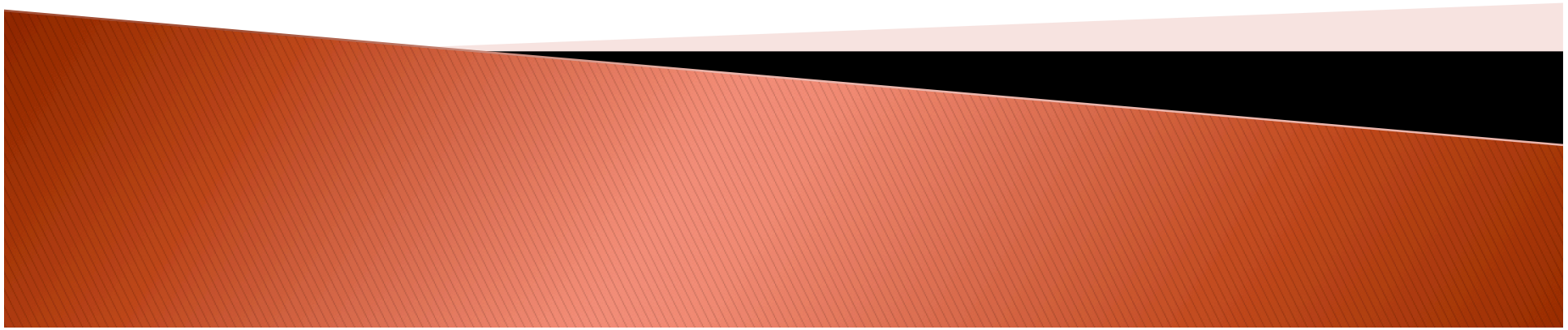
TYTO – A Collaborative Research Tool for Linked Linguistic Data

Andrea C. Schalley, Griffith University, Australia

LDL 2012, Frankfurt, 8 March 2012



Acknowledgments

- ARC Discovery Grant project DP0878126
  Social cognition and language: the design resources of grammatical diversity
- Project Members and Affiliates:

ANU:
Nicholas Evans
Alan Rumsey
Tom Honeyman
Stef Spronck
Aung Si
Darja Hoenigman
Anneliese Kuhle
Yusuf Sawaki

Griffith University:
Andrea Schalley
Alexander Borkowski

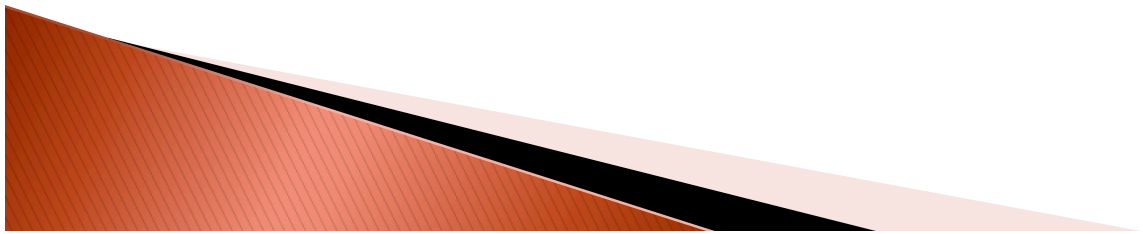
University of Melbourne:
Barbara Kelly
Murray Garde
Lauren Gawne
Sara Ciesielski

