Linguistic Linked Open Data
From collection to application
(for under-resourced languages)

Christian Chiarcos
chiarcos@uni-frankfurt.de

Steven Moran
steven.moran@uzh.ch
Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era

- Collaboration
- Sustainability
- Publication and maintenance
- Benefits of Linked Data
Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era

- Collaboration
- Sustainability
- Publication and maintenance

Benefits of Linked Data

- How can research on underresourced languages benefit from Semantic Web technologies, and specifically the Linked Data framework?
Defining under-resourced Languages

- Lack of access to **language data**
  - General lack of language documentation, e.g., dictionaries
    - e.g., Chalkan (Turkic, Altay, 1180 speakers)
Defining under-resourced Languages

- Lack of access to **language data**
  - General lack of language documentation, e.g., dictionaries

- Lack of access to **digital** language data
  - Standardized orthography & encoding (ASCII, KOI-8, SAMPA)
  - Web resources (Wikipedia, Wiktionary, ...)
    - e.g., Shor (Turkic, Siberia, 2800 speakers)
Defining under-resourced Languages

- Lack of access to **language data**
  - General lack of language documentation, e.g., dictionaries

- Lack of access to **digital** language data
  - Standardized orthography & encoding (ASCII, KOI-8, SAMPA)
  - Web resources (Wikipedia, Wiktionary, ...)

- Lack of **IT/NLP support**
  - Localized text processing software
  - Basic Language Resource Kit (http://www.blark.org/)
    - e.g., Hausa [2010] (Chadic, West Africa, 34-53 mio speakers)
Defining under-resourced Languages

- Lack of access to **language data**
  - General lack of language documentation, e.g., dictionaries

- Lack of access to **digital** language data
  - Standardized orthography & encoding (ASCII, KOI-8, SAMPA)
  - Web resources (Wikipedia, Wiktionary, ...)

- Lack of **IT/NLP support**
  - Localized text processing software
  - Basic Language Resource Kit (http://www.blark.org/)

- Limited **interoperability** of data and tools
  - tools & annotations use different formats and conventions
    - e.g., Russian [2005] (Slavic, Eurasia, 150 mio speakers)
Linked Data & under-resourced Languages

- Linked Data
  - rules of best practice for publishing data on the web
    - protocols and standards
    - links between data sets
Linked Data & under-resourced Languages

- Linked Data
  - rules of best practice for publishing data on the web
  
=> Information integration

- Structural interoperability
  - comparable formats and protocols to access data
  
=> use the same query language for different data sets
Linked Data & under-resourced Languages

- Linked Data
  - rules of best practice for publishing data on the web

=> Information integration

- Structural interoperability
- Conceptual interoperability
  - develop and (re-)use a shared vocabularies for equivalent concepts

=> the same query on different data sets
Linked Data & under-resourced Languages

- Linked Data
  - rules of best practice for publishing data on the web

=> Information integration

- Structural interoperability
- Conceptual interoperability
- Federation
  - data published on the web
    - under an open license
    - with a query interface (SPARQL end point)

=> use a single query to query different datasets
Linked Data & under-resourced Languages

- Linked Data
  - rules of best practice for publishing data on the web

=> Information integration
  - Structural interoperability
  - Conceptual interoperability
  - Federation

Now: Non-technical intro to Linked Data
Later: How does this help under-resourced languages?
Linked Data

A non-technical introduction
From Tables to RDF to Linked Data

- PHOnetics Information Base and LExicon (PHOIBLE)

- Phoneme inventories and phonological features
  - Covers ~20% of the world’s spoken languages
  - Compiled from various sources, originally a flat table (list)
From Tables ...

