Problem Space
How do we leverage Web scale linguistic data and make it interoperable for phonetic data?
How do we query across such a resource?
What other features are desirable?
How do we merge cyberinfrastructure efforts in linguistics with the larger (semantic) Web?

Linguistic challenges for interoperation
Orthographies may differ graphemically for same phoneme
Transcription systems and orthographies may have complex grapheme-to-phoneme relationships
Varying levels of phonological abstraction and ambiguity

Linguistic definitions
Transcription system: system of symbols and rules for graphically transcribing the sounds of a language variety
Practical orthography: phonemic writing system for competent speakers
Orthography: specifies the symbols, punctuation, and the rules for a standard written language all orthographies are language-specific
Practical orthographies and transcription systems are both kinds of writing systems
Writing system: symbolic system of visible or tactile signs that represents a language in a systematic way

Writing System
Logographic Syllabic Phonetic Figural

Script: a collection of symbols (or distinct marks) as employed by a writing system
Script often confused with writing system
-writing system may be written with different scripts (e.g. alphabet writing system can be written in Roman and Cyrillic scripts)

OATS
Knowledge base that supports interoperation over disparate transcription systems
-ontology to define the structure of stored data
-data
Facilitates resource discovery and intelligent search over linguistic data
Current knowledge base includes an ontological description of writing systems and specifies relations for mapping segments of transcription systems to IPA equivalents
OATS is an ontological description of writing systems, relations, and interlingual pivot
Leverages technologies for ontological description, query and multilingual character encoding
Implements several useful case studies

Interlingual pivot - IPA as interlingua
OATS uses the International Phonetic Alphabet as an interlingua (the "pivot")
-All elements of systems of transcription are mapped to IPA
-IPA is used as the interlingua pivot that provides the ability to query across all resources in the knowledge base
-IPA has broad coverage of the sounds of the world's languages
-Mainstream adoption for transcription by linguists
-Encoded (at least mostly) in Unicode

Case studies implemented with OATS
-Querying disparate linguistic data via IPA
-Querying for phonetic data via orthography
-Error checking and data integrity
-Conversion

Query OATS knowledge base via IPA
-example: query for voiced palatal nasal /ɲ/ results: accounted for in 136 of 203 languages (67%)
Graphemes Languages % of Data
<ɲ> 114 84%
<ɲ> 11 8%
<p> 8 6%
<r> 2 1%
<n> 1 5%

Querying for phonemic data via orthography
-example: query "<ky>" which represents the voiceless alveo-palatal affricate /ʃ/ in Sisala [sɔlɔ]
-results: 98 languages have /ʃ/ written in 7 different ways
Graphemes Languages % of Data
<ʃ> 60 62%
<ʃ> 29 29%
<ʃ> 3 3%
<ʃ> 2 2%
<ʃ> 2 2%
<ʃ> 1 1%
<ʃ> 1 1%

To solve the problem we need...
Abstract modeling of linguistic structures
Precise definitions of linguistic concepts
Explicit encoding of concepts related to writing systems
Explicit encoding of these entities and their relationships
Absorptions that allow computers to process and encode data

Technological challenges for interoperation
Encoding multilingual text in an interoperable format
Defining linguistic entities and their relationships logically
Computationally tractable data structures
Resolving ambiguity between mapping relations

Linguistic definitions
Grapheme: abstract representation of a symbol employed by a writing system
-grapheme is a minimal distinctive symbol of a writing system
-contrastive graphical unit
-multiple graphemes may represent a single phoneme (digraph)
-<ck> and <ch> in English represent one phoneme <ːk>
-a single grapheme may represent two or more phonemes
-<sh> in English is a combination of the phonemes /ʃ/ and /s/:
Scripteme: grapheme within a writing system with particular semantics (i.e. pronunciation)
-definition of scripteme is necessary for interoperability
-graphemes may be homophonic across scripts and languages
-semantics of a grapheme is dependent on the writing system using it
-Russian <!<p> is a dental or alveolar trill /r/
-English <p> is a voiceless bilabial stop /p/
-provides a level for mapping a writing system specific grapheme to the phonological level
-allows the same grapheme to represent different sounds across different transcription and writing systems

OATS data model - an ontology
Formalized relationships that hold between entities in a given domain
-low level ontology: taxonomy and a set of inference rules
-higher-level ontology: collections of info that have formalized relationships that hold between entities in a given domain
Provides the basis for automated reasoning by computer software
Content is given meaning in the sense of interpreting data and disambiguating entities

Example graph (there are over 100,000 triples in the knowledge base):

Technologies used in OATS
RDF (Resource Description Framework)
-subject-predicate-object (or entity-relationship-entity)
-identifies unique resources through Uniform Resource Identifiers (URIs)
-triples form a graph data structure of nodes and arcs
-non-hierarchical and can be complexly connected
XML http://www.w3.org/XML
Unicon Standard http://uniconcode.org/
Python http://python.org/
RDFLib http://rdflib.net/
Protege http://protege.stanford.edu/
SPARQL http://www.w3.org/TR/rdf-sparql-query/

Data sources
OATS includes over 203 languages from 23 language families
-for a full list go to http://phoible.org
Data sources include
-Alphabets des langues africaines (Hartell 1993)
-Max Planck Intercontinental Dictionary Series
Additional languages added by hand

Complex query via IPA
-relationships are encoded by ontological structure
-‘which languages have the phoneme /gb/ and of those languages which lack its voiceless counterpart /kp/’

Code Language Name Genetic Affiliation
enk Mankankan Mande
kza Karaboro Gur
lia Limba Atlantic
mif Mofis-Gudur Chadic
sld Sisala Gur
ssl Sisala Gur
sus Sasa Mande
ted Krumen Kru
tem Themne Atlantic