Using Linked Data to Create a Typological Knowledge Base

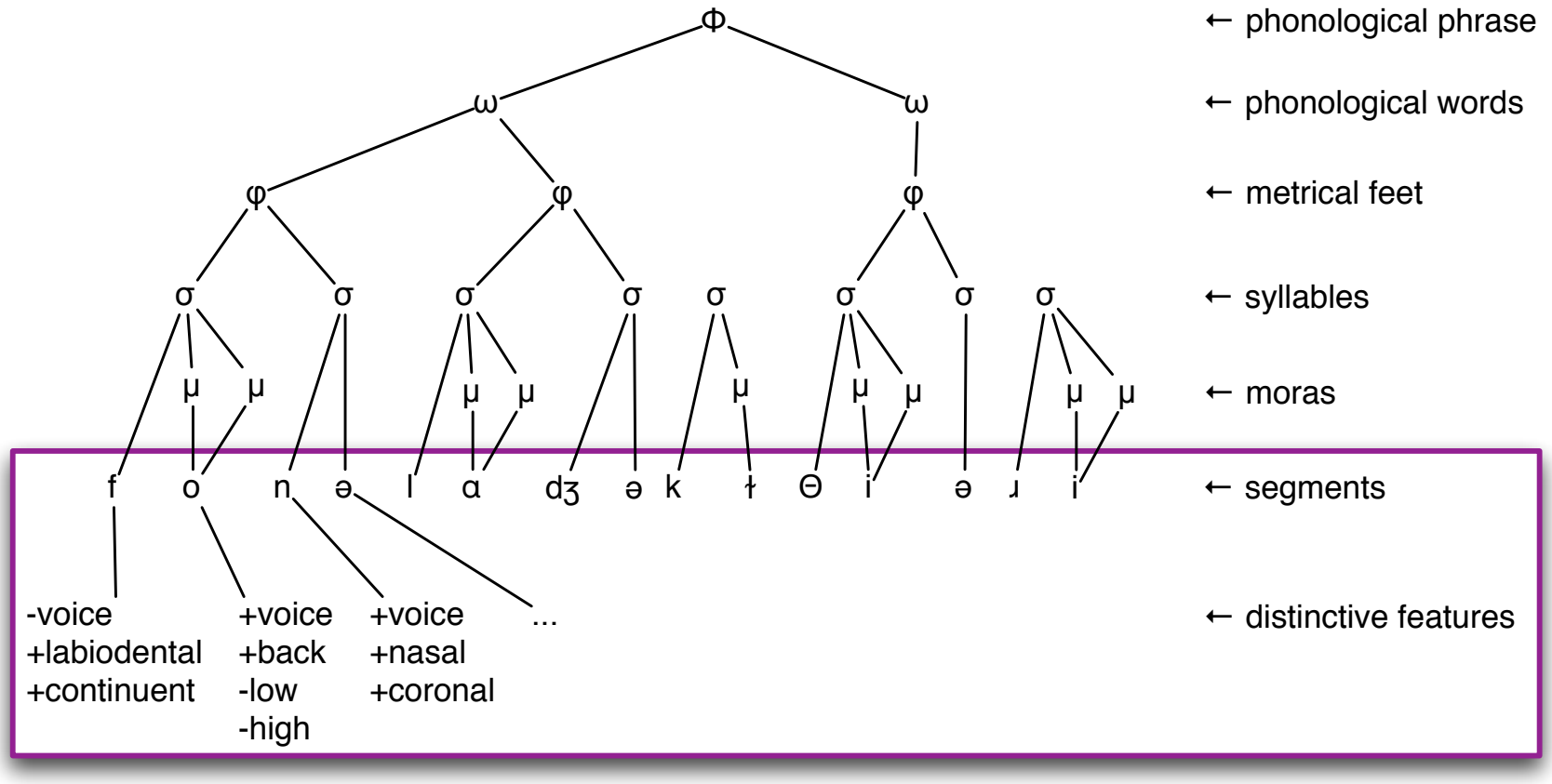
Steven Moran
LMU München
steve.moran@lmu.de

March 8, 2012

Talk Map

- Overview
- Investigation
- Challenges
- Conclusion

Where are we in linguistic structure?



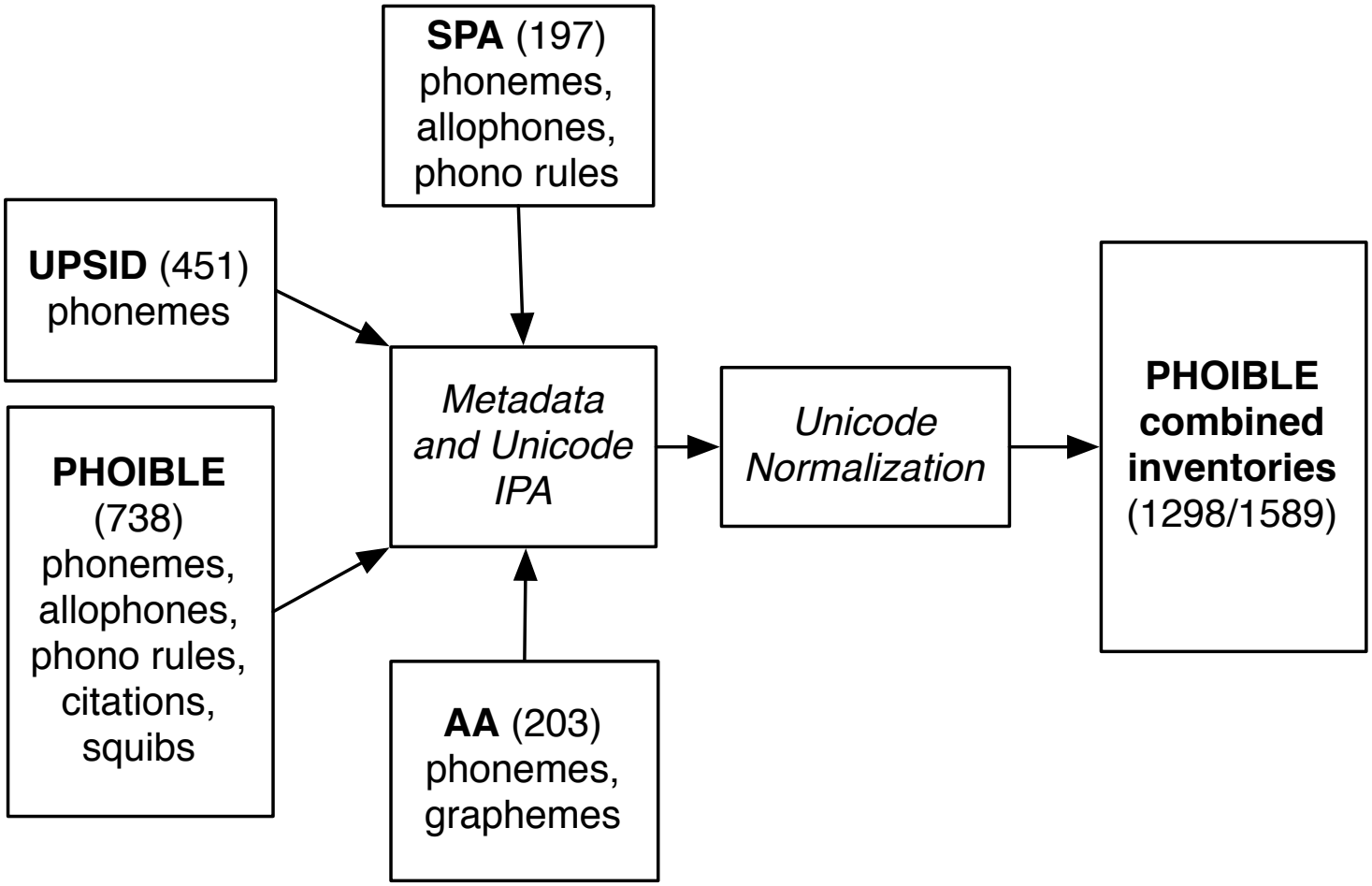
PHOnetics Information Base and LExicon (PHOIBLE)

- A typological data set of segment inventories with linguistic and non-linguistic information
- Linguistic info
  - segment inventories
  - distinctive features
  - genealogical data (language stock and genus)
- Non-linguistic info
  - population figures
  - geographic location (geo-coordinates, country and region)
  - per-capita GDP, etc.