<table>
<thead>
<tr>
<th>Source</th>
<th>id</th>
<th>ISO639-3</th>
<th>trump</th>
<th>root</th>
<th>wals_genus</th>
<th>population</th>
<th>latitude</th>
<th>longitude</th>
<th>phoneme_id</th>
<th>glyph_id</th>
<th>glyph</th>
<th>class</th>
<th>comb</th>
<th>num</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPA</td>
<td>1</td>
<td>kor</td>
<td>1</td>
<td>asis</td>
<td>Korean</td>
<td>42,000,000</td>
<td>37:30</td>
<td>128:0</td>
<td>1</td>
<td>1</td>
<td>ñ</td>
<td>cons</td>
<td>c-d-c-c</td>
<td>4</td>
</tr>
<tr>
<td>SPA</td>
<td>3</td>
<td>lbe</td>
<td>1</td>
<td>ncau</td>
<td>Lak-Dargwa</td>
<td>157,000</td>
<td>42:0</td>
<td>47:0</td>
<td>124</td>
<td>1</td>
<td>ñ</td>
<td>cons</td>
<td>c-d-c-c</td>
<td>4</td>
</tr>
<tr>
<td>SPA</td>
<td>5</td>
<td>kat</td>
<td>1</td>
<td>kart</td>
<td>Kartvelian</td>
<td>3,900,000</td>
<td>42:0</td>
<td>44:0</td>
<td>203</td>
<td>1</td>
<td>ñ</td>
<td>cons</td>
<td>c-d-c-c</td>
<td>4</td>
</tr>
<tr>
<td>SPA</td>
<td>6</td>
<td>bsk</td>
<td>1</td>
<td>asis</td>
<td>Burushaski</td>
<td>87,000</td>
<td>36:30</td>
<td>74:30</td>
<td>240</td>
<td>1</td>
<td>ñ</td>
<td>cons</td>
<td>c-d-c-c</td>
<td>4</td>
</tr>
<tr>
<td>SPA</td>
<td>14</td>
<td>khm</td>
<td>1</td>
<td>ausa</td>
<td>Khmer</td>
<td>12,300,000</td>
<td>12:30</td>
<td>105:0</td>
<td>632</td>
<td>19</td>
<td><em>MIC</em></td>
<td>vowel</td>
<td>v-d</td>
<td>2</td>
</tr>
<tr>
<td>SPA</td>
<td>27</td>
<td>tha</td>
<td>1</td>
<td>taik</td>
<td>Kam-Tai</td>
<td>20,200,000</td>
<td>15:00</td>
<td>100:40</td>
<td>1150</td>
<td>19</td>
<td>OMIC_</td>
<td>vowel</td>
<td>v-d</td>
<td>2</td>
</tr>
<tr>
<td>Source</td>
<td>id</td>
<td>ISO639-3</td>
<td>trump</td>
<td>root</td>
<td>wals_genus</td>
<td>population</td>
<td>latitude</td>
<td>longitude</td>
<td>phoneme.id</td>
<td>glyph_id</td>
<td>glyph</td>
<td>class</td>
<td>comb</td>
<td>num</td>
</tr>
<tr>
<td>--------</td>
<td>----</td>
<td>----------</td>
<td>-------</td>
<td>------</td>
<td>------------</td>
<td>------------</td>
<td>----------</td>
<td>-----------</td>
<td>------------</td>
<td>----------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>SPA</td>
<td>1</td>
<td>kor</td>
<td>1</td>
<td>asis</td>
<td>Korean</td>
<td>42,000,000</td>
<td>37:30</td>
<td>128:0</td>
<td>1</td>
<td>1</td>
<td>t</td>
<td>r</td>
<td>cons</td>
<td>c-d-c-c</td>
</tr>
<tr>
<td>SPA</td>
<td>3</td>
<td>lbe</td>
<td>1</td>
<td>ncau</td>
<td>Lak-Dargwa</td>
<td>157,000</td>
<td>42:0</td>
<td>47:0</td>
<td>124</td>
<td>1</td>
<td>t</td>
<td>r</td>
<td>cons</td>
<td>c-d-c-c</td>
</tr>
<tr>
<td>SPA</td>
<td>5</td>
<td>kat</td>
<td>1</td>
<td>kart</td>
<td>Kartvelian</td>
<td>3,900,000</td>
<td>42:0</td>
<td>44:0</td>
<td>203</td>
<td>1</td>
<td>t</td>
<td>r</td>
<td>cons</td>
<td>c-d-c-c</td>
</tr>
<tr>
<td>SPA</td>
<td>6</td>
<td>bsk</td>
<td>1</td>
<td>asis</td>
<td>Burushaski</td>
<td>87,000</td>
<td>36:30</td>
<td>74:30</td>
<td>240</td>
<td>1</td>
<td>t</td>
<td>r</td>
<td>cons</td>
<td>c-d-c-c</td>
</tr>
<tr>
<td>SPA</td>
<td>14</td>
<td>khm</td>
<td>1</td>
<td>ausa</td>
<td>Khmer</td>
<td>12,300,000</td>
<td>12:30</td>
<td>105:0</td>
<td>632</td>
<td>19</td>
<td>u:</td>
<td>vowel</td>
<td>v-d</td>
<td>2</td>
</tr>
<tr>
<td>SPA</td>
<td>27</td>
<td>tha</td>
<td>1</td>
<td>taik</td>
<td>Kam-Tai</td>
<td>20,200,000</td>
<td>15:00</td>
<td>100:40</td>
<td>1150</td>
<td>19</td>
<td>u:</td>
<td>vowel</td>
<td>v-d</td>
<td>2</td>
</tr>
</tbody>
</table>

Subject
(primary key)
From Tables to RDF ...

