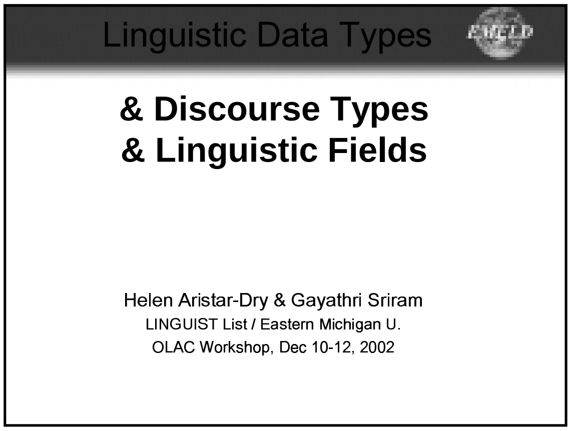
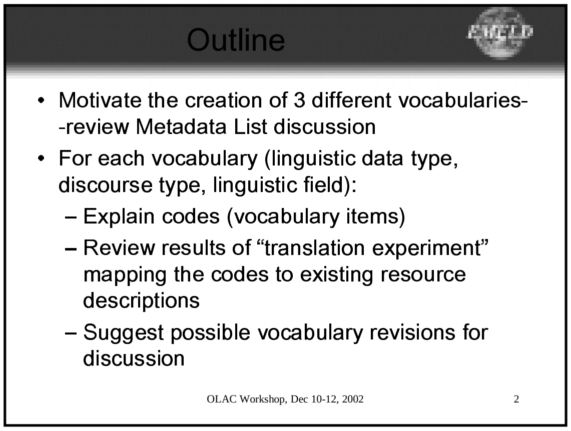


Linguistic Data Types

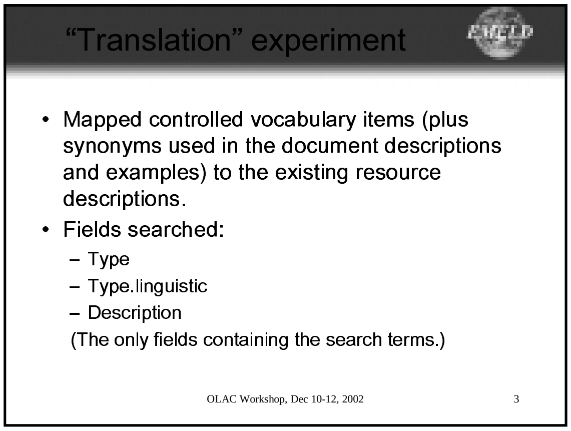
& Discourse Types
& Linguistic Fields

Helen Aristar-Dry & Gayathri Sriram
LINGUIST List / Eastern Michigan U.
OLAC Workshop, Dec 10-12, 2002



Outline

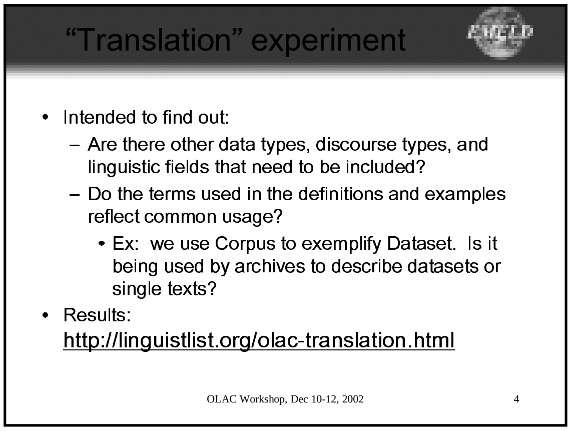
- Motivate the creation of 3 different vocabularies--review Metadata List discussion
- For each vocabulary (linguistic data type, discourse type, linguistic field):
  - Explain codes (vocabulary items)
  - Review results of "translation experiment" mapping the codes to existing resource descriptions
  - Suggest possible vocabulary revisions for discussion



"Translation" experiment

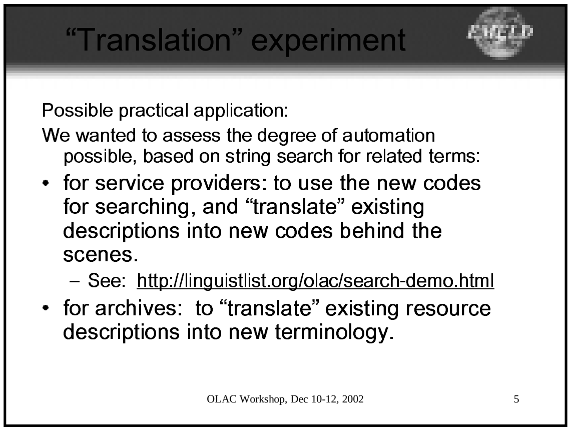
- Mapped controlled vocabulary items (plus synonyms used in the document descriptions and examples) to the existing resource descriptions.
- Fields searched:
  - Type
  - Type.linguistic
  - Description

  (The only fields containing the search terms.)