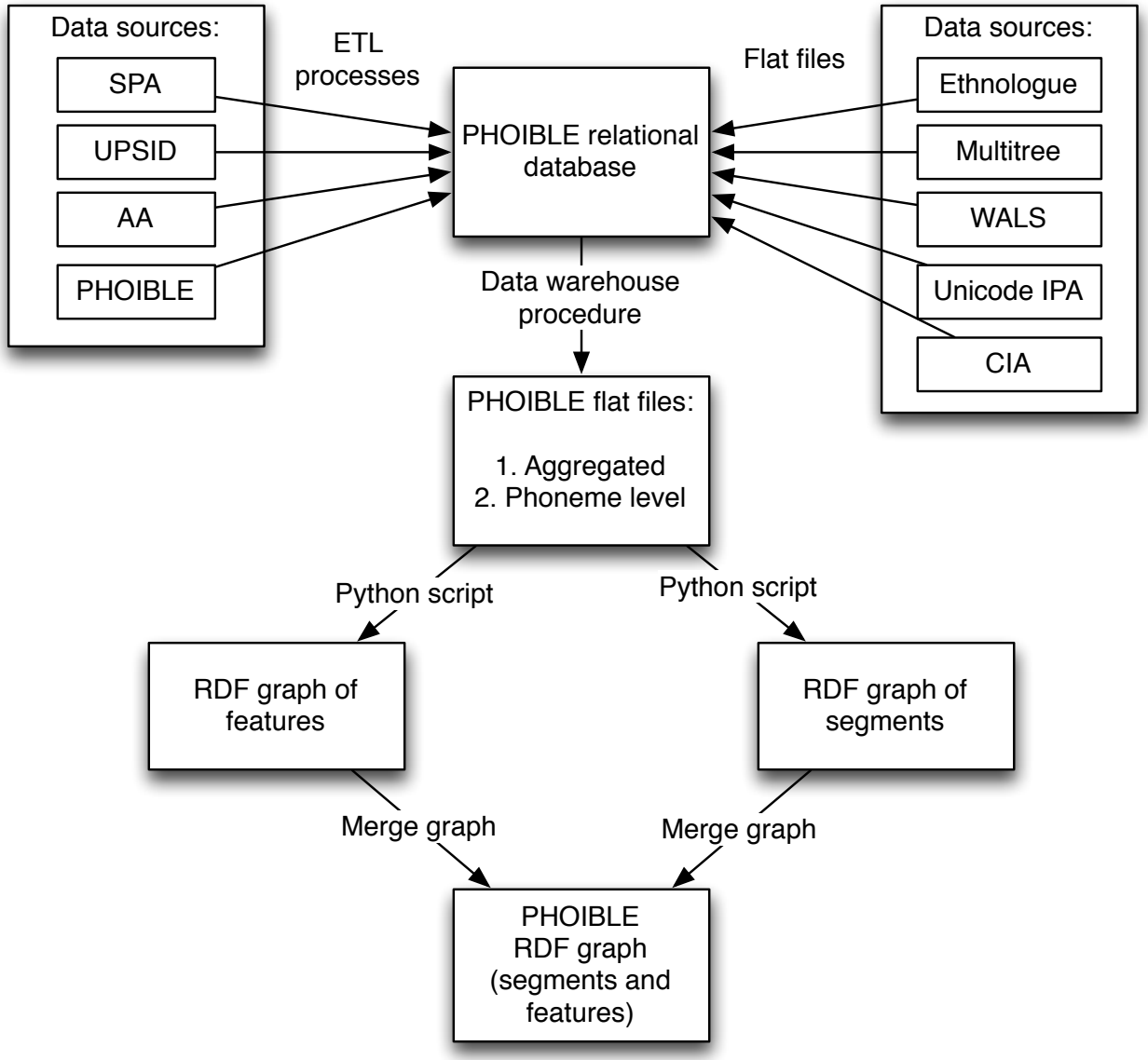
Segment inventory databases

- Stanford Phonology Archive (**SPA**; Crothers et al 1979)
- UCLA Phonological Segment Inventory Database (**UPSID**; Maddieson 1984; Maddieson & Precoda 1990)
- Systèmes alphabétiques des langues africaines (**AA**; Hartell 1993; Chanard 2006)
- **PHOIBLE** data (Moran 2009-2012)

PHOIBLE resources



How was PHOIBLE developed?



PHOIBLE data warehouse flat file table – phoneme level

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	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	3	lbe	1	ncau	Lak-Dargwa	157,000	42:0	47:0	124	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	5	kat	1	kart	Kartvelian	3,900,000	42:0	44:0	203	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	6	bsk	1	asis	Burushaski	87,000	36:30	74:30	240	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	14	khm	1	ausa	Khmer	12,300,000	12:30	105:0	632	19	uː	vowel	v-d	2
SPA	27	tha	1	taik	Kam-Tai	20,200,000	15:00	100:40	1150	19	uː	vowel	v-d	2

Building the RDF graph

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	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	3	lbe	1	ncau	Lak-Dargwa	157,000	42:0	47:0	124	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	5	kat	1	kart	Kartvelian	3,900,000	42:0	44:0	203	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	6	bsk	1	asis	Burushaski	87,000	36:30	74:30	240	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	14	khm	1	ausa	Khmer	12,300,000	12:30	105:0	632	19	uː	vowel	v-d	2
SPA	27	tha	1	taik	Kam-Tai	20,200,000	15:00	100:40	1150	19	uː	vowel	v-d	2

Subject

Building the RDF graph

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	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	3	lbe	1	ncau	Lak-Dargwa	157,000	42:0	47:0	124	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	5	kat	1	kart	Kartvelian	3,900,000	42:0	44:0	203	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	6	bsk	1	asis	Burushaski	87,000	36:30	74:30	240	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	14	khm	1	ausa	Khmer	12,300,000	12:30	105:0	632	19	uː	vowel	v-d	2
SPA	27	tha	1	taik	Kam-Tai	20,200,000	15:00	100:40	1150	19	uː	vowel	v-d	2

Subject

Object

Building the RDF graph

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	1	$t\underline{ʃ}^{h}$	cons	c-d-c-c	4
SPA	3	lbe	1	ncau	Lak-Dargwa	157,000	42:0	47:0	124	1	$t\underline{ʃ}^{h}$	cons	c-d-c-c	4
SPA	5	kat	1	kart	Kartvelian	3,900,000	42:0	44:0	203	1	$t\underline{ʃ}^{h}$	cons	c-d-c-c	4
SPA	6	bsk	1	asis	Burushaski	87,000	36:30	74:30	240	1	$t\underline{ʃ}^{h}$	cons	c-d-c-c	4
SPA	14	khm	1	ausa	Khmer	12,300,000	12:30	105:0	632	19		vowel	v-d	2
SPA	27	tha	1	taik	Kam-Tai	20,200,000	15:00	100:40	1150	19	uː	vowel	v-d	2

Subject

Object

Building the RDF graph

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	1	$t\underline{ʃ}^h$	cons	c-d-c-c	4
SPA	3	lbe	1	ncau	Lak-Dargwa	157,000	42:0	47:0	124	1	$t\underline{ʃ}^h$	cons	c-d-c-c	4
SPA	5	kat	1	kart	Kartvelian	3,900,000	42:0	44:0	203	1	$t\underline{ʃ}^h$	cons	c-d-c-c	4
SPA	6	bsk	1	asis	Burushaski	87,000	36:30	74:30	240	1	$t\underline{ʃ}^h$	cons	c-d-c-c	4
SPA	14	khm	1	ausa	Khmer	12,300,000	12:30	105:0	632	19		vowel	v-d	2
SPA	27	tha	1	taik	Kam-Tai	15:00	100:40	1150	19		uː	vowel	v-d	2