MPI Nijmegen:
Stephen Levinson
Nick Enfield
Lila San Roque

Stockholm University:
Henrik Bergquist

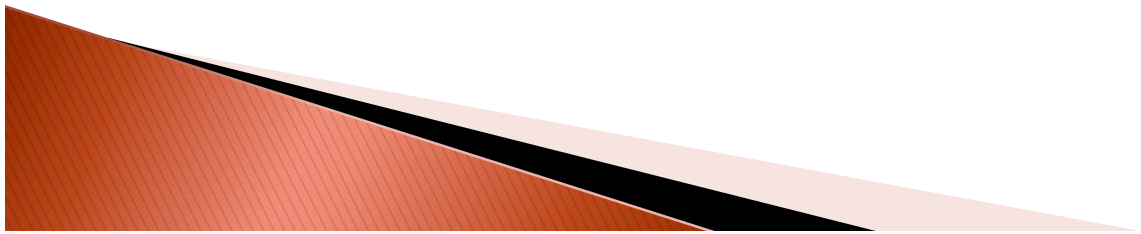
Outline

- Introduction
- Linked data in typology
- Related projects
- TYTO
- Conclusion



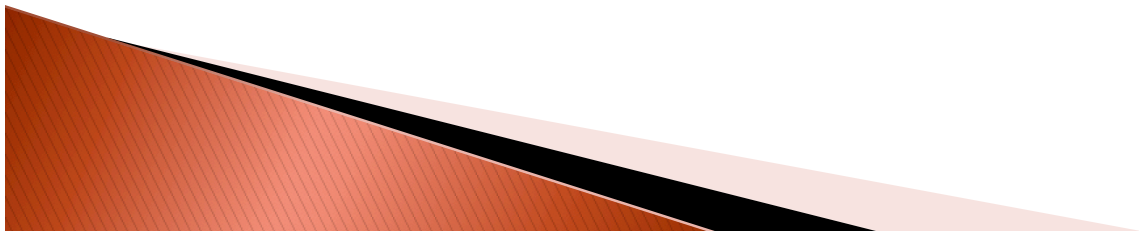
Introduction

- typology:
  - branch of linguistics
  - studies language from a comparative, cross-linguistic point of view
- pre-requisite for successful typological comparison:
  availability of reliable and readily accessible
  - data on specific languages
  - analyses of these data



Competency questions

- Which languages are know to have suffixes that express past tense? List them and provide an overall number.
- Is there any evidence for Language X marking categories of knowledge sources? Give all relevant examples of this language, and list the knowledge source categories as well as their morphological and constructional realisations.
- Which languages in North America are know to encode senior kin and ingroup (such as belonging to the same ethnic group) in a suffixal case marking system? Provide a list of the languages and outline where they are spoken.

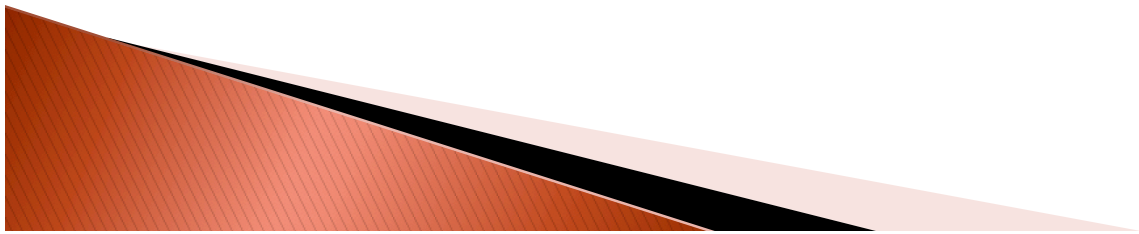


Linked data in typology

- Cross-linguistic data
  - comprehensive
  - form and meaning
  - raw data and analyses
- Grounding in linguistic examples
  - source of data
- Data analysis
  - reanalysis (correction and expansion; history)
  - fine-grained (dimensions of typological variation)

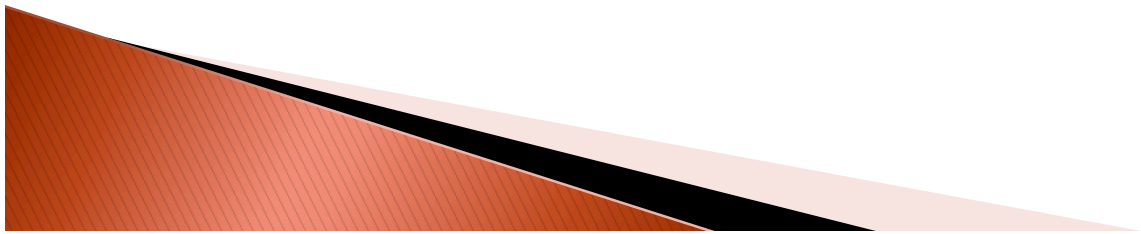
Linked data in typology

- Querying and reporting
  - highly targeted querying (cf. competency questions)
  - flexibility of accessing the data and their analyses
    - variation dimensions
    - representation format of reports
  - intuitive query formulation
- Scope
  - form and meaning (semasiological vs. onomasiological view)