"Translation" experiment

- Intended to find out:
  - Are there other data types, discourse types, and linguistic fields that need to be included?
  - Do the terms used in the definitions and examples reflect common usage?
    - Ex: we use Corpus to exemplify Dataset. Is it being used by archives to describe datasets or single texts?
- Results:
  http://linguistlist.org/olac-translation.html

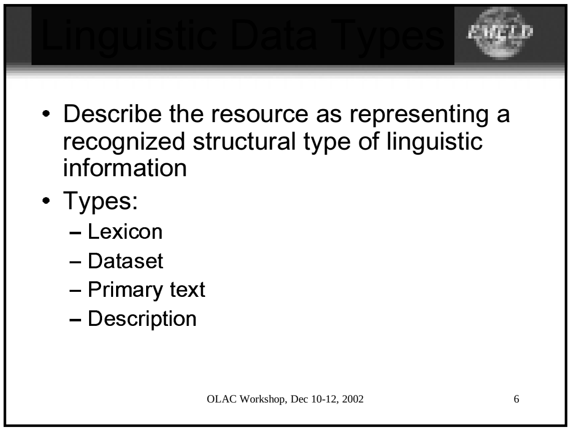


"Translation" experiment

Possible practical application:

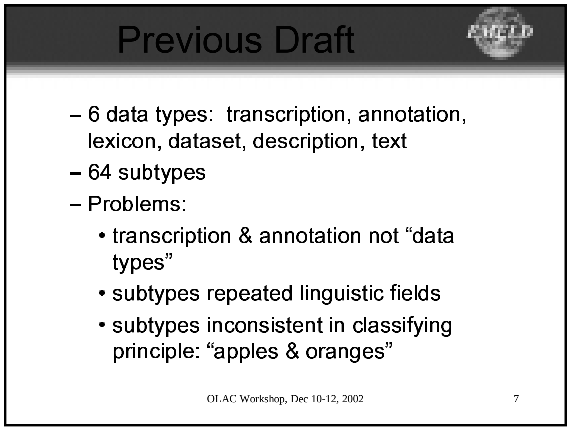
We wanted to assess the degree of automation possible, based on string search for related terms:

- for service providers: to use the new codes for searching, and "translate" existing descriptions into new codes behind the scenes.
  - See: http://linguistlist.org/olac/search-demo.html
- for archives: to "translate" existing resource descriptions into new terminology.



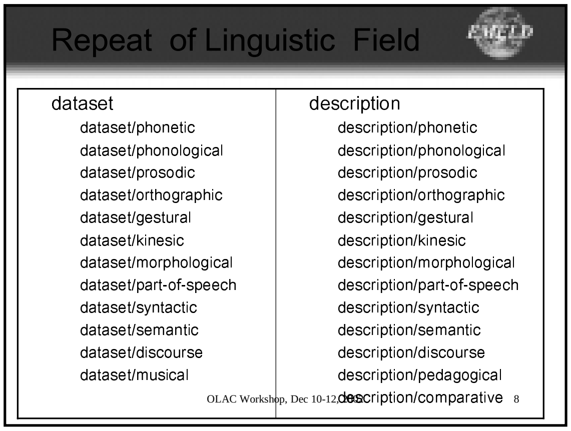
Linguistic Data Types

- Describe the resource as representing a recognized structural type of linguistic information
- Types:
  - Lexicon
  - Dataset
  - Primary text
  - Description



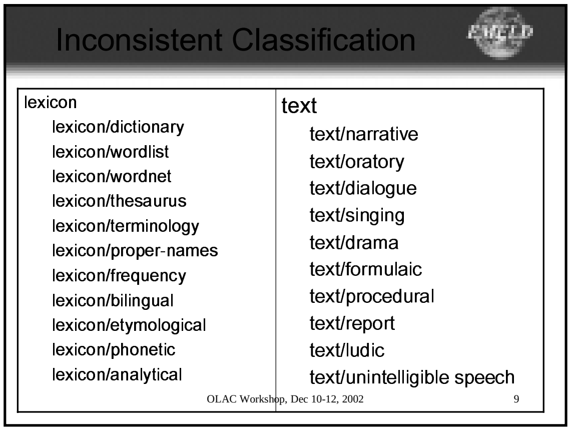
Previous Draft

- 6 data types: transcription, annotation, lexicon, dataset, description, text
- 64 subtypes
- Problems:
  - transcription & annotation not "data types"
  - subtypes repeated linguistic fields
  - subtypes inconsistent in classifying principle: "apples & oranges"



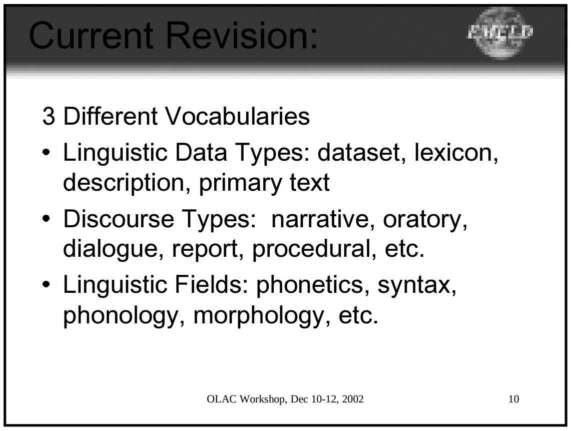
Repeat of Linguistic Field

dataset	description
dataset/phonetic	description/phonetic
dataset/phonological	description/phonological
dataset/prosodic	description/prosodic
dataset/orthographic	description/orthographic
dataset/gestural	description/gestural
dataset/kinesic	description/kinesic
dataset/morphological	description/morphological
dataset/part-of-speech	description/part-of-speech
dataset/syntactic	description/syntactic
dataset/semantic	description/semantic
dataset/discourse	description/discourse
dataset/musical	description/pedagogical
	description/comparative



Inconsistent Classification

