Typology with graphs and matrices

Steven Moran, Michael Cysouw

Philipps University Marburg, University of Zurich
Deutschhausstrasse 3 35037 Marburg, Plattenstrasse 54 8032 Zürich
steven.moran@uzh.ch, cysouw@uni-marburg.de

Abstract

In this paper we show how the same data source can be represented in three different data formats – graphs, tables and matrices. After extracting table data from aggregated graphs data sources in the Linguistic Linked Open Data cloud, we convert these tables into numerical matrices to which we can apply mathematical formulations of linear algebra. As one example of the application of matrix algebra for language comparison, we identify clusters of association between disparate typological databases by leveraging the transformation of different data formats and Linked Data.

Keywords: linguistics, typology, language comparison

1. Introduction

In this paper we show how to access federated linguistic databases through Linked Data graphs so that we can extract data for typological analyses and then apply efficient computation for association measures through linear algebra calculations on matrix data to do language comparison. First we demonstrate how to leverage Semantic Web technologies to transform data in any number of typological databases, e.g. WALS (Haspelmath et al., 2008), AUTOTYP (Bickel and Nichols, 2002), PHOIBLE (Moran, 2012), ODIN (Lewis, 2006), or language-specific databases – along with metadata from Ethnologue (Lewis et al., 2013), LLMAP (LINGUIST List, 2009a), Multitree (LINGUIST List, 2009b) and Glottolog (Nordhoff et al., 2013) – into Linked Data. This is the vision of the Linguistic Linked Open Data Cloud (LLOD; (Chiarcos et al., 2012)).

Once data from these databases are converted into a homogeneous format, i.e. RDF graph data structures, the contents of these disparate datasets can be merged into one large graph, which allows for their data to be queried in a federated search fashion, in line with how we currently search the content of the Web through popular search engines. We illustrate how users can query and retrieve information about a particular language, from multiple databases, e.g. via a languages ISO 639-3 code. For example, a user might be interested in accessing all typological variables described by various databases for a particular language, e.g. word order data from WALS, genealogical information and phonological word domains from AUTOTYP, and phoneme inventory data from PHOIBLE.

Further, we show how the results of such queries can be combined and output into a matrix format that mirrors recent work in multivariate typology (cf. (Witzlack-Makarevich, 2011; Bickel, 2011a)). By outputting the results of users queries across different databases into table-based matrix formats, the results can be directly loaded into statistical packages for statistical analyses, and published algorithms can be directly applied to them and tested, e.g. statistical sampling procedures (cf. (Cysouw, 2005)) and statistical approaches to determine universal (language) preferences, e.g. Family Bias Theory (Bickel, 2011b). Furthermore, when typological data are output into tables, state-of-the-art approaches using linear algebra to transform matrices into new datasets can be applied (Mayer and Cysouw, 2012; Cysouw, In prep).

2. Graphs and matrices

Graphs and matrices are two representations of data that can encode the same things. We use the term *graph* in its mathematical sense, i.e. an ordered pair comprising of a set of vertices together with a set of edges, or in other words, a set of objects in which some objects are connected by links. By *table* data, we simply mean data in a table format. And by *matrix*, we mean purely numerical table data. Some illustrations will make these definitions clear.

Table 1 illustrates what we mean by table data; it provides a set of data, here observations about the last symbol in several words, where each word's class is also given.

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

Table 1: Table data

If we want to transform the table data in Table 1 into a matrix, we can use numerical values to indicate the presence or absence of features, as illustrated in Table 2.[1]

Table 2 can algorithmically be transformed into a graph by assigning the column labels as vertices and connecting them via edges for cells that a "1". The result of this transformation is illustrated in Figure 1.