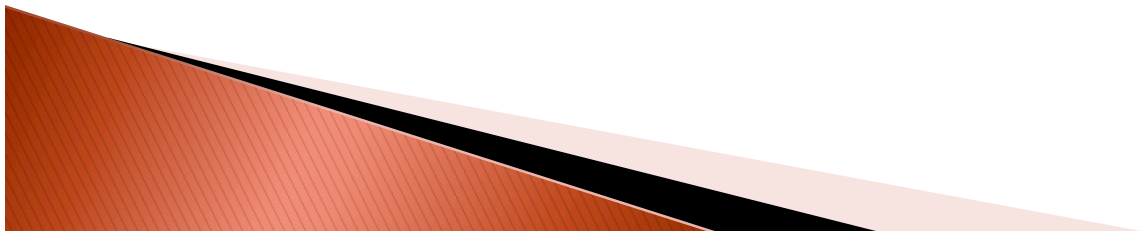
Linked data in typology

- Multi-user contributions (collaboration)
  - handling of diverse contributions at same or at different times
  - automatic integration of contributions
  - immediate access to submitted information as part of the system
- Fieldwork compatibility
  - local copy, independent of Internet
  - data entry in field
  - querying in field; generation of reports
  - fast automatic integration into central data store on return



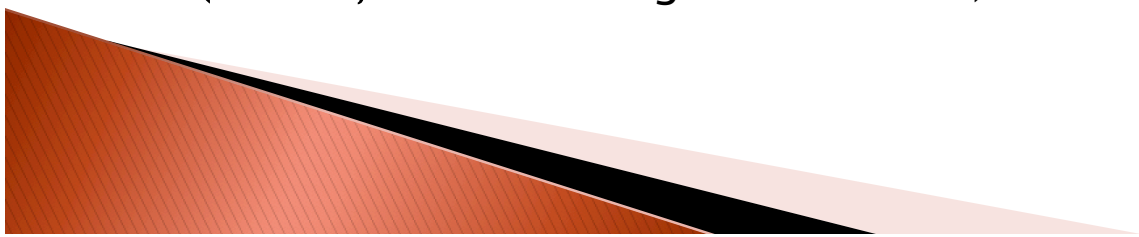
Linked data in typology

- Data entry
  - userfriendly, fast, efficient
  - automatic parsing of interlinear glossing
  - interfaces for non-anticipated data
- Expandability
  - new analytical concepts
  - terminological controversies catered for
  - positive and negative evidence



Related projects

- *Cross-linguistic Reference Grammar* (CRG) (Comrie et al. 1993; Zaefferer 2003, 2006)
- *The World Atlas of Language Structures* (WALS, Dryer & Haspelmath 2011)
- *Database of Syntactic Structures of the World's Languages* (SSWL, http://sswl.railsplayground.net/)
- *Galoes* (http://www.galoes.org/; Nordhoff, 2008)
- *Typological Database System* (TDS, Dimitriadis et al. 2009)
- *Generalized Ontology for Linguistic Description* (GOLD, Farrar & Langendoen 2003)



TYTO

(*Tyto alba*)