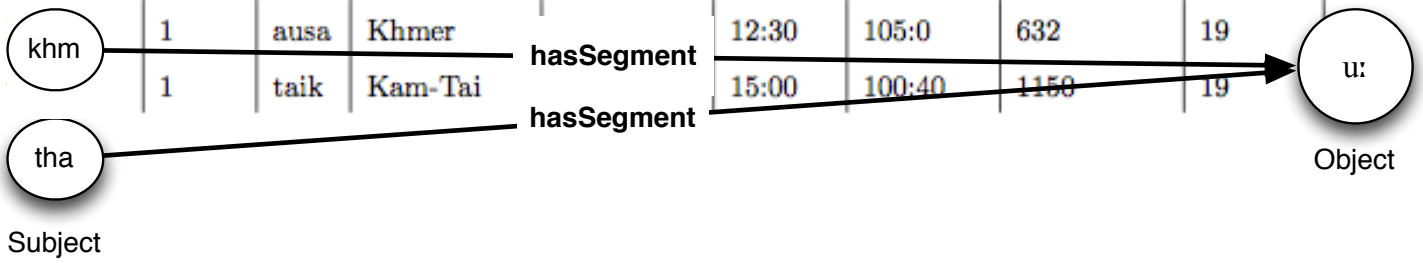
hasSegment

Subject

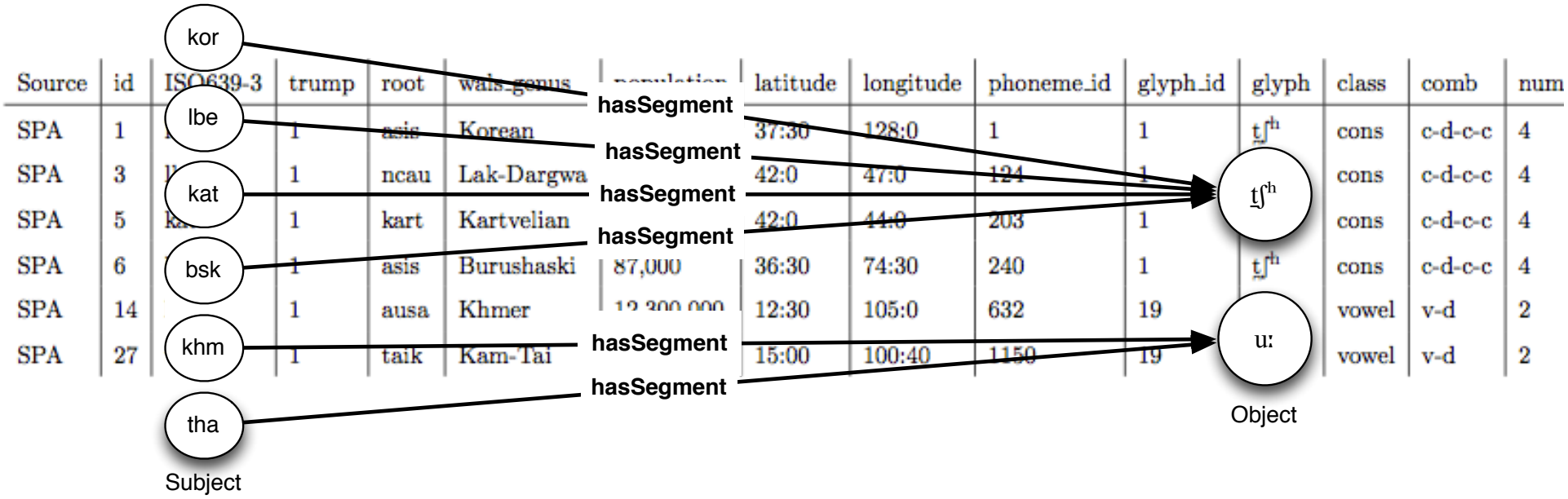
Object

Building the RDF graph

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	1	$t\underset{\frown}{ʃ}^{h}$	cons	c-d-c-c	4
SPA	3	lbe	1	ncau	Lak-Dargwa	157,000	42:0	47:0	124	1	$t\underset{\frown}{ʃ}^{h}$	cons	c-d-c-c	4
SPA	5	kat	1	kart	Kartvelian	3,900,000	42:0	44:0	203	1	$t\underset{\frown}{ʃ}^{h}$	cons	c-d-c-c	4
SPA	6	bsk	1	asis	Burushaski	87,000	36:30	74:30	240	1	$t\underset{\frown}{ʃ}^{h}$	cons	c-d-c-c	4
SPA	14		1	ausa	Khmer		12:30	105:0	632	19		vowel	v-d	2
SPA	27		1	taik	Kam-Tai		15:00	100:40	1150	19		vowel	v-d	2

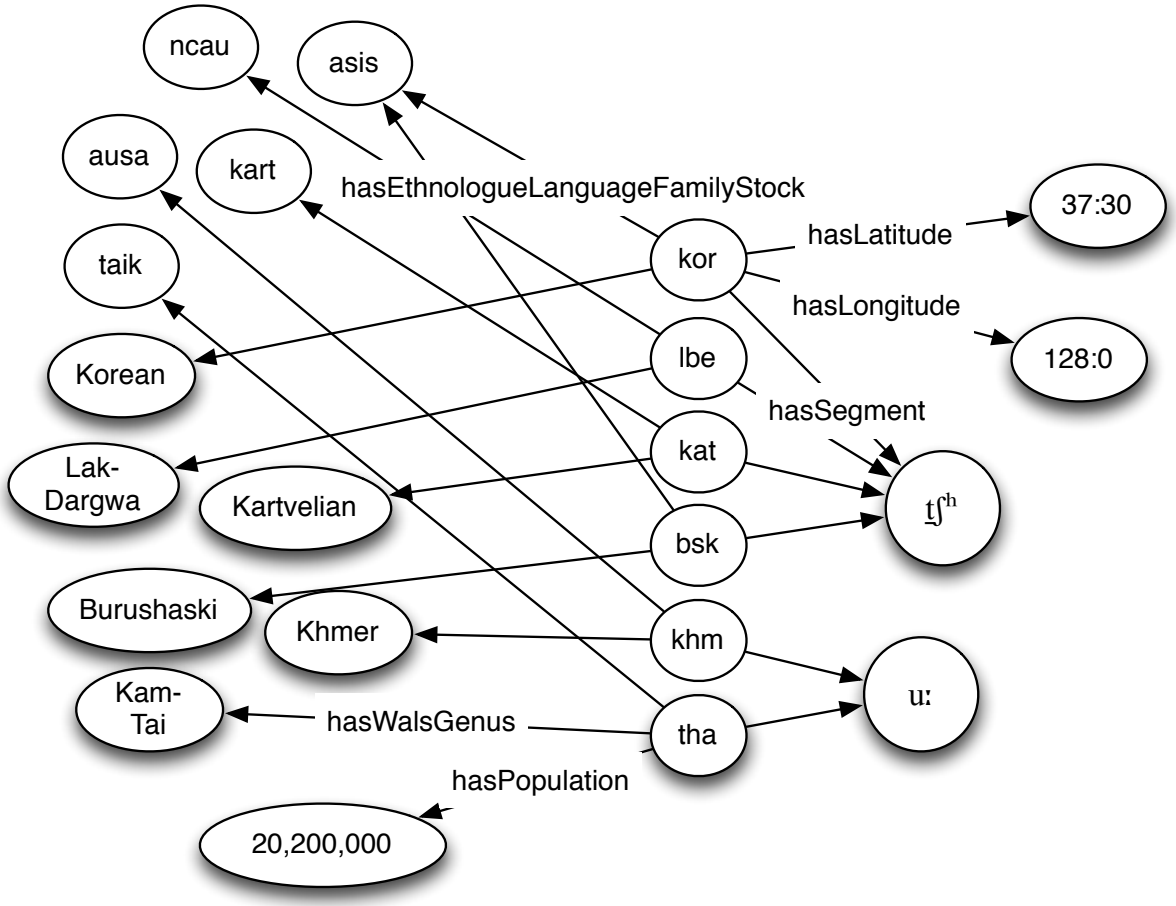


Building the RDF graph

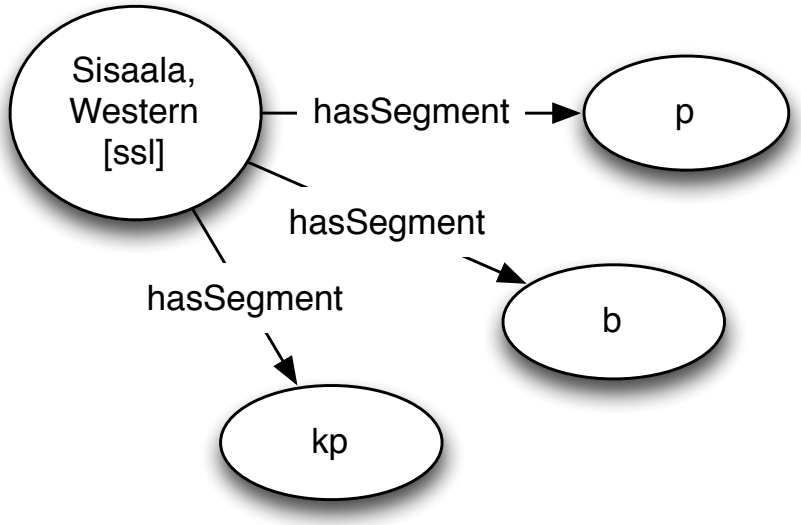


Source	id	ISO639-3	trump	root	wals_genus	population	latitude	longitude	phoneme_id	glyph_id	glyph	class	comb	num
SPA	1		1	asis	Korean		37:30	128:0	1	1	tʃʰ	cons	c-d-c-c	4
SPA	3		1	ncau	Lak-Dargwa		42:0	47:0	124	1		cons	c-d-c-c	4
SPA	5		1	kart	Kartvelian		42:0	44:0	203	1		cons	c-d-c-c	4
SPA	6		1	asis	Burushaski	87,000	36:30	74:30	240	1	tʃʰ	cons	c-d-c-c	4
SPA	14		1	ausa	Khmer	12,300,000	12:30	105:0	632	19		vowel	v-d	2
SPA	27		1	taik	Kam-Tai		15:00	100:40	1150	19	uː	vowel	v-d	2