<table>
<thead>
<tr>
<th>Source</th>
<th>id</th>
<th>ISO639-3</th>
<th>trump</th>
<th>root</th>
<th>wals_genus</th>
<th>population</th>
<th>latitude</th>
<th>longitude</th>
<th>phoneme_id</th>
<th>glyph_id</th>
<th>glyph</th>
<th>class</th>
<th>comb</th>
<th>num</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPA</td>
<td>1</td>
<td>kor</td>
<td>1</td>
<td>asis</td>
<td>Korean</td>
<td>42,000,000</td>
<td>37:30</td>
<td>128:0</td>
<td>1</td>
<td>1</td>
<td>тн</td>
<td>cons</td>
<td>c-d-c-c</td>
<td>4</td>
</tr>
<tr>
<td>SPA</td>
<td>3</td>
<td>lbe</td>
<td>1</td>
<td>ncau</td>
<td>Lak-Dargwa</td>
<td>157,000</td>
<td>42:0</td>
<td>47:0</td>
<td>124</td>
<td>1</td>
<td>тн</td>
<td>cons</td>
<td>c-d-c-c</td>
<td>4</td>
</tr>
<tr>
<td>SPA</td>
<td>5</td>
<td>kat</td>
<td>1</td>
<td>kart</td>
<td>Kartvelian</td>
<td>3,900,000</td>
<td>42:0</td>
<td>44:0</td>
<td>203</td>
<td>1</td>
<td>тн</td>
<td>cons</td>
<td>c-d-c-c</td>
<td>4</td>
</tr>
<tr>
<td>SPA</td>
<td>6</td>
<td>bsk</td>
<td>1</td>
<td>asis</td>
<td>Burushaski</td>
<td>87,000</td>
<td>36:30</td>
<td>74:30</td>
<td>240</td>
<td>1</td>
<td>тн</td>
<td>cons</td>
<td>c-d-c-c</td>
<td>4</td>
</tr>
<tr>
<td>SPA</td>
<td>14</td>
<td>khm</td>
<td>1</td>
<td>ausa</td>
<td>Khmer</td>
<td>12,300,000</td>
<td>12:30</td>
<td>105:0</td>
<td>632</td>
<td>19</td>
<td>у</td>
<td>vowel</td>
<td>v-d</td>
<td>2</td>
</tr>
<tr>
<td>SPA</td>
<td>27</td>
<td>tha</td>
<td>1</td>
<td>taik</td>
<td>Kam-Tai</td>
<td>20,200,000</td>
<td>15:00</td>
<td>100:40</td>
<td>1150</td>
<td>19</td>
<td>у</td>
<td>vowel</td>
<td>v-d</td>
<td>2</td>
</tr>
</tbody>
</table>

Subject

Property ("Relation")
From Tables to RDF ...