[1] We provide the headers for convenience, but strictly speaking, a matrix in this work contains purely numerical data in a tabular structure.

observations	adj	noun	prep	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	1	0

Table 2: Matrix data

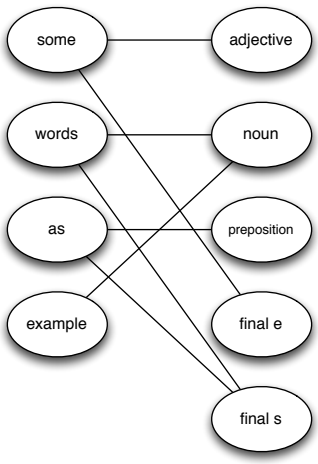

Figure 1: Matrix transformed into a graph

3. Connection to Linked Data

Linked Data refers to Semantic Web framework practices for publishing and connecting structured data.[2] Linked Data uses uses a graph-based model for data interchange, whereby data, specifically Web content, is connected using the Resource Description Framework (RDF), uniform resource identifiers (URIs) and content negoiation. Using graphs, anyone can describe "knowledge" in statements encoded in subject-predicate-object triples; a hypothetical example is given in Figure 2 of a concept "language" having several phonological "segment(s)".

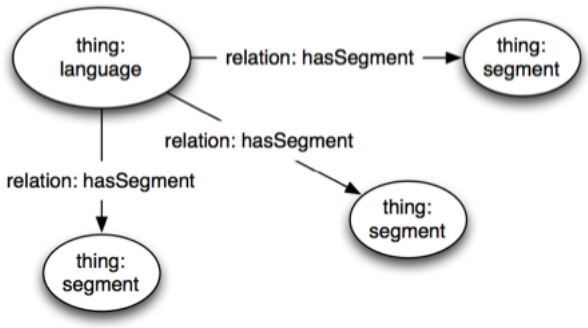

Figure 2: Linked Data example

The aims of a Semantic Web are to attain syntactic and semantic intoperability of data (cf. (Ide and Pustejovsky, 2010)). Syntactic interoperability means a consistent interpretation of exchanged data, which is achieved through graph data structures that allow for data access, aggregation and manipulation. Semantic interoperability is the ability to automatically interpret exchanged information meaningfully. Content must be unambiguously defined and is dependent on common definitions and concepts in a vocabulary or ontology. In this paper we are mainly concerned with syntactic interoperability for data aggregation and transformation.

There are several technological issues with Linked Data that are worth pointing out. First, anyone can say anything about anything, i.e. anyone can define their own naming conventions, devise their own models, etc. This is of course problematic when striving to attain semantic interoperability between resources. Another issue is the open world assumption that is built into the design of the Semantic Web. This assumption states that the truth value of a statement is independent of whether or not it is known to be true. Or in other words, not knowing whether or not a statement is explicitly true, does not imply that the statement is false. Although this stipulation is an important factor in attaining semantic interoperability of data sources, it is also directly relevant to academic research that uses Linked Data. Data as it currently stands in resources like the Linguistics Linked Open Data cloud (LLOD)[3] cloud must be problematically taken at face-value.

There are also practical problems with Linked Data, such as it is difficult to deploy, host and maintain. Furthermore, accessing the underlying structures is not necessarily transparent (i.e. most resources, say, in the LLOD are not published with information about their underlying data models). Technology to federate queries across endpoints is still immature, so that in reality Linked Data sets typically have to be hosted on the same server.[4]

Using an endpoint, such as one set up by the Open Working Group on Open Data in Linguistics (OWLG),[5] we can query data sources already in the LLOD, such as Glottolog, WALS, PHOIBLE, Wordnet, IDS, WOLD and Lexvo. By querying the LLOD via an endpoint, users can extract data from disparate but connected Linked Data graphs, to get information (metadata, typological data, etc), aggregated data (e.g. extract wordlists from different lexical sources such as WOLD, IDS and QuantHistLing) and to contrast data from different sources, e.g. published language geo-coordinates or language genealogical classifications.

Extracting information from Linked Data graphs is as simple as the query given in Example 1, which says 'show me all sources linked in the cloud'.[6] Some results of this query

[2] `http://linkeddata.org`

[3] `http://linguistics.okfn.org/files/2013/10/llod-colored-current.png`

[4] The SPARQL query language involves learning how to match sets of triple patterns that match concepts and their relations by binding variables to match graph patterns. An online query services can be made accessible through the browser via a SPARQL "endpoint".

[5] `http://linguistics.okfn.org/`

[6] This is a simplification because Linked Data federated queries do not yet work across disparately hosted data sources. As is, we query data sources *hosted* on a single server and acces-

are shown in Table 3.