Building the RDF graph

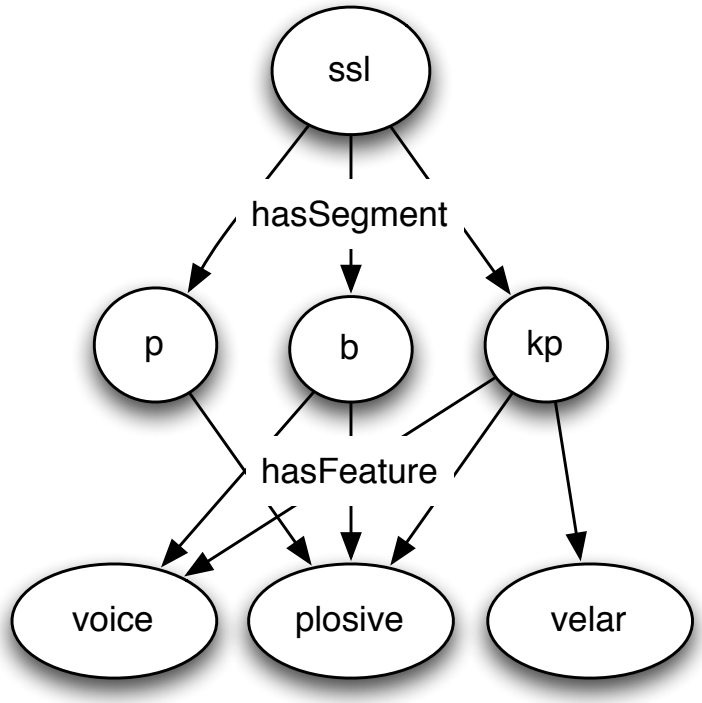


Knowledge base with segments

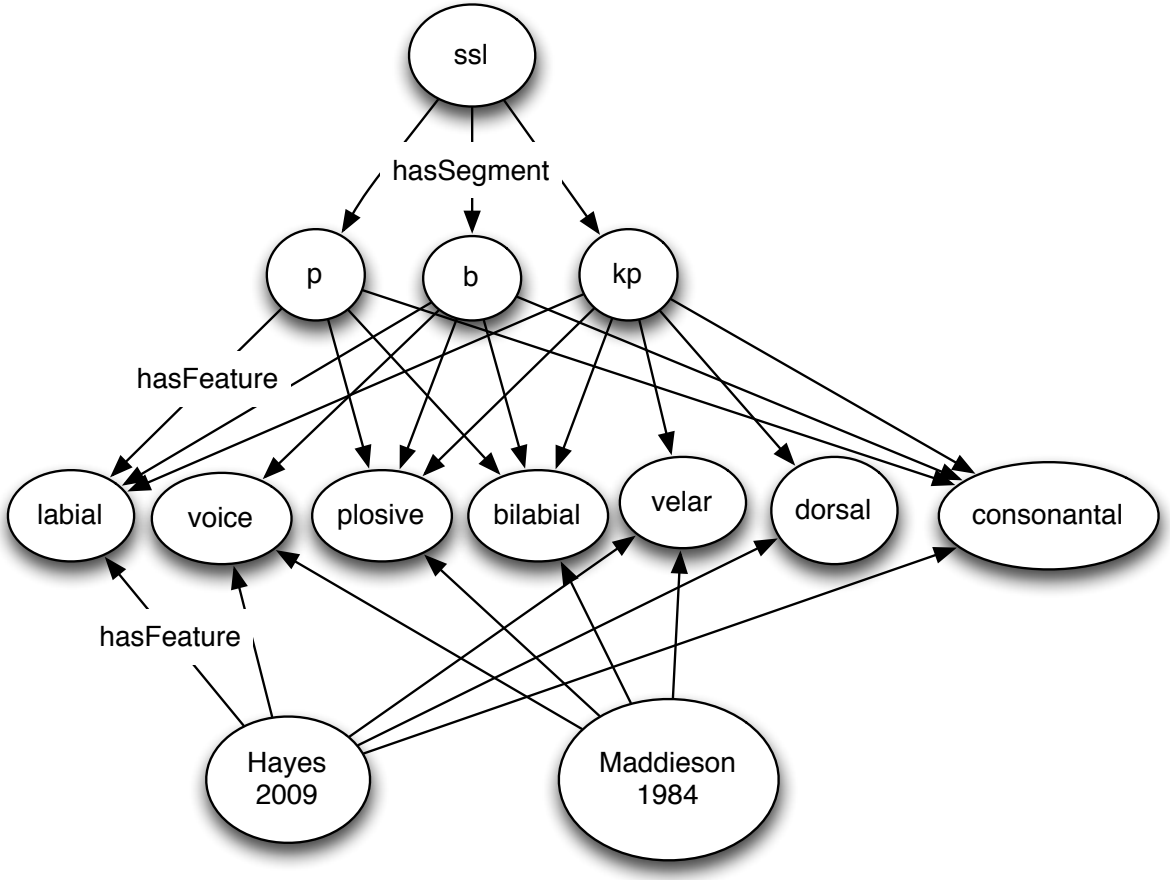


Knowledge base with segments and features

- Features added to the graph by linking them from each segment



Knowledge base with segments, features and feature sets



Mapping distinctive features to segment types

- Features are the atoms that combine compositionally to form a segment
- Query features for natural classes of sounds
- Model in RDF/OWL to hierarchically organize features into a feature geometry

Simple and complex segment feature resolution

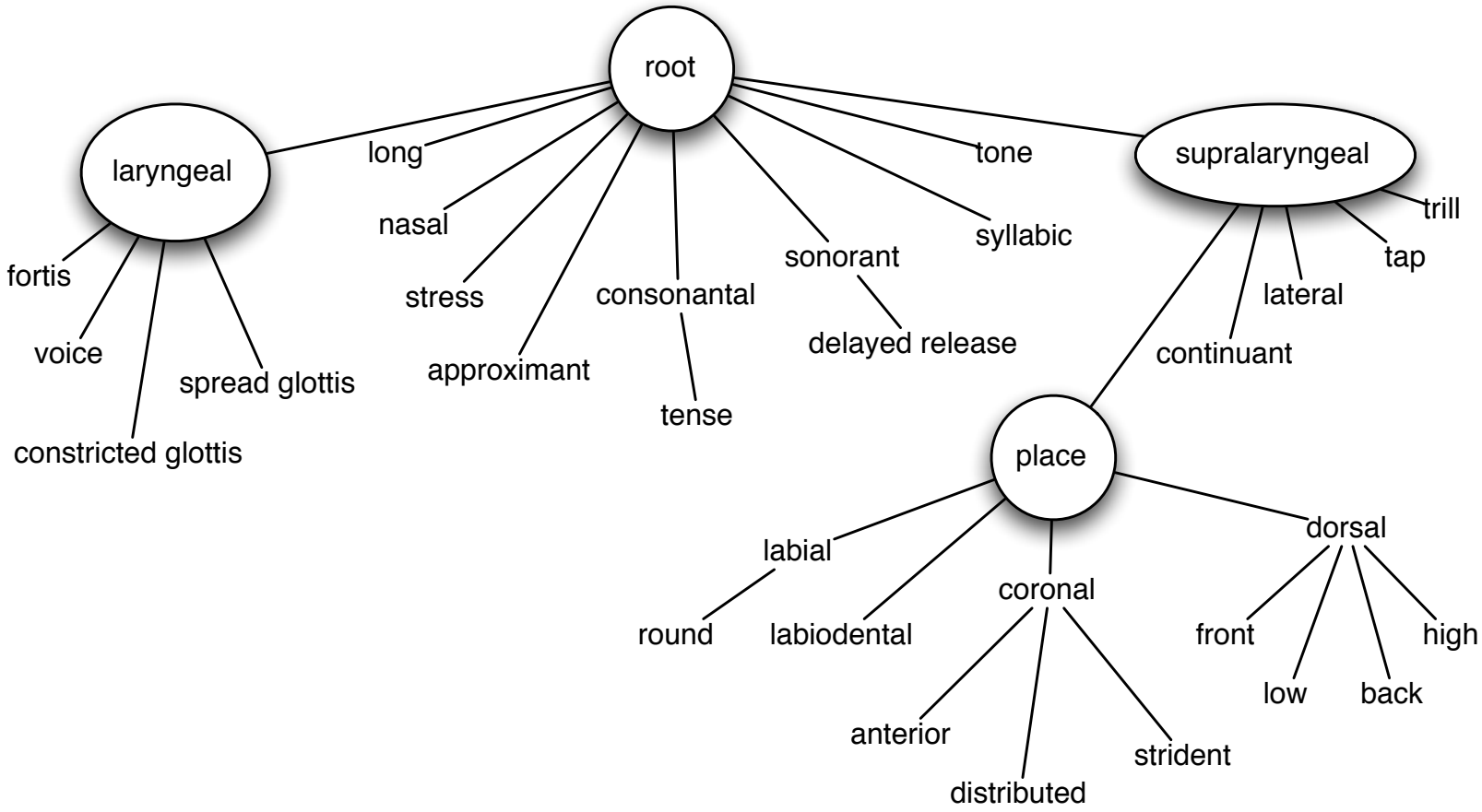
segment	labial	coronal	anterior	distributed	strident	dorsal	high	low	front	back
k	-	-	0	0	0	+	+	-	-	-
p	+	-	0	0	0	-	0	0	0	0
kp	+	-	0	0	0	+	+	-	-	-
p	+	-	0	0	0	-	0	0	0	0
t	-	+	+	-	-	-	0	0	0	0
pt	+	+	+	-	+	-	0	0	0	0
b	+	-	0	0	0	-	0	0	0	0
d	-	+	+	-	-	-	0	0	0	0
bd	+	+	+	-	+	-	0	0	0	0
g	-	-	0	0	0	+	+	-	-	-
b	+	-	0	0	0	-	0	0	0	0
gb	+	-	0	0	0	+	+	-	-	-