<table>
<thead>
<tr>
<th>Source</th>
<th>id</th>
<th>ISO639-3</th>
<th>trump</th>
<th>root</th>
<th>wals_genus</th>
<th>population</th>
<th>latitude</th>
<th>longitude</th>
<th>phoneme_id</th>
<th>glyph_id</th>
<th>glyph</th>
<th>class</th>
<th>comb</th>
<th>num</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPA</td>
<td>1</td>
<td>kor</td>
<td>1</td>
<td>asis</td>
<td>Korean</td>
<td>42,000,000</td>
<td>37:30</td>
<td>128:0</td>
<td>1</td>
<td>1</td>
<td>t̂h</td>
<td>cons</td>
<td>c-d-c-c</td>
<td>4</td>
</tr>
<tr>
<td>SPA</td>
<td>3</td>
<td>lbe</td>
<td>1</td>
<td>ncau</td>
<td>Lak-Dargwa</td>
<td>157,000</td>
<td>42:0</td>
<td>47:0</td>
<td>124</td>
<td>1</td>
<td>t̂h</td>
<td>cons</td>
<td>c-d-c-c</td>
<td>4</td>
</tr>
<tr>
<td>SPA</td>
<td>5</td>
<td>kat</td>
<td>1</td>
<td>kart</td>
<td>Kartvelian</td>
<td>3,900,000</td>
<td>42:0</td>
<td>44:0</td>
<td>203</td>
<td>1</td>
<td>t̂h</td>
<td>cons</td>
<td>c-d-c-c</td>
<td>4</td>
</tr>
<tr>
<td>SPA</td>
<td>6</td>
<td>bsk</td>
<td>1</td>
<td>asis</td>
<td>Burushaski</td>
<td>87,000</td>
<td>36:30</td>
<td>74:30</td>
<td>240</td>
<td>1</td>
<td>t̂h</td>
<td>cons</td>
<td>c-d-c-c</td>
<td>4</td>
</tr>
<tr>
<td>SPA</td>
<td>14</td>
<td>khm</td>
<td>1</td>
<td>ausa</td>
<td>Khmer</td>
<td>12,300,000</td>
<td>12:30</td>
<td>105:0</td>
<td>632</td>
<td>19</td>
<td>uː</td>
<td>vowel</td>
<td>v-d</td>
<td>2</td>
</tr>
<tr>
<td>SPA</td>
<td>27</td>
<td>tha</td>
<td>1</td>
<td>taik</td>
<td>Kam-Tai</td>
<td>20,200,000</td>
<td>15:00</td>
<td>100:40</td>
<td>1150</td>
<td>19</td>
<td>uː</td>
<td>vowel</td>
<td>v-d</td>
<td>2</td>
</tr>
</tbody>
</table>

Property („Relation“)
From Tables to RDF ...

1. Decompose tables into triples, i.e.,
   - entity   attribute   value       resp.
   - Subject   Property   Object
From Tables to RDF ...

1. Decompose tables into triples, i.e.,

- entity, attribute, value, resp.

- Subject, Property, Object

```
Source | id | ISO639-3 | trump | root | wals_genus | population | latitude | longitude | phoneme_id | glyph_id | glyph | class | comb | num
SPA    | 1  | kor      | 1     | asis | Korean     | 42,000,000 | 37:30     | 128:0      | 1          | tj     | cons   | c-d-c-c | 4
SPA    | 3  | lbe      | 1     | ncau | Lak-Dargwa | 157,000    | 42:0      | 47:0       | 124        | tj     | cons   | c-d-c-c | 4
SPA    | 5  | kat      | 1     | kart | Kartvelian | 3,900,000  | 42:0      | 44:0       | 203        | tj     | cons   | c-d-c-c | 4
SPA    | 6  | bsk      | 1     | asis | Burushaski | 87,000     | 36:30     | 74:30      | 240        | tj     | cons   | c-d-c-c | 4
SPA    | 14 | kham     | 1     | ausa | Khmer     | 12,300,000 | 12:30     | 105:0      | 632        | u:     | vowel  | v-d    | 2
SPA    | 27 | tha      | 1     | taik | Kam-Tai   | 20,200,000 | 15:00     | 100:40     | 1150       | u:     | vowel  | v-d    | 2
```
1. Decompose tables into triples, i.e.,
   - entity      attribute      value      resp.
   - Subject     Property       Object
From Tables to RDF ...

1. Decompose tables into triples
2. Multiple triples constitute a graph
From Tables to RDF ...

1. Decompose tables into triples
2. Multiple triples constitute a graph
1. Decompose tables into triples
2. Multiple triples constitute a graph
3. A graph can aggregate triples from other sources, as well
From Tables to RDF ...

Graphs can be represented in other ways, but RDF allows us to

1. Provide explicit semantics (RDF Schema, Ontology)
2. Check consistency and infer implicit information
3. Merge (not only syntactically, but semantically)
4. Query
5. Link (enrich with external data)
From Tables to RDF ...