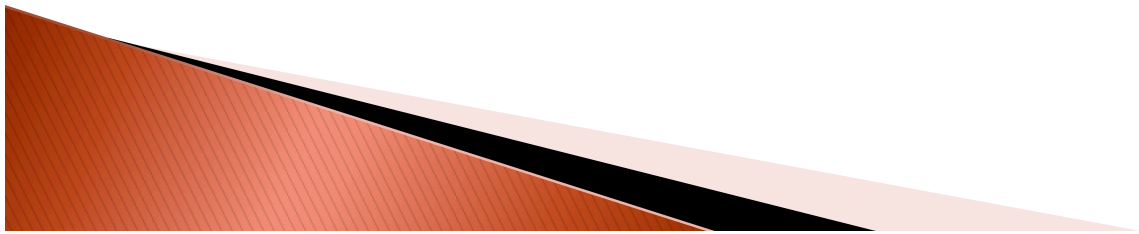
- typology tool
- ontology backbone
- data-driven
- input system
- querying
- reporting
- collaborative
- reasoner
- fieldwork
- revisions

TYTO – Linked data in typology

- Cross-linguistic data ✓
  - comprehensive ✗
  - form and meaning ✓
  - raw data and analyses ✓
- Grounding in linguistic examples ✓
  - source of data ✓
- Data analysis
  - reanalysis (correction and expansion; history) ✓
  - fine-grained (dimensions of typological variation) (✓)

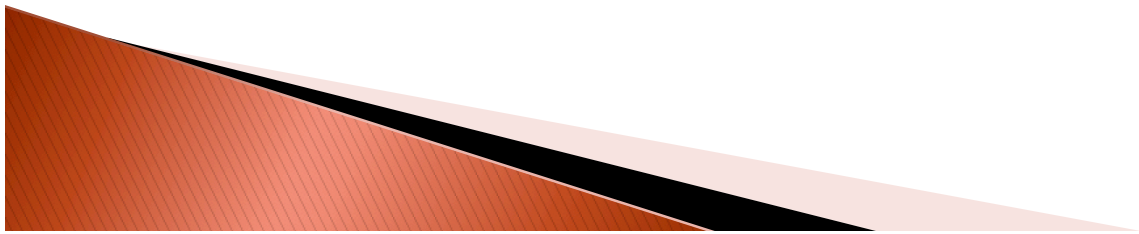
TYTO – Linked data in typology

- Querying and reporting
  - highly targeted querying (cf. competency questions) ✓
  - flexibility of accessing the data and their analyses ✓
    - variation dimensions ✓
    - representation format of reports ✓
  - intuitive query formulation (x)
- Scope
  - form and meaning (semasiological vs. onomasiological view) ✓



TYTO – Linked data in typology

- Multi-user contributions (collaboration)
  - handling of diverse contributions at same or at different times ✓
  - automatic integration of contributions (✓)
  - immediate access to submitted information as part of the system (✓)
- Fieldwork compatibility
  - local copy, independent of Internet ✓
  - data entry in field ✓
  - querying in field; generation of reports ✓
  - fast automatic integration into central data store on return (✓)