Feature specifications for natural classes of sounds

Class of sounds	Feature specification
Vowels	[+syllabic] [-consonantal]
Vowels & Syllabic Consonants	[+syllabic]
Glides	[-syllabic] [-consonantal]
Liquids	[+consonantal] [+approximant]
Nasals	[+sonorant] [-approximant]
Fricatives	[-sonorant] [+continuant]
Affricates	[-continuant] [+delayed release]
Stops	[-delayed release]
Stops & Affricates	[-continuant]
Liquids & Glides	[-syllabic] [+approximant]
Liquids, Glides, & Nasals	[-syllabic] [+sonorant]

Feature geometry

Feature geometry

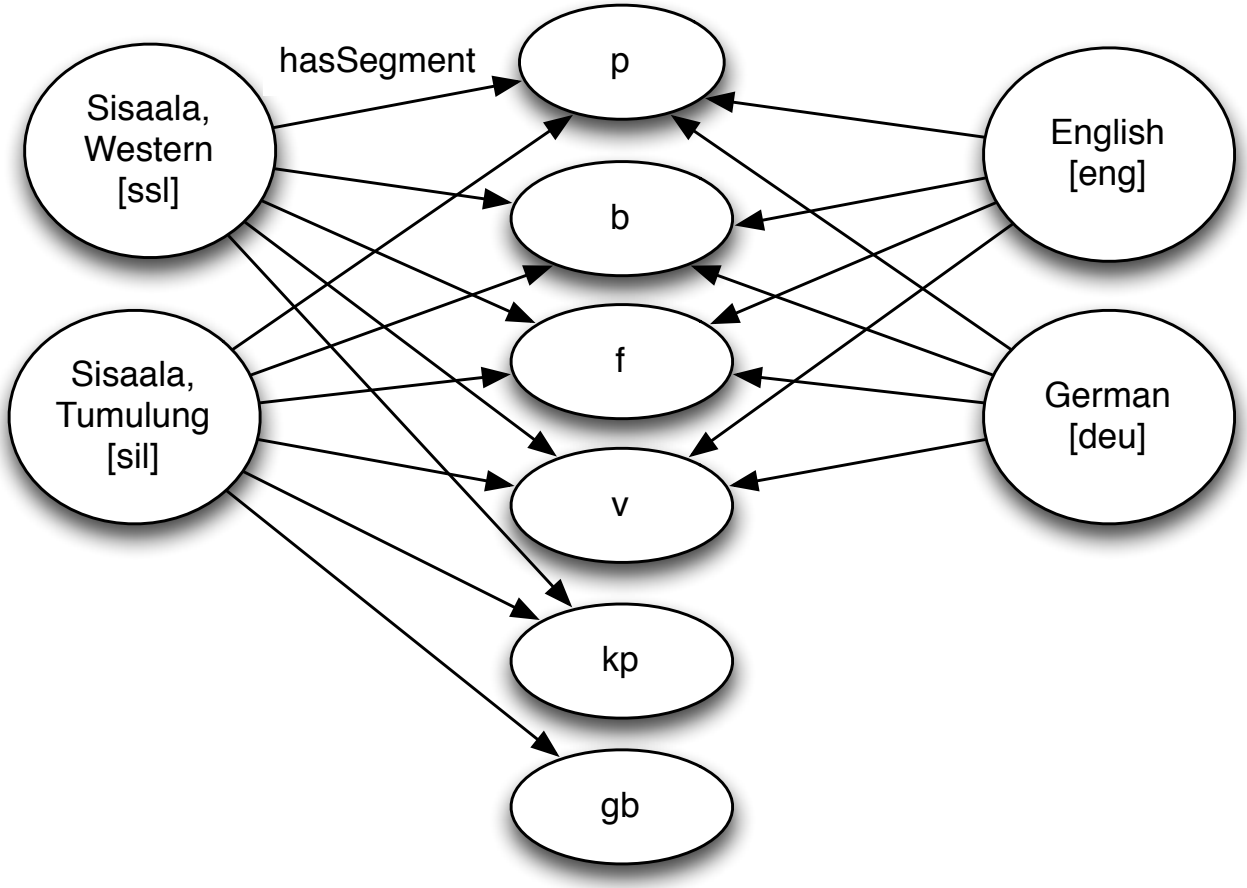


SPARQL

- **S**PARQL **P**rotocol **A**nd **R**DF **Q**uery **L**anguage
- RDF query language (Prud'Hommeaux and Seaborne, 2006)
- consist of triple patterns that match concepts and their relations by binding variables to match graph patterns

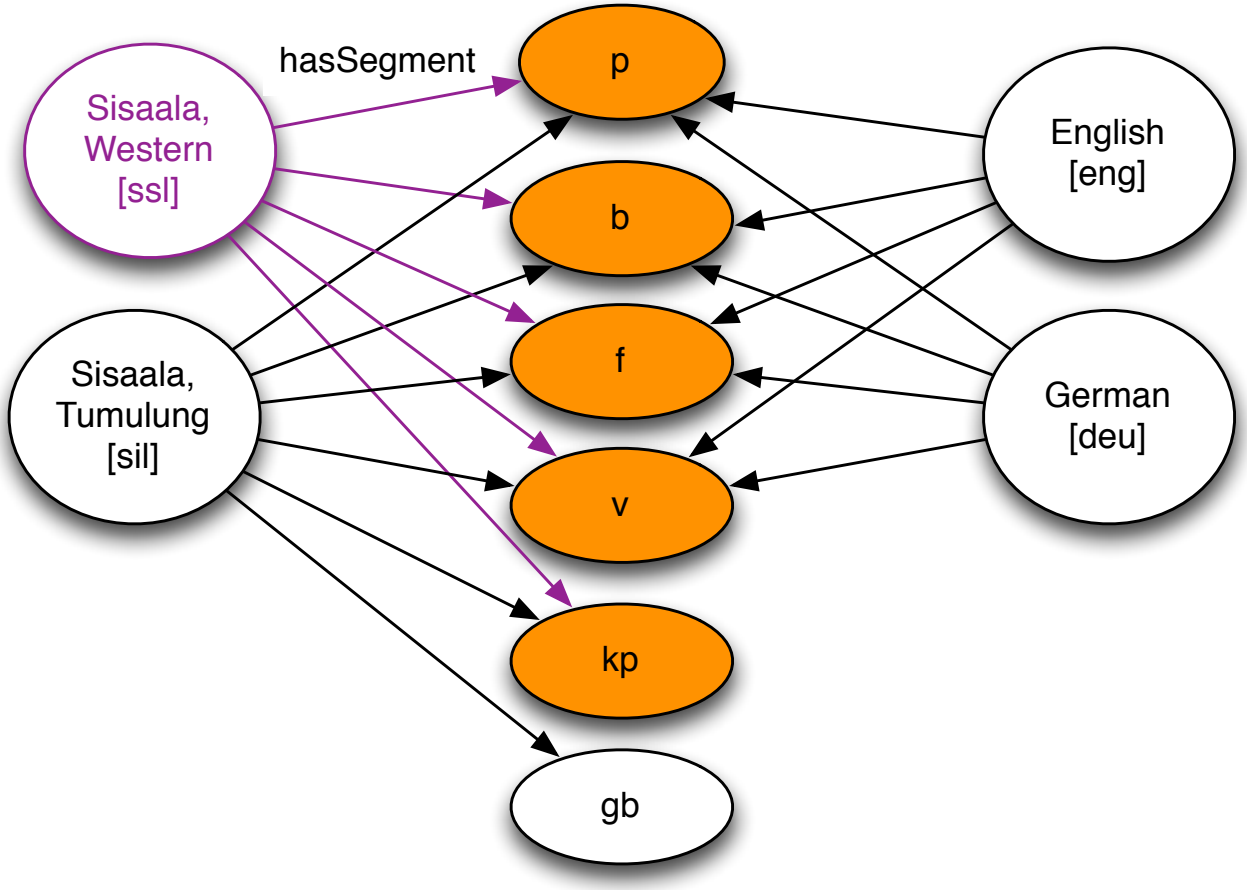
Select segments of a particular language

```
SELECT ?segments
WHERE {ssl hasSegment ?segments}
```