1.
```
select distinct ?graph
where {GRAPH ?graph {?s ?p ?o}}
```

graph
http://wiktionary-en.dbpedia.org/
http://linked-data.org/resource/wals/
http://lexvo.org/
http://linked-data.org/resource/ids/
http://quanthistling.info/lod/
http://mlode.nlp2rdf.org/resource/ids/
http://mlode.nlp2rdf.org/resource/wals/
http://wold.livingsources.org/
http://example.org/
http://wiktionary.dbpedia.org/
http://lemon-model.net/

Table 3: Some results from simple query

Moving a step forward towards querying linguistic data, we can ask for all data sources linked the LLOD that have information for a given WALS code (as associated with an ISO 639-3 language name identifier) with the query given in Example 2 for WALS code chr (language name Chrau; ISO 639-3 crw). Some query results are given in Table 4.

2.
```
PREFIX wals:
<http://mlode.nlp2rdf.org/
resource/wals/language/>
PREFIX dcterms:
<http://purl.org/dc/terms/>
select distinct ?relation where {
wals:chr dcterms:relation ?relation . }
```

relation
llmap.org/maps/by-code/crw.html
ethnologue.com/show_language.asp?code=crw
en.wikipedia.org/wiki/ISO_639:crw
lexvo.org/data/iso639-3/crw
sil.org/iso639-3/documentation.asp?id=crw
multitree.org/codes/crw
scriptsource.org/lang/crw
language-archives.org/language/crw
odin.linguistlist.org/igt_urls.php?lang=crw
glottolog.org/resource/languoid/id/chra1242

Table 4: Some results from aggregated query

Digging deeper, we can extend this query so that we return all information for a given WALS code, as shown in Example 3. Example results are given in Table 5.

3.
```
PREFIX wals: <http://mlode.nlp2rdf.org/
resource/wals/language/>
PREFIX walsVocab: <http://mlode.nlp2rdf.org/
resource/wals/vocabulary/>
PREFIX wgs84: <http://www.w3.org/2003/01/geo/
wgs84_pos#>
PREFIX dcterms: <http://purl.org/dc/terms/>

select distinct ?label ?descr ?ref
?area ?lat ?long ?genus where {
?s dcterms:subject wals:chr .
?s walsVocab:hasValue ?value .
?value dcterms:description ?descr .
wals:chr wgs84:lat ?lat ;
         wgs84:long ?long ;
         ?feature ?datapoint ;
         rdfs:label ?label ;
         walsVocab:hasGenus ?genus ;
         walsVocab:altName ?name .

?datapoint dcterms:references ?ref .
?feature dcterms:isPartOf ?chapter .
?chapter walsVocab:chapterArea ?area .
}
```

The idea of federated queries across Linked Data graphs allows us to combine different data sources and not only aggregate the results, but to use information from different linked sources to filter results. In Example 4, we leverage the World Geodetic System (WGS) standard to query for language data within specific geographic coordinates, a common task and useful function in cross-linguistic investigations.

4.
```
prefix phoible:
<http://mlode.nlp2rdf.org/resource/phoible/>
prefix wgs84:
<http://www.w3.org/2003/01/geo/wgs84_pos#>
select distinct ?iso ?segRes where {
GRAPH <http://mlode.nlp2rdf.org/
resource/phoible/> {
?langRes phoible:hasSegment ?segRes;
phoible:iso639-3 ?iso;
wgs84:lat ?lat;
wgs84:long ?long.
FILTER(?lat < 12.57 && ?lat > -56.24 &&
?long > -81.57 && ?long < -34.15) }
}
```

This query returns data on information about phonological inventories, from the PHOIBLE database, for languages spoken in South America. Some results are illustrated in Table 6.

iso	segRes
teh	http://mlode.nlp2rdf.org/resource/phoible/segment/j
teh	http://mlode.nlp2rdf.org/resource/phoible/segment/a
teh	http://mlode.nlp2rdf.org/resource/phoible/segment/k
teh	http://mlode.nlp2rdf.org/resource/phoible/segment/o