Graphs can be represented in other ways, but RDF allows us to

1. Provide explicit semantics (RDF Schema, Ontology)
2. Check consistency and infer implicit information  
   
3. Merge (not only syntactically, but semantically)
4. Query
5. Link (enrich with external data)
Graphs can be represented in other ways, but RDF allows us to

1. Provide explicit semantics (RDF Schema, Ontology)
2. Check consistency and infer implicit information
3. Merge (not only syntactically, but semantically)
4. Query
5. Link (enrich with external data)

URIs & SPARQL
Uniform Resource Identifiers (URIs)

- Agree on a common vocabulary and names for entities
- **URIs** provide globally unique identifiers

```
"hasSegment"
```

```
@prefix phoible: <http://mlode.nlp2rdf.org/resource/phoible/>

... phoible:hasSegment ...
```
Merge data and query it using the W3C standard SPARQL (SPARQL Protocol and Query Language)

“the SQL of the Semantic Web”

```
SELECT DISTINCT ?language
WHERE {
  ?language phoible:hasSegment ?segment .
  ?segment phoible:hasFeature phoible:delayed_release
}
```
use URIs as names for things  

- links to external URIs (links) allow us to retrieve more information from these sites

if they can be resolved via HTTP

and provide information as RDF*

and they include links to other URIs

⇒ then, this is Linked Data (informally)

@prefix phoible: <http://mlode.nlp2rdf.org/resource/phoible/>

phoible:khm phoible:hasSegment "u:".


Turtle notation

http://www.w3.org/DesignIssues/LinkedData.html
From Tables to RDF to Linked Data

@prefix phoible: <http://mlode.nlp2rdf.org/resource/phoible/>.

phoible:khm phoible:hasSegment "u:".


Turtle notation
Linked Open Data: The 5 star plan

★ Make your data available on the Web under an open license
★★ Make it available as structured data
(Excel sheet instead of image scan of a table)
★★★ Use a non-proprietary format
(CSV file instead of an Excel sheet)
★★★★ Use Linked Data format
(URIs to identify things, RDF to represent data)
★★★★★ Link your data to other people’s data to provide context

Linked Open Data cloud: Sep 2011

Source http://lod-cloud.net
Linguistically relevant LOD resources

Source http://lod-cloud.net
Linked Data for Linguistics

Chiarcos, Littauer, Mendes, Moran & Nordhoff (2013)
Linked Data for Linguistics

- Representation and modelling
- Dynamic Import
- Structural interoperability
- Conceptual interoperability
- Federation
- Community and ecosystem
Linked Data for Linguistics

- Representation and modelling
- Dynamic Import
- Structural interoperability
- Conceptual interoperability
- Federation
- Community and ecosystem
Information Integration

- Structural interoperability
  - *same query language* for different data sets
- Conceptual interoperability
  - *same query* for different data sets
- Federation
  - *a single query* for different, distributed data sets

(simplified)
Community and Ecosystem

- RDF has been used in different contexts
  - Active community of users and developers
  - Rich technological infrastructure
  - Semantic Web: applied to **lexical** resources

- Also, it was applied to other linguistic resources
  - linguistic terminology  (Farrar & Langendoen 2003)
  - corpora  (Burchardt et al. 2005)
  - typological databases  (Saulwick et al. 2005)

=> Linguistic Linked Open Data cloud  (Chiarcos et al. 2012)
Linguistic Linked Open Data cloud

- a collection of linguistic resources
  - published under open licenses
  - as linked data
  - decentralized developed and maintained
  - meta data at http://datahub.io
    => cloud diagram

- developed as a community effort in the context of the Open Linguistics Working Group of the Open Knowledge Foundation
Open Knowledge Foundation (OKFN, http://okfn.org)

- non-profit organization
- founded in 2004
- promote open knowledge in all its forms
  - e.g., publication of government data (UK, US)
- provide infrastructural support for several working groups
OKFN Open Linguistics Working Group (OWLG)

- founded in Oct 2010 in Berlin, Germany
- open network of individuals interested in
  - linguistic resources and/or
  - their publication under open licenses
- multi-disciplinary
  - NLP/CL, typology/language documentation, IT, ...
- infrastructure
  - mailing list, web site/blog, wiki
  - http://linguistics.okfn.org
Important OWLG goals (http://linguistics.okfn.org)

1. **Promote open data** in relation to language data
2. **Facilitate communication** between researchers who use / distribute / maintain open linguistic data
3. **Mediate between providers and users** of technical infrastructures
4. Build and maintain an **index of open linguistic data sources**
Workshop series

Linked Data in Linguistics (LDL)

Multilingual Linked Open Data for Enterprises (MLODE)

Linked Data in Linguistic Typology (LDLT)

- point-to-point cooperations between individual members
- regular telcos/meetings
- workshops
- joint publications and presentations
- LLOD cloud development
new diagram, introduced tomorrow at LDL-2014
Building the Cloud: Examples

- Each data provider has different incentives to use Linked Data and/or RDF
- Concepts of RDF and Linked Data have been brought up to solve open problems in different subcommunities of linguistics and neighboring fields
- Examples
  - Corpora
  - Lexicons
  - Linguistic term and data bases
Building the Cloud: Examples

- Each data provider has different incentives to use Linked Data and/or RDF
- Concepts of RDF and Linked Data have been brought up to solve open problems in different subcommunities of linguistics and neighboring fields
- Examples
  - Corpora
  - Lexicons
  - Linguistic term and data bases

TODAY: Underresourced Languages
Case Studies

Linked Data for Underresourced Languages
Under-resourced Languages

- Lack of access to **language data**
  - General lack of language documentation, e.g., dictionaries

- Lack of access to **digital** language data
  - Standardized orthography & encoding (ASCII, KOI-8, SAMPA)
  - Web resources (Wikipedia, Wiktionary, ...)

- Lack of **IT/NLP support**
  - Localized text processing software
  - Basic Language Resource Kit (http://www.blark.org/)

- Limited **interoperability** of data and tools
  - tools & annotations use different formats and conventions
Linked Data may

1. Improve conceptual and structural interoperability

- Lack of access to **language data**
  - General lack of language documentation, e.g., dictionaries

- Lack of access to **digital** language data
  - Standardized orthography & encoding (ASCII, KOI-8, SAMPA)
  - Web resources (Wikipedia, Wiktionary, ...)

- Lack of **IT/NLP support**
  - Localized text processing software
  - Basic Language Resource Kit (http://www.blark.org/)

- Limited **interoperability** of data and tools
  - tools & annotations use different formats and conventions
Linked Data may

- Lack of access to **language**
  - General lack of language documentation, e.g., dictionaries

- Lack of access to **digital** language data
  - Standardized orthography & encoding (ASCII, KOI-8, SAMPA)
  - Web resources (Wikipedia, Wiktionary, ...)

- Lack of **IT/NLP** support
  - Localized text processing software
  - Basic Language Resource Kit (http://www.blark.org/)

- Limited **interoperability** of data and tools
  - tools & annotations use different formats and conventions

1. Improve conceptual and structural interoperability
   1.a between languages => Projection
Linked Data may

- Lack of access to **language** data
  - General lack of language documentation, e.g., dictionaries

- Lack of access to **digital** language data
  - Standardized orthography & encoding (ASCII, KOI-8, SAMPA)
  - Web resources (Wikipedia, Wiktionary, ...)

- Lack of **IT/NLP** support
  - Localized text processing software
  - Basic Language Resource Kit (http://www.blark.org/)

- Limited **interoperability** of data and tools
  - tools & annotations use different formats and conventions

1. Improve conceptual and structural interoperability
2. Guide digitization efforts
Linked Data may

- Lack of access to **language data**
  - General lack of language documentation, e.g., dictionaries
- Lack of access to **digital language data**
  - Standardized orthography & encoding (ASCII, KOI-8, SAMPA)
  - Web resources (Wikipedia, Wiktionary, ...)
- Lack of **IT/NLP support**
  - Localized text processing software
  - Basic Language Resource Kit (http://www.blark.org/)
- Limited **interoperability** of data and tools
  - tools & annotations use different formats and conventions

1. Improve conceptual and structural interoperability
2. Guide digitization efforts
3. (Partially) compensate the lack of lexical resources
Case Studies

1. Improve conceptual and structural interoperability

(A) Shared vocabularies
- lemon: lexicons
- lexvo, Glottolog: languages
- PHOIBLE: phonemes
- OLiA: annotations

(B) Link and query multiple dictionaries
- QHL, PanLex, GermLex, ...
- Towards a Comparative-Lexicographical Workbench

2. Guide digitization efforts

3. (Partially) compensate the lack of lexical resources
Case Studies

1. Improve conceptual and structural interoperability
2. Guide digitization efforts
3. (Partially) compensate the lack of lexical resources

(A) Shared vocabularies
- lemon: lexicons
- lexvo, Glottolog: languages
- PHOIBLE: phonemes
- OLiA: annotations

(B) Link and query multiple dictionaries
- QHL, PanLex, GermLex, ...
- Towards a Comparative-Lexicographical Workbench

Today
Case Studies

- Linking collections of dictionaries, e.g.,
  - PanLex (http://panlex.org/)
    - Dictionaries for all languages in the world
  - QuantHistLing (http://quanthistling.info/)
    - South America
  - GermLex (http://datahub.io/dataset/germlex)
    - Germanic languages
Case Studies

- Linking collections of dictionaries, e.g.,
  - PanLex (http://panlex.org/)
    - dictionaries for all languages in the world
  - QuantHistLing (http://quanthistling.info/)
    - South America (Moran and Brümmer 2013)
  - GermLex (http://datahub.io/dataset/germlex)
    - Germanic languages (tomorrow @ LDL-2014)
• Team: Michael Cysouw (PI), Jelena Prokić, Johann Mattis-List, Peter Bouda, Steven Moran, Ramon Rodriguez, Ioana Fugaru

• Project aims:
  • to digitize around 200 works, most of which are currently only available in print and many of which are the only resources available for the poorly described and under-resourced languages that they describe
  • http://quanthistling.info/index.php?id=resources
  • to develop new and innovative computer-assisted methods to quantitatively analyze this information
  • to uncover and clarify phylogenetic relationships between native South American languages using quantitative methods
QuantHistLing: Source Data
QuantHistLing: Extraction

- Digitization pipeline (prepares the data for analysis)
  - [http://quanthistling.info/data/](http://quanthistling.info/data/)
- We digitize the whole resource
- 80 dictionaries down, 120 to go...
- Simple data output format that contains metadata (prefixed with “@”) and tab-delimited lexical output

```
@date: 2012-11-23
@source_title: Analise descritiva e teorica do Katukino-Pano
@source_author: de Aguiar, Maria Sueli
@source_year: 1994
@doculect: Katukina, n/a, Katukina, Panoan
@doculect: Portugues, por, Portugues, Panoan
QLCID HEAD HEADDDOCULECT TRANSLATION TRANSLATIONDOCULECT
aguiar1994/329/1 ai Katukina presente Portugues
aguiar1994/329/2 aima Katukina solteiro Portugues
aguiar1994/329/3 ain Katukina esposa Portugues
```
QuantHistLing: From Data to Database using Linked Data and *lemon*

- We convert the QLC data into Linked Data that conforms to the Lemon model with a simple Python script.

- Lemon is an ontological model for modeling lexicons and machine-readable dictionaries for linking to the Semantic Web and the Linked Data cloud.
  - [http://lemon-model.net/](http://lemon-model.net/)

- Lemon developers also active in the W3C Ontology-Lexica Community Group.
  - Goal is to "develop models for the representation of lexica (and machine readable dictionaries) relative to ontologies".
  - [http://www.w3.org/community/ontolex/](http://www.w3.org/community/ontolex/)
Why lemon:
(Relatively) widely used & actively maintained
lemon Core
QuantHistLing: *lemon* Sample

- We convert the QLC data into Linked Data that conforms to the Lemon model with a simple Python script.
QuantHistLing: Search

- As a first step, we have converted the QHL data into RDF and it is available online through a SPARQL endpoint
  - http://linked-data.org/sparql/ (preliminary)
  - http://linked-data.org/datasets/ (data dump)
- Querying the combined dictionaries and lexicons is straightforward
  - Return all triples:
    - SELECT * WHERE
      {GRAPH <http://quanthistling.info/lod/>}
        {?s ?p ?o}
    
- Returns over 3.8 million triples
Pairs of languages in the translation graph that contain written forms for the lexical sense “casa”

```sql
PREFIX lemon: <http://www.monnet-project.eu/lemon#>
PREFIX lexinfo: <http://lexinfo.net/ontology/2.0/lexinfo#>

SELECT ?wordForm1 ?language1 ?wordForm2 ?language2 WHERE {
  GRAPH <http://quanthistingling.info/lod/> {
    ?word1 a lemon:LexicalForm;
    lemon:writtenRep ?wordForm1.
    ?entry1 lemon:form ?word1;
    lemon:sense ?sense1.
    ?word2 a lemon:LexicalForm;
    lemon:writtenRep ?wordForm2.
    ?entry2 lemon:form ?word2;
    lemon:sense ?sense2.
    FILTER(str(?wordForm1)="casa")
  }
}
```
QuantHistLing: Search

Works, but maybe not exactly convenient ...
Linked Open Dictionaries (LiODi)
Towards a Workbench

Scenario: Language contact studies
- query for a lexeme across multiple dictionaries
  - filter for source and target languages and language families
- query across diverse resources available in the LLOD cloud
  - glosses to be linked to existing lemon resources, e.g., DBnary, WordNet

Currently in preparation
- Chiarcos, C. (in prep.), Linked Open Dictionaries. Towards a Workbench for Comparative Lexicography
- Early implementation efforts in Frankfurt
Linked Open Dictionaries (LiODi) Towards a Workbench
Given a lexeme in the source variety:

Retrieve (a) all direct matches from the target varieties, and (b) every other word from the target varieties that is either (b.1) linked with a result from (a), or (b.2) has the same gloss as a result from (a)
Linked Open Dictionaries (LiODi)
Towards a Workbench

Visualize the results such that
(a) lemma and gloss are shown,
(b) matches are grouped according to some (externally provided) pylogenetic tree, and
(c) the path of dictionaries consulted is shown
What’s in for underresourced languages?

- Language documentation
  - Material collected on field trips is usually *afterwards* analysed, e.g., using annotation tools like ELAN or Toolbox
  - For the analysis of difficult words, it may not be possible to get in contact with native speakers
  - A distributional analysis of the word form and its meaning in related or neighboring varieties may help to disambiguate

=> partially compensates the lack of lexical resources
But wait!

If a single query is to be applied on different resources, then relying on *lemon* is not enough

- *lemon* provides data structures, **but**
  - for content and metadata, it relies on external vocabularies

Interoperability depends on a *bundle* of vocabularies

- WordNet, DBpedia, *any* ontology (lexical senses)
- lexvo (language identifiers)
- glottolog (languoid identifiers *from linguistic typology*)
- PHOIBLE (phoneme inventories and phonological structures)
- OLiA (annotations)
- ISOcat (resource metadata)
- GOLD (grammatical concepts)
Discussion

Problems and Questions
Summary

- Linked Data
  - General introduction
  - Benefits for linguist(ic)s
- Linguistic Linked Open Data
  - Community activities
- Use cases
  - Querying multiple dictionaries, filter and visualize by structured language metadata
    - Independently developed resources, shared vocabularies
Problems and Questions, and what to do about them

- RDF is misunderstood
  - RDF/XML is hard to read and process
  - As an alternative format, Turtle may be a compromise

- SPARQL is complicated
  - but not meant to be used by linguists in the field – it can nevertheless be used to develop tools for them

- Federation is a great concept, but causes too much traffic
  - Maintain your own sync‘ed copy of relevant external resources
Problems and Questions, and what to do about them

- *lemon* is neither developed for nor by linguists
  - but a vocabulary under development, so giving linguists a voice may be an option

- How can I publish my data as Linked Data?
  - Ask, e.g., on the OWLG mailing list. Most likely, someone may help, and maybe, this will be a linguist, as well.

- Who could host my data?
  - That’s a problem we can only solve as a community. If you write your next proposal, think of an end point for your data and help others to host (some of) their data.
Problems and Questions, and what to do about them

- How do I get into the LLOD cloud (diagram)?
  - Convert your data to RDF and put it under an open license
  - Create an entry at datahub.io
    - provide URL of a data dump or a SPARQL end point
  - Tag it as „linguistic“
  - Specify „triples“ and „links:xy“ (for datahub dataset xy).
  - Join the mailing list and wait for the next diagram generation announcement to make sure all went well.
  - Make sure your URLs are alive.
Problems and Questions, and what to do about them

- I encountered technical issues with datahub.io
  - Possible. It is not a perfect solution, and some colleagues are working on an alternative, but for the moment, we have to rely on it.

- Can I actually *do* anything with the LLOD cloud?
  - No, the diagram is merely a snapshot of the datahub.io metadata. It helps you to discover datasets and their dependencies.
  - But it tells you where to retrieve data dumps for local use or how to call SPARQL end points
Thank you!

Special thanks to
Laurette Pretorius & Claudia Soria,
The Open Linguistics Working Group,
Martin Brümmer, John McCrae,
Robert Forkel, Martin Haspelmath,
Sebastian Hellmann, Sebastian Nordhoff

Christian Chiarcos’ work was partially supported by the LOEWE cluster „Digital Humanities“ funded by the federal state of Hesse.
Steven Moran's work was supported by the ERC starting grant 240816 „Quantitative modeling of historical-comparative linguistics“
Sources

Nontechnical Introduction


PHOIBLE example

Sources

**Linked Data for Linguistics**

**Case Studies: QuantHistLing**
Sources

- **Case Studies: *lemon* Core Model**

- **Case Studies: Comp-Lex Workbench**
  - Chiarcos, C. (in prep.), *Linked Open Dictionaries. Towards a Workbench for Comparative Lexicography*