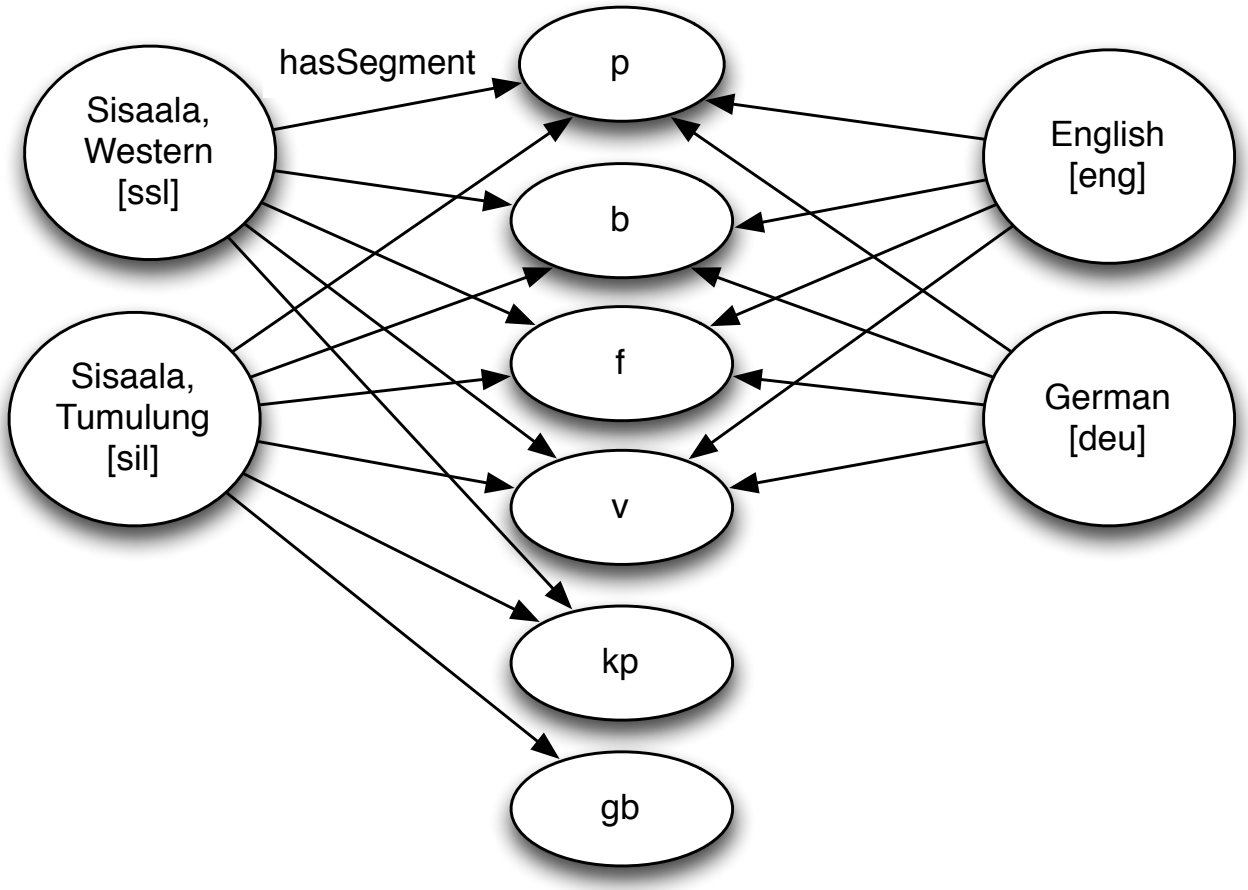
Select segments of a particular language

```
SELECT ?segments
WHERE {ssl hasSegment ?segments}
```



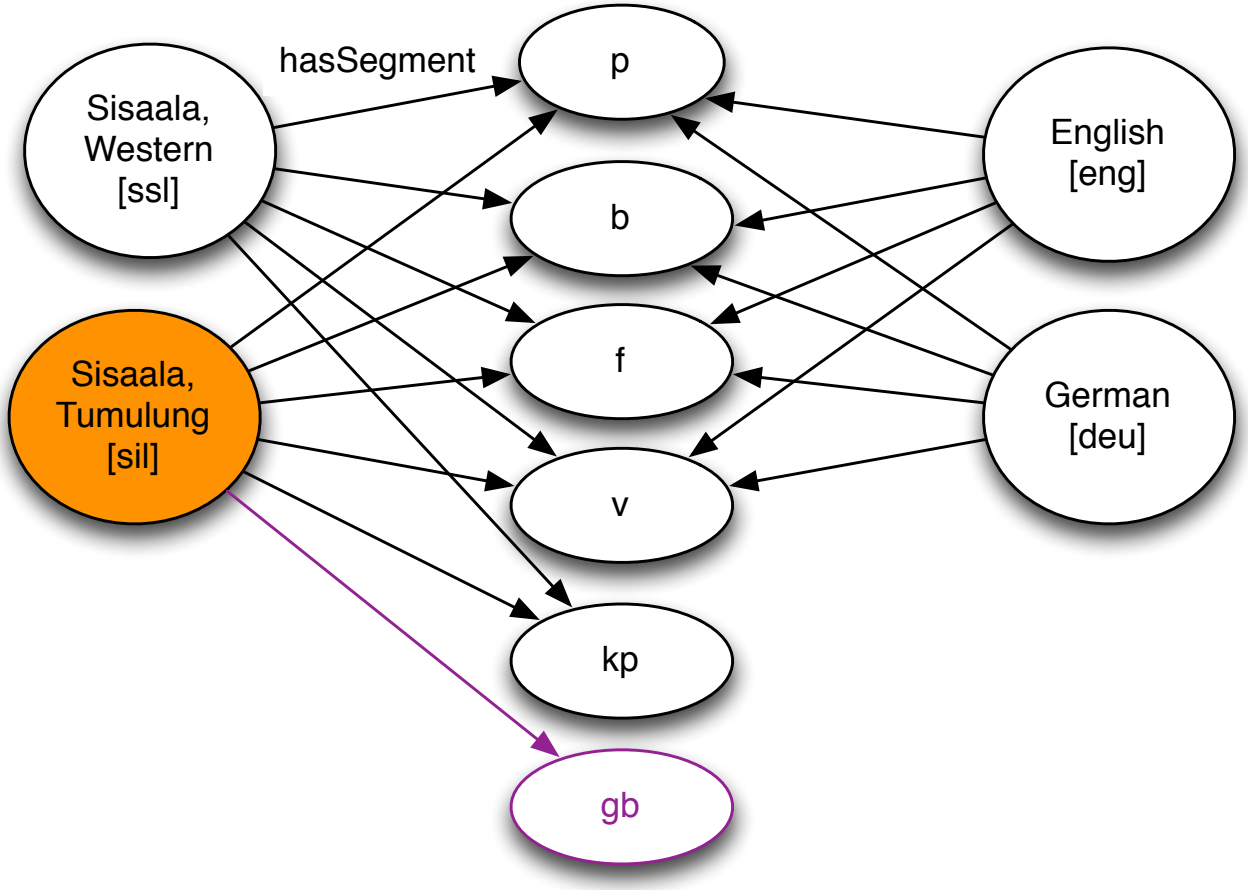
Select languages that have a particular segment

```
SELECT ?languages
WHERE { ?languages hasSegment gb }
```



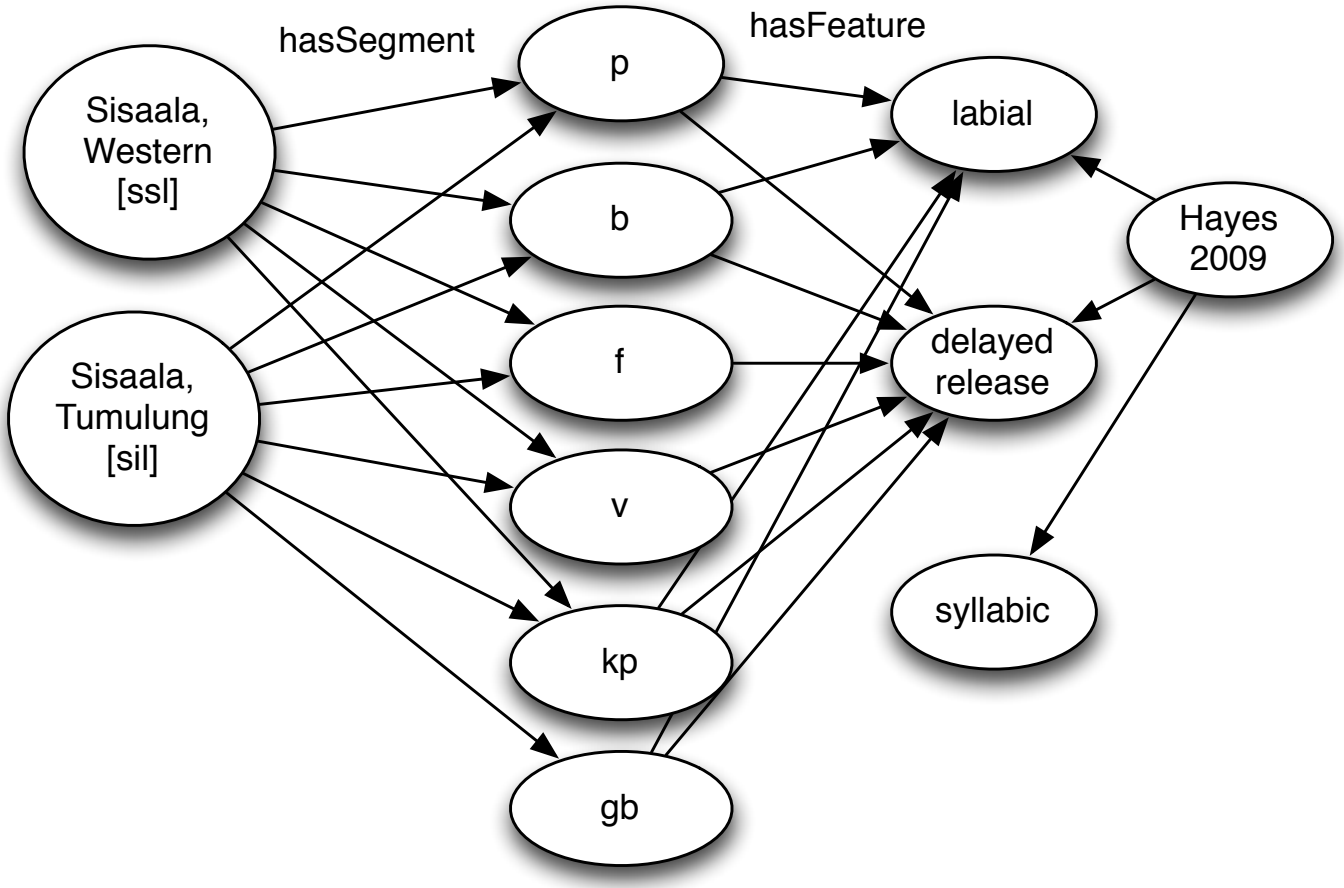
Select languages that have a particular segment

```
SELECT ?languages
WHERE { ?languages hasSegment gb }
```