Table 6: Some results from aggregated query

4. Extract and convert

We have demonstrated how to extract table data from the Linked Data graph. In Section 2. we explained how table

sible through an endpoint. In this work we use the endpoint hosted by Martin Brümmer: `linked-data.org/sparql`. There is a URL to access the LLOD's endpoint at `http://llod.info`, but again, hosting Linked Data sources and true federate query is difficult.

label	description	reference	area	lat	long	genus
Chrau	The language has no morphologically dedicated second-person imperatives at all	Thomas 1971	Verbal Categories	10.75	107.5	bahnaric
Chrau	Differentiation: one word denotes 'hand' and another, different word denotes 'finger' (or, very rarely, 'fingers')	Thomas 1971	Verbal Categories	10.75	107.5	bahnaric
Chrau	Identity: a single word denotes both 'hand' and 'arm'	Thomas 1971	Verbal Categories	10.75	107.5	bahnaric

Table 5: Some results from aggregated query

data can be transformed into numerical matrices. An illustration is given in Figure 3, which contrasts the graph, table and matrix formats.

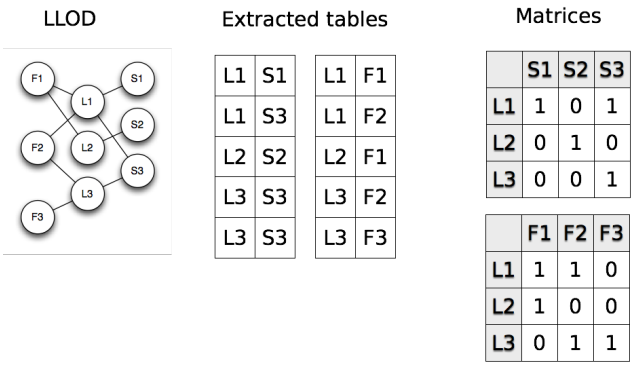

Figure 3: The caption of the figure.

Once graph data have been extracted into table format and transformed into numerical matrices, a straightforward transformation in statistical packages, matrix algebra calculations can be applied for the comparison of languages (Cysouw, In prep). One example of matrix manipulation is to take the dot product of two matrices, as illustrated in Figure 4. Here the matrix LS, Languages by Symbols, is multiplied by the linearly transformed matrix LF, i.e. languages by features, resulting in newly derived data in the segment by features matrix (SF).[7]

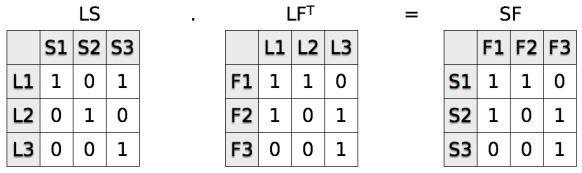

Figure 4: Dot product

The application of linear equations and linear transformations on matrices (vectors) has numerous applications across mainly fields. The reformulation of various research methods from the field of language comparison into matrix algebra highlights many parallels across methods and we believe it promises a deeper understanding of the methodology of language comparison in general. Additionally, the implementation of matrix algebra, as just one example from linear algebra, is highly efficient and fast. This makes computation on large datasets, like those that can be extracted from the LLOD, easier to manage and to perform. Furthermore, computations used in research projects can be straightforwardly written own in the form of formulas, which can both simplify instantiations in computer code as well as documentation of the research in published papers.

Using linear algebra on matrices, measures of association (similarity) can be computed. For association measures, we can compute the assocaition between all rows of, say, matrix A and matrix B by taking the dot product of the two. Depending on the form of normalization applied, we can for example take Pearson's correlation coefficient, with or without weighting, or we can calculate Cosine similarity (Cysouw, In prep). Identifying missing data, a substantial problem in linguistics, is also relatively easy using matrices (identify the gaps or "0"s) and matrix manipulations hold promise for adding data correction methods, such as normalization or estimating expected values by taking into account the distribution of missing information.

5. Testing the approach

To test our approach, we first extracted data from WALS and PHOIBLE from the LLOD. There are a total of 117,279 links between WALS codes and linguistic characteristics in PHOIBLE. Extraction from the LLOD goes quick – a few seconds with a good internet connection. Transformation from the extracted tables into sparse matrices is also very fast.[8] Correlation of all pairs of characteristics (3263x3263) via sparse matrix manipulation is extremely fast (0.18 sec. on a MacBook Air). Correction for genealogical relationship is also no problem. The biggest problem we encounter is how to analyze such large correlation matrices!