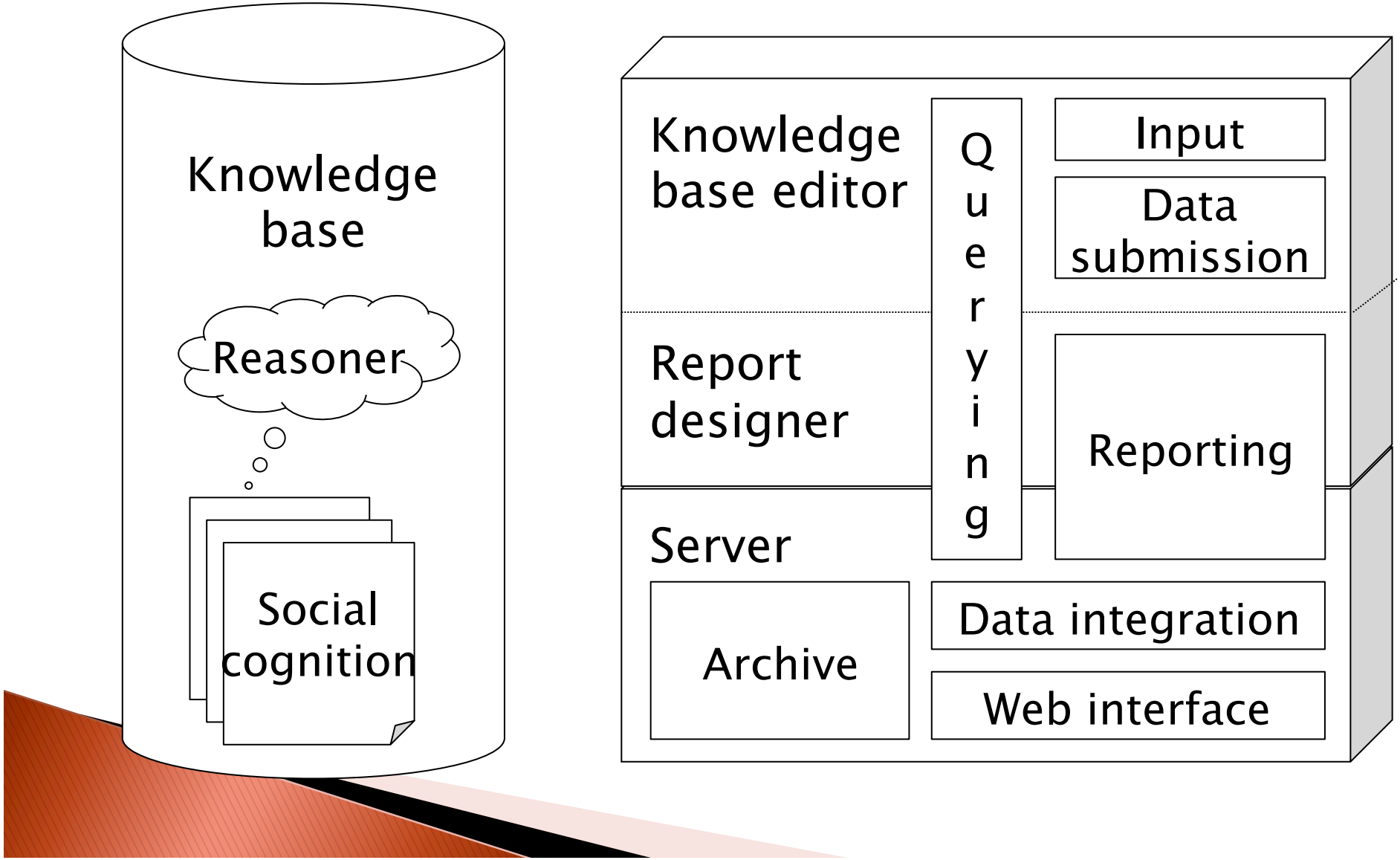


TYTO – Linked data in typology

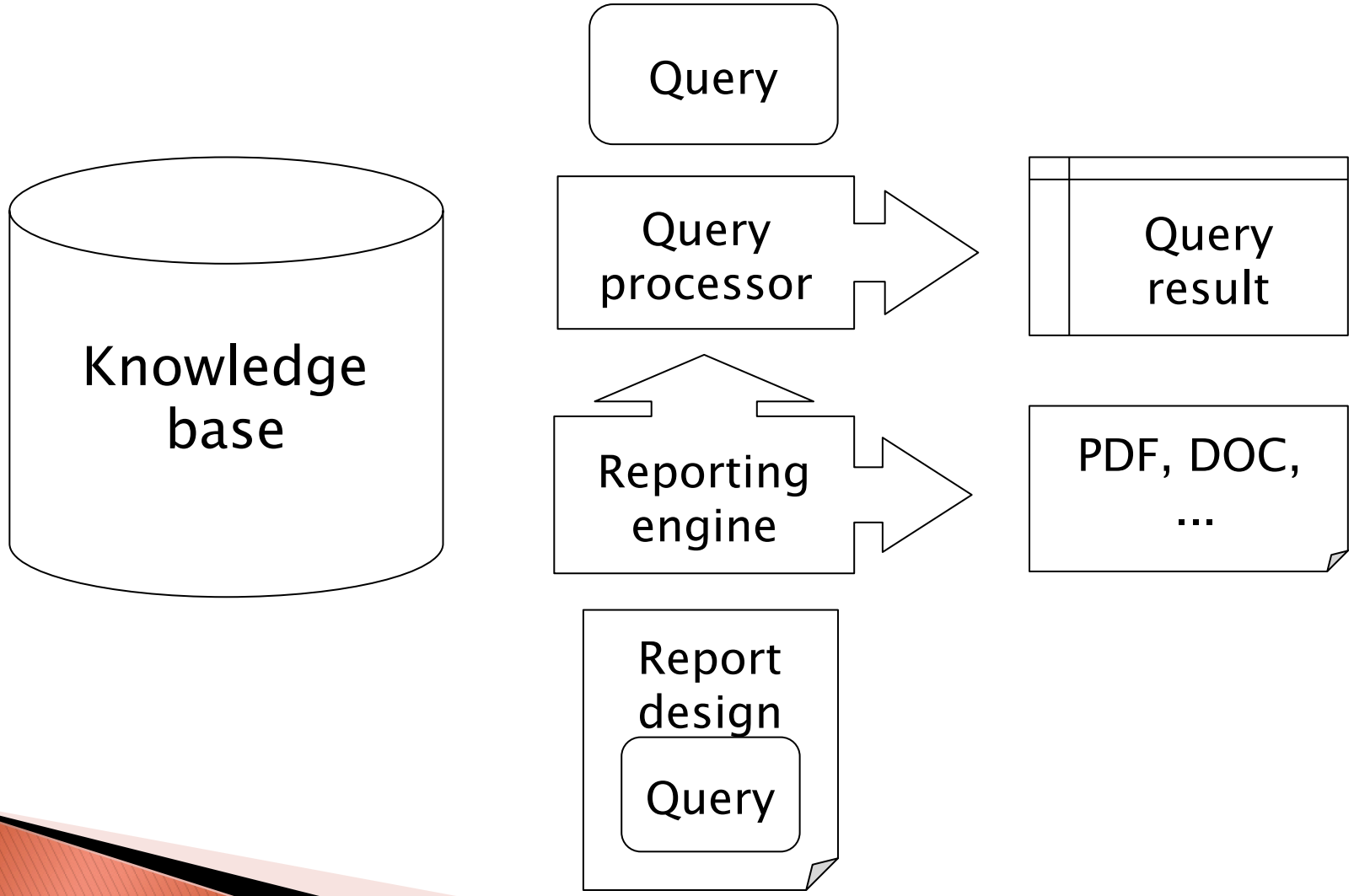
- Data entry
  - userfriendly, fast, efficient ✓
  - automatic parsing of interlinear glossing ✓
  - interfaces for non-anticipated data ✓
- Expandability
  - new analytical concepts ✓
  - terminological controversies catered for ??
  - positive and negative evidence

✓ (x)