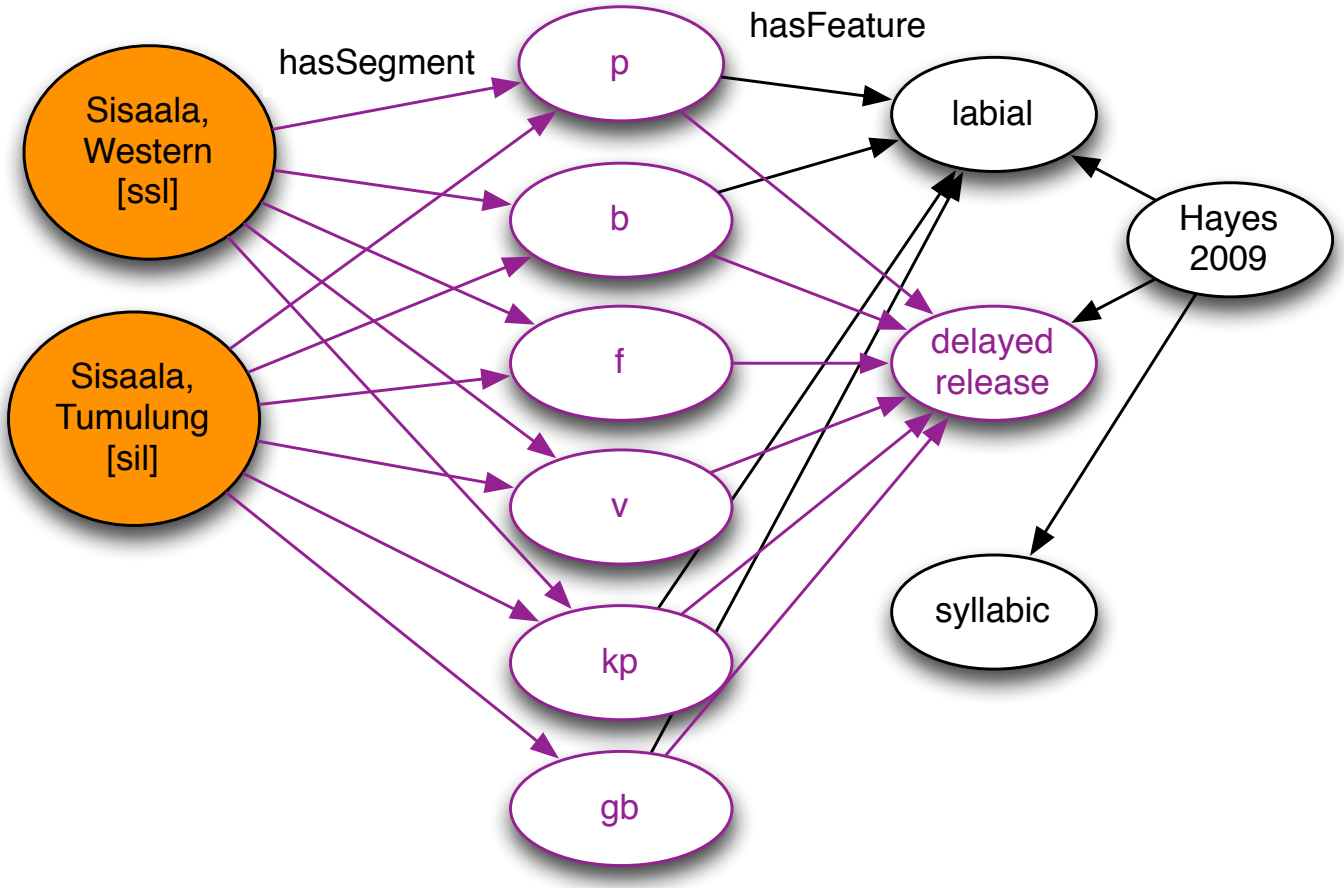
Select languages that have a class of segments

```
SELECT ?languages
WHERE { ?languages hasSegment ?segments .
?segments hasFeature DELAYED_RELEASE }
```



Select languages that have a class of segments

```
SELECT ?languages
WHERE { ?languages hasSegment ?segments .
?segments hasFeature DELAYED_RELEASE }
```



Investigating phonological universals

- Hyman (2008): Every phonological system has...
  - stops
  - at least one unrounded vowel
  - at least one front vowel or the palatal glide /j/
  - coronal phoneme(s)

Every phonological system has stops

```
SELECT ?languages
WHERE {
?languages phoible:hasSegment ?segments .
?segments phoible:notHasFeature feature:DELAYED_RELEASE
}
```

Phonological system with at least one unrounded vowel – and what are those languages and their segments?

```
SELECT ?languages ?segments
WHERE {
?languages phoible:hasSegment ?segments .
?segments phoible:hasFeature feature:SYLLABIC .
?segments phoible:notHasFeature feature:CONSONANTAL .
?segments phoible:notHasFeature feature:ROUND
}
```

Every phonological system has at least one front vowel or the palatal glide /j/

```
SELECT ?languages
WHERE {
?languages phoible:hasSegment ?segments .
?segments phoible:hasFeature feature:SYLLABIC .
?segments phoible:notHasFeature feature:CONSONANTAL .
?segments phoible:hasFeature feature:FRONT .
UNION { ?languages phoible:hasSegment segment:j }
}
```

Every phonological system has coronal phonemes... nope!

```
SELECT ?languages
WHERE {
?languages phoible:hasSegment ?segments .
?segments phoible:hasFeature feature:CORONAL
}
```

- "Another Universal Bites the Dust" (Blevins, 2009)
- Northwest Mekeo [mek] /p, β, m, w, g, ŋ, j, i, e, a, o, u/

Linguistic challenges

- Which language is this? ("A Grammar of Haida" - Northern? Southern?)
- Different theoretical models are used in describing languages
- Diacritic ordering
  - creaky voiced syllabic dental nasal: ñ̪̰ː
  - labialized aspirated long alveolar plosive: t^{wh}ː

