

Typology with graphs and matrices

Steven Moran,^{1,2} Martin Brümmer³ & Michael Cysouw¹

¹University of Marburg, ²University of Zurich, ³University of Leipzig

Goals

- Access federated linguistic databases through graphs
- Extract data for typological analyses
- Efficient computation through matrix data calculations

Talk map

- Overview of the technologies (the why)
- Overview of the LLOD (the how)
 - What is it?
 - What's in it?
 - How to access it
- What can you do with the data?

Graphs and matrices

- are two representations of data that encode the same thing
- table data (what we all know and use)
- matrix is purely numerical table data
- graph (mathematical sense)

Table data

observations	word class	last symbol
some	adjective	e
words	noun	s
as	preposition	s
example	noun	e

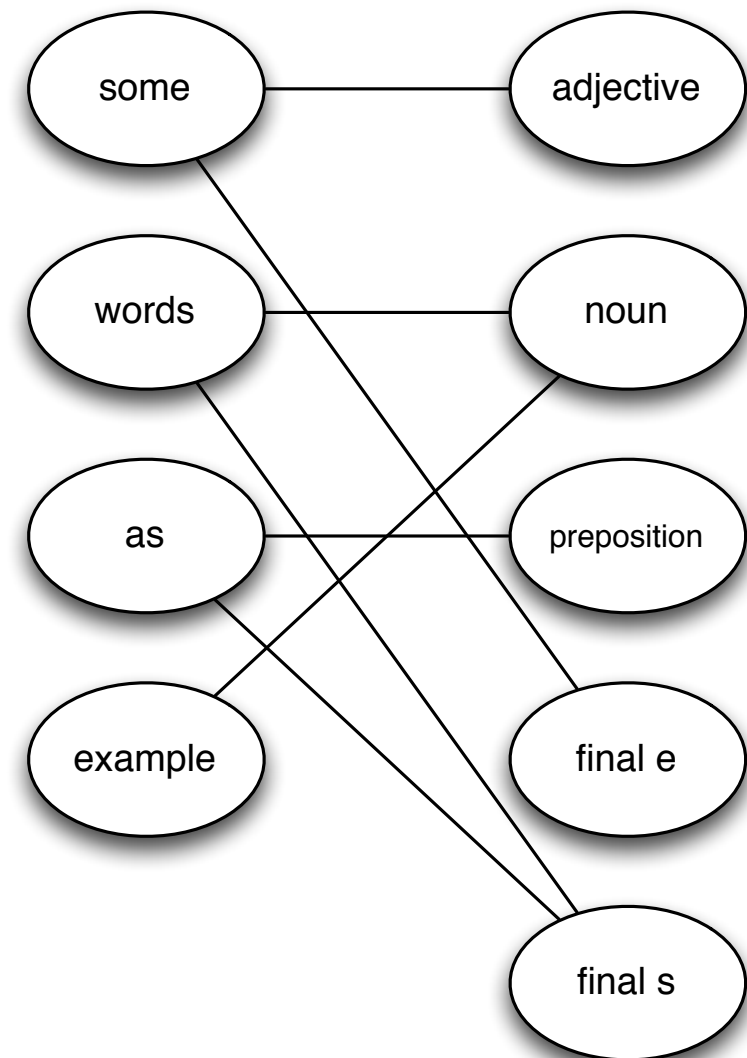
Matrix

observations	word class	last symbol
some	adjective	e
words	noun	s
as	preposition	s
example	noun	e

observations	adjective	noun	preposition	final e	final s
some	1	0	0	1	0
words	0	1	0	0	1
as	0	0	1	0	1
example	0	1	0	0	1

Graph

observations	adjective	noun	preposition	final e	final s
some	1	0	0	1	0
words	0	1	0	0	1
as	0	0	1	0	1
example	0	1	0	0	1



Linked Data

- “Linked Data” refers to Semantic Web framework practices for publishing and connecting structured data
 - uses a graph-based model for data interchange
 - encodes “knowledge” in statements encoded in subject-predicate-object triples
- syntactic and semantic interoperability
 - data aggregation, access and manipulation
 - consistent interpretation of exchanged data
 - dependent on common definitions and concepts in a vocabulary or ontology

Technological issues with Linked Data

- Anyone can say anything about anything
 - anyone can define their own naming conventions; devise their own models
- Open world assumption
 - the truth value of a statement is independent of whether or not it is known to be true
 - not knowing whether or not a statement is explicitly true, does not imply that the statement is false
- No unique names assumption
 - users cannot assume that any resources (concepts or relations) identified by URIs are actually different