We decided to identify major clusters of association between WALS and PHOIBLE. Using Pearson's correlation coefficient, we identify levels of high association between clusters of features in WALS chapters and PHOIBLE phonological inventory data. These are visualized as heatmaps in Figures 5, 6 and 7.

What we find are several clusters between data in WALS chapters (grouped into buckets) and sets of segments from

[7] Here we use superscript $< ^t >$ to denote the transformed matrix, LF.

[8] We use R for the conversion and most of the time is spent reading in the data.

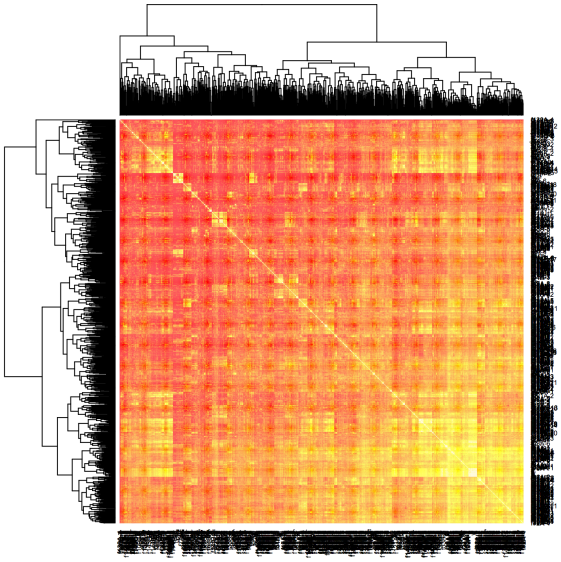

Figure 5: Heatmap for all characteristics with frequency more than 10 (˜1000 characteristics)

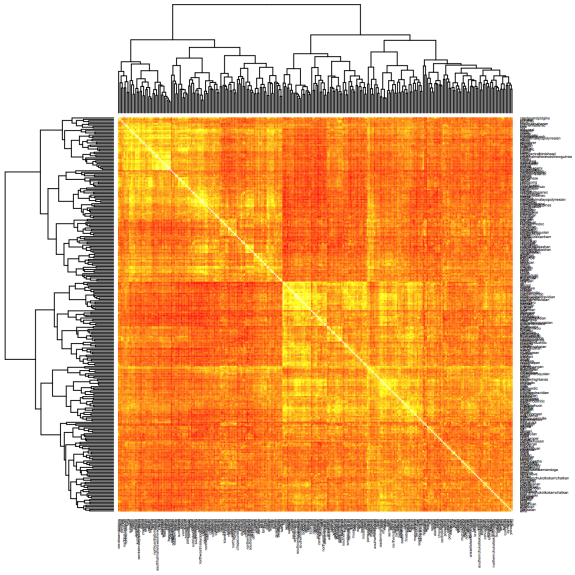

Figure 7: Heatmap for genera with most data in WALS only

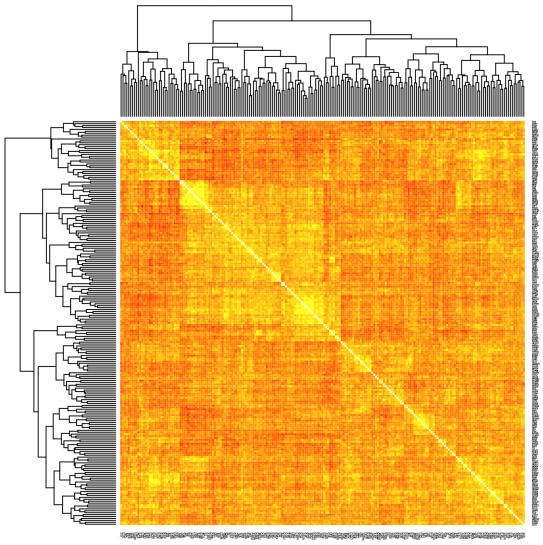

Figure 6: Heatmap for languages with most data in WALS only

cross-linguistic phonological inventory data in PHOIBLE. The first is the association between WALS feature 13A-3 (complex tone system) and high and low tone segments (a subset of tones) found in PHOIBLE's 1600+ languages. In another highly associated cluster, WALS feature 7A-2 (glottalized consonants, ejectives only) corresponds with languages in PHOIBLE that contain ejective segments /k', p', q', ts', tʃ'/. Our approach also identifies similarity between WALS feature 10A-1 (vowel nasalization present) and the set of languages in PHOIBLE that contain the cardinal nasalized vowels /ã, ẽ, ɛ̃, ĩ, õ, ɔ̃ , ũ/.