TYTO – System overview

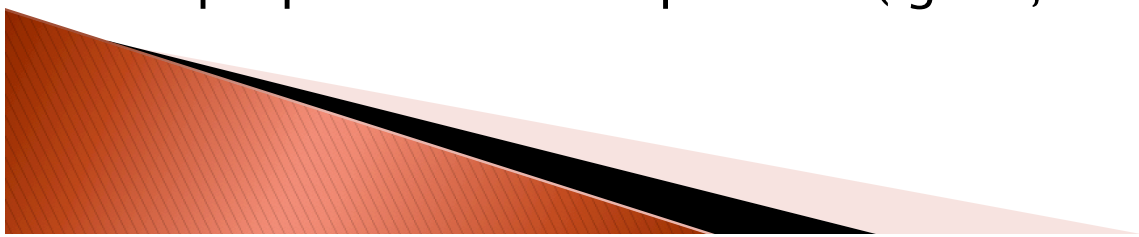


TYTO – System output



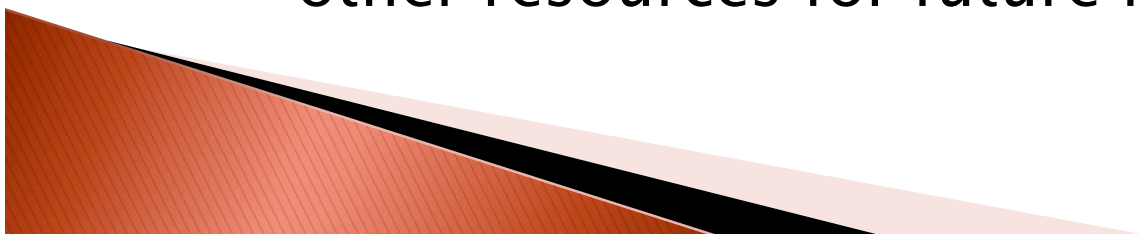
TYTO – Technologies

- URI, XML (and XML schemata) (ontology; example data, source information, and reports)
- RDF and OWL (ontology)
- SPARQL (query language)
- Apache Jena (Semantic Web framework)
- Protégé (ontology editor)
- Jena's rule reasoner (software reasoner)
- JasperReports (reporting engine)
- iReport (report designer)
- Mercurial (distributed version control system)
- purpose-built components ('glue', interfaces, data entry parser)



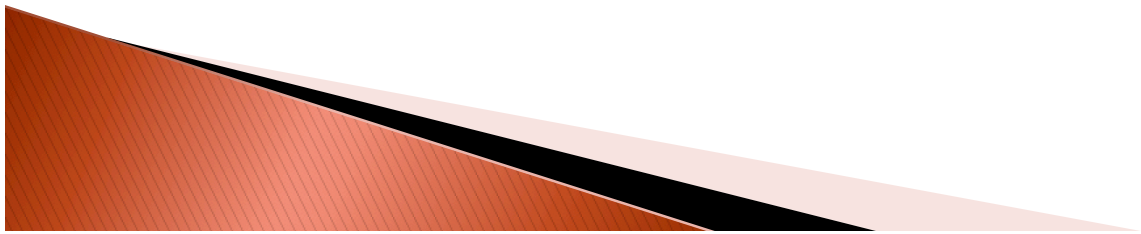
Conclusion

- four points that lie at the core of Linked Data [http://www.w3.org/DesignIssues/LinkedData.html]:
  1. URIs used as names for things ✓
  2. HTTP URIs used so that people can look up those names ✓
  3. Standards used (RDF, SPARQL) ✓
  4. Include links to other URIs, so that people can discover more things (✓)
     [so far only within tool, but plans for linking to other resources for future implementation]



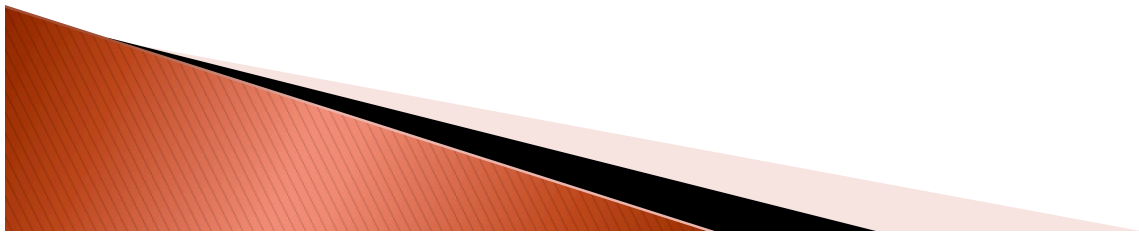
Conclusion

- 5-star ranking:
  - Make your data available on the Web under an open license ✓
  - Make it available as structured data ✓
  - Use a non-proprietary format ✓
  - Use linked data format ✓
  - Link your data to other people's data to provide context (×) [not yet]



Conclusion

- collaborative typology tool: tool to inform language comparison and linguistic theory building
- TYTO not intended to replace grammar writing
- modular tool, reusability of components
- major roadblocks:
  - terminological controversies
    (in particular: tension between single-language descriptors and cross-linguistic comparative concept)
  - establishment of trust (last layer in Semantic Web architecture), i.e. documentation of information source and assessing its reliability (this is closely connected to question of how such contributions can be counted as research output)



Thank you!