Practical problems with Linked Data

- Difficult to deploy and maintain
- Accessing the underlying structures not so transparent
- Technology for federate queries still immature
- SPARQL query language involves learning
 - matches sets of triples patterns that match concepts and their relations by binding variables to match graph patterns
 - accessible through the browser via a “SPARQL Endpoint”

Linguistics Linked Open Data cloud (LLOD)

- Open Working Group in Linguistics
 - leading development and implementation of LLOD
- Data sources already in Linked Data
 - Glottolog
 - WALS, WOLD, IDS
 - PHOIBLE
 - etc...

Extracting table/matrix data from Linked Data graphs

- Give me all sources linked in the
- **select distinct ?graph**
where {GRAPH ?graph {?s ?p ?o}}
- results:
 - <http://mlode.nlp2rdf.org/resource/phoible/>
 - <http://quanthistling.info/lod/>
 - <http://mlode.nlp2rdf.org/resource/ids/>
 - <http://mlode.nlp2rdf.org/resource/wals/>
 - <http://wold.livingsources.org/>

Extracting table/matrix data from Linked Data graphs

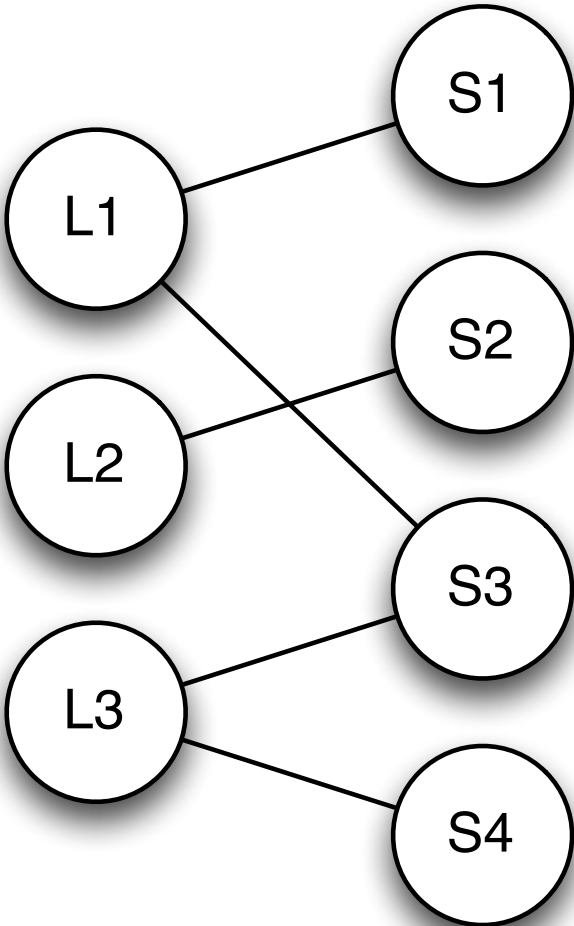
- Give me all data across resources for a given ISO 639-3 language name identifier

Example of extracting table/matrix data

- multivariate typology data
- this would be an example of querying features...
- “Don’t categorize languages into gross types, use fine-grained matrices.”

Extracting data from graph to matrix

- Return all data in PHOIBLE and WALS



L1	S1
L1	S3
L2	S2
L3	S3
L3	S4

	S1	S2	S3	S4
L1	1	0	1	0
L2	0	1	0	0
L3	0	0	1	1

Example of extracting table/matrix data

- phoible data
- wals data
- combined data

Testing the approach on WALS and PHOIBLE

- total 117.279 links between WALS codes and linguistic characteristics
- extracting data from the LLOD goes quick (STEVE: do you have any timing on the SPRQL query?)
- transformation into sparse matrix is only rewriting (very quick: most time lost in reading data)
- correlation all pairs of characteristics (3263x3263) via sparse matrix manipulation is quick (0.18 sec. on a MacBook Air)
- correction for genealogical relationship is no problem
- biggest problem: how to analyse such large correlation matrices!