This is just a simple demonstration of identifying association using Pearson's correlation coefficient between two richly-annotated typological databases. One can imagine expanding the search for associations across other data sources, and even more exciting, apply the wealth of possibilities afforded by matrix algebra for language comparison, such as normalization of entities to be compared, the application of other measures of association and correction for missing data through evaluation of expected and observed results.

6. Conclusion

We have shown that the same data source can be represented in different data structures. Linguistic data often starts its life stored in tables, e.g. database tables. Table data can be converted into mathematical graphs, which can be used to overcome the problem of syntactic interoperability for data aggregation. Linked Data is the classic example. Linked Data graphs can be combined into larger graphs with links between them, thus enhancing data aggregation. In this paper we have illustrated how combined data graphs in the form of the LLOD can be queried and how data can be extracted and transformed into matrices. Matrix data gives us a data format to leverage mathematic formulations of linear algebra, the surface of which we have only scratched in this paper. We provide a simple example of how to manipulate data and to find clusters of association in combined datasets for research in language comparison and typology.

7. References

Balthasar Bickel and Johanna Nichols. 2002. Autotypologizing databases and their use in fieldwork. In *Proceedings of the LREC 2002 Workshop on Resources and Tools in Field Linguistics*.

Balthasar Bickel. 2011a. Grammatical relations typology. In J J Song, editor, *The Oxford Handbook of Language Typology*, Oxford Handbooks in Linguistics, pages 399 – 444. Oxford University Press, Oxford.

Balthasar Bickel. 2011b. Statistical modeling of language universals. *Linguistic Typology*, 15(2):401–413.

Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann. 2012. *Linked Data in Linguistics*. Springer.

Michael Cysouw. 2005. Quantitative Methods in Typology. In Gabriel Altmann, Reinhard Köhler, and Rajmund G. Piotrowski, editors, *Quantitative Linguistics: An International Handbook*, pages 554–578. Berlin: Walter de Gruyter.

Michael Cysouw. In prep. Matrix algebra for language comparison. In preperation.

Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie. 2008. The world atlas of language structures online. Munich: Max Planck Digital Library. Available online at http://wals.info/.

Nancy Ide and James Pustejovsky. 2010. What does interoperability mean, anyway? toward an operational definition of interoperability for language technology. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources. Hong Kong, China*.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig, editors. 2013. *Ethnologue: Languages of the world*, volume 17. SIL International, Dallas, Texas.

William D Lewis. 2006. Odin: A model for adapting and enriching legacy infrastructure. In *e-Science and Grid Computing, 2006. e-Science'06. Second IEEE International Conference on*, pages 137–137. IEEE.

LINGUIST List. 2009a. Ll-map: Language and location - map accessibility project. Online: `http://llmap.org/`.

LINGUIST List. 2009b. Multitree: A Digital Library of Language Relationships. Institute for Language Information and Technology (LINGUIST List), Eastern Michigan University. Ypsilanti, MI. Online: `http://multitree.org/`.

Thomas Mayer and Michael Cysouw. 2012. Language comparison through sparse multilingual word alignment. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 54–62. Association for Computational Linguistics.

Steven Moran. 2012. *Phonetics information base and lexicon*. Ph.D. thesis, University of Washington.

Harald Nordhoff, Sebastian andHammarström, Robert Forkel, and Martin (eds.) Haspelmath. 2013. Glottolog 2.2. leipzig: Max planck institute for evolutionary anthropology. Available online at http://glottolog.org.

Alena Witzlack-Makarevich. 2011. *Typological variation in grammatical relations*. Ph.D. thesis, University of Leipzig.