Different analyses – same language

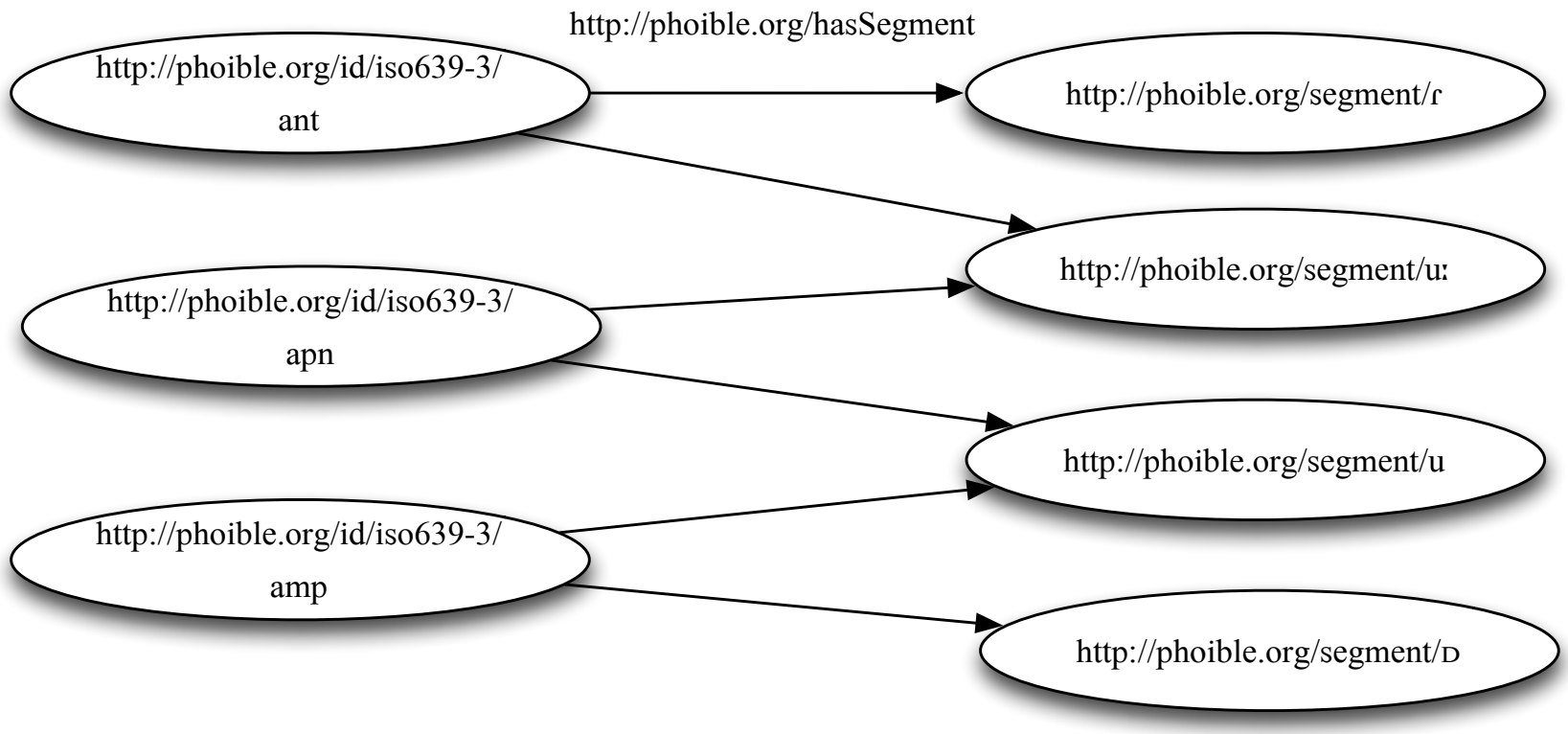
Table 2.1: Analyses of vertical vowel system of Kabardian (Hyman 2008:99)

Ladefoged & Maddieson 1996	/i ə a/
Halle 1970	/ə a/
Anderson 1978	/a/
Kuipers 1960	No vowels

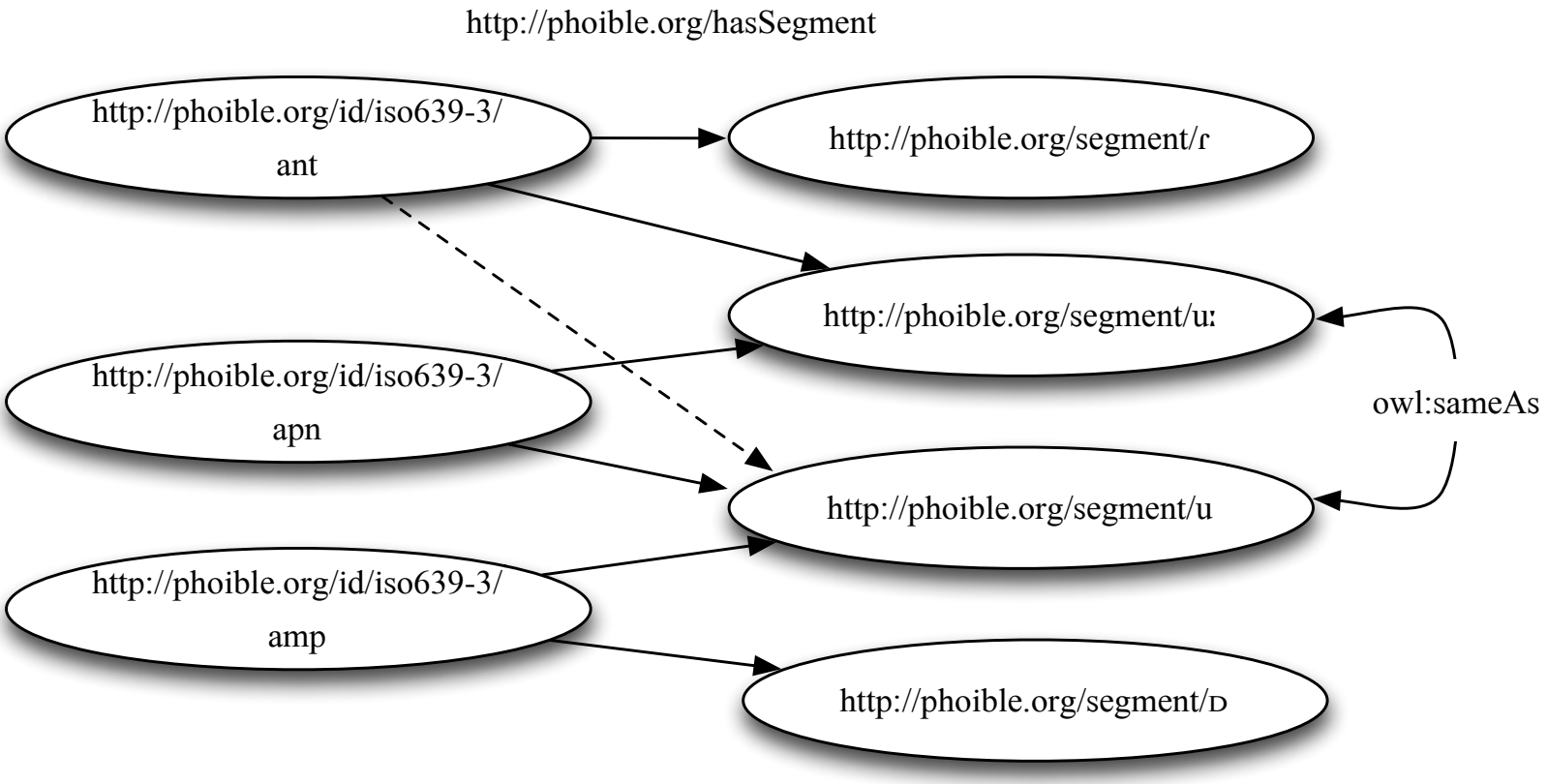
Using inference

- RDF model we can add additional knowledge **to the model**
- Change knowledge base without changing our query
- Logically-defined properties in OWL; merge OWL and RDF graphs
- Establish relationships between resources as inferred by a reasoner
- Reasoner evaluates logic statements in graph and adds inferred triples
- Ability to manipulate the ontology and to specify how to derive logical consequences and to create new entailments

Using OWL logic to extend the knowledge base



Using OWL logic to extend the knowledge base



Computational challenges

- Adherence to Unicode IPA:
  - g/g, !/!, a/ɑ, p/p
- Rendering sequences of Unicode characters as the same segment

ã	ã
U+0061 + U+0330 + U+0303	U+0061 + U+0303 + U+0330
LATIN SMALL LETTER A + COMBINING TILDE BELOW + COMBINING TILDE	LATIN SMALL LETTER A + COMBINING TILDE + COMBINING TILDE BELOW

Metadata

- DCMI RDF gets you most of OLAC
- Two big things missing: resource type and language identification
  - http://www.sil.org/iso639-3/documentation.asp?id=aar
  - http://www.ethnologue.com/show_language.asp?code=aar
  - http://lexvo.org/id/iso639-3/aar
  - http://wals.info/languoid/by_code/iso_639_3_aar
  - http://phoible.org/id/iso639-3/aar
  - http://resource dc:subject GOLD:Language
  - http://glottolog.livingsources.org/languoid/id/25785.xhtml

Poornima & Good 2010

```
<lego:hasCounterpart>
    <gold:LinguisticSign rdf:about=
     "http://www.purl.org/linguistics/North_Asmat_Voorhoeve/13">
      <gold:inLanguage>
       <gold:Language rdf:about=
       "http://www.sil.org/ISO639-3/documentation.asp?id=nks"/>
      </gold:inLanguage>
      <gold:hasForm>
       <gold:formUnit>
         <gold:stringRep>afak</gold:stringRep>
       </gold:formUnit>
      </gold:hasForm>
      <lego:hasSource>Voorhoeve 1980</lego:hasSource>
    </gold:LinguisticSign>
</lego:hasCounterpart>
```

Other questions moving forward...

- How do we access the Linguistic Linked Open Data cloud?
  - Download the RDF/OWL files and run locally?
  - In our code and point to the files online?
- What about a publicly accessible interface?
- SPARQL endpoint to make the LLOD data widely available

Summary

- PHOIBLE provides a large sample of segment inventories and additional phonological information from the world's languages
- It uses RDF and OWL graph data structures to capture knowledge about segments and distinctive features
- Can be used to ask questions of phonological systems... and more

Many thanks to the participants and organizers of LDL

And... Emily Bender, Michael Cysouw, Morgana Davids, Scott Drellishak, Shauna Eggers, David Ellison, Scott Farrar, Christopher Green, Sharon Hargus, Richard John Harvey, Jeff Good, Kelley Kilanski, William Lewis, Michael McAuliffe, Dan McCloy, Brandon Plasters, Tristan Purvis, Cameron Rule, Daniel Smith, Daniel Veja & Richard Wright