Correlating WALS with PHOIBLE

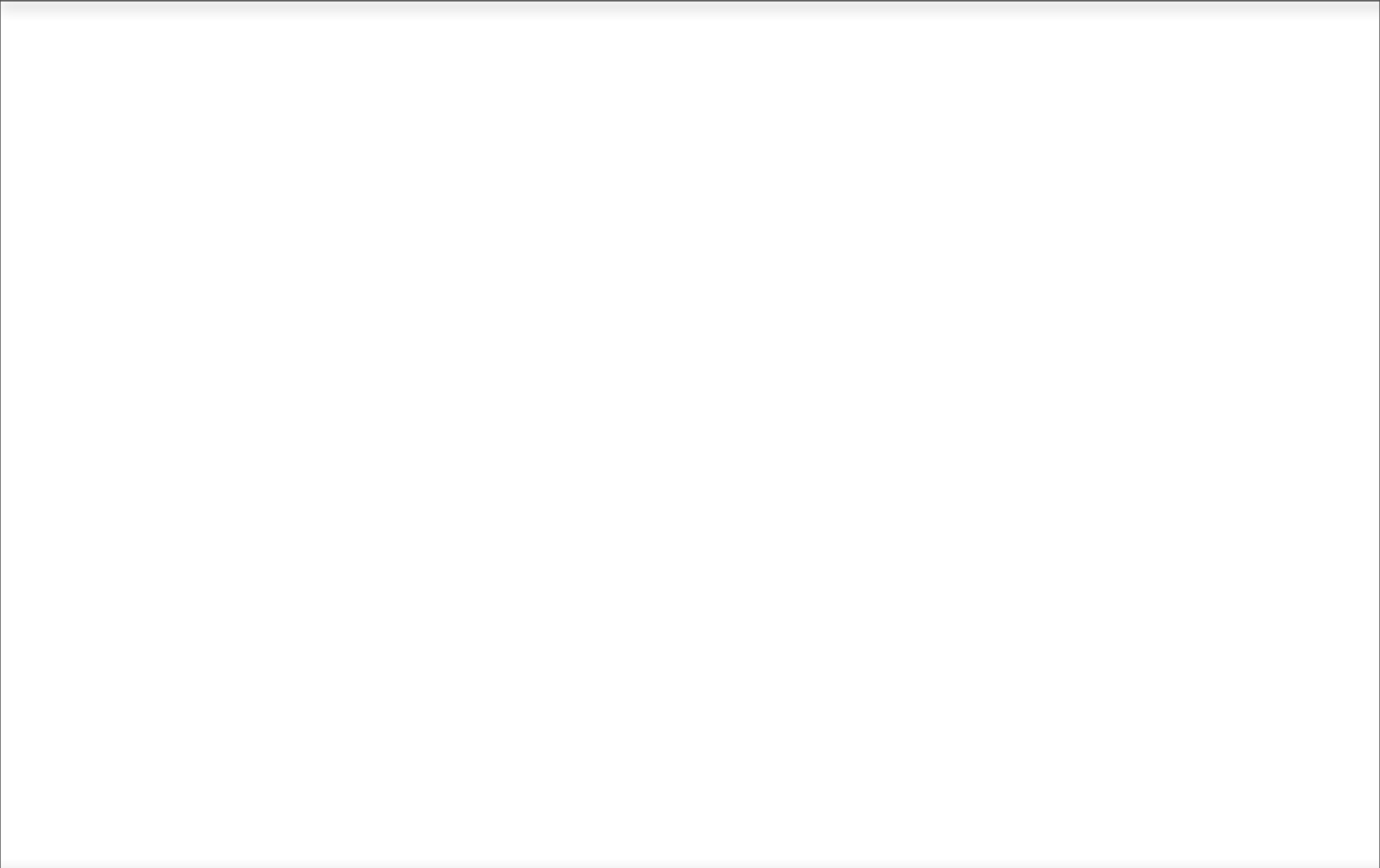
- Clustering results in a few interesting clusters across WALS and PHOIBLE:
- Cluster 42
 - WALS: f13A-3 (complex tone system)
 - PHOIBLE: / ɿ, ɿ /
- Cluster 47
 - WALS: f7A-2 (glottalized consonants, ejectives only)
 - PHOIBLE: / k', p', q', ts', tʃ' /
- Cluster 54
 - WALS: f10A-1 (vowel nasalization present)
 - PHOIBLE: / ã, ã, ã, ã, ã, ã , ã /

Heatmap for all characteristics with frequency more than 10 (~1000 characteristics)



Heatmap for languages with most data in WALS only

Heatmap for genera with most data in WALS only



Conclusion

- Data often starts life as a table
- Disparate table data can be converted into graphs
- Graphs can be combined into larger graphs with links between them
- Combined data graphs can be queried and data extracted into matrices
- Matrices are efficient for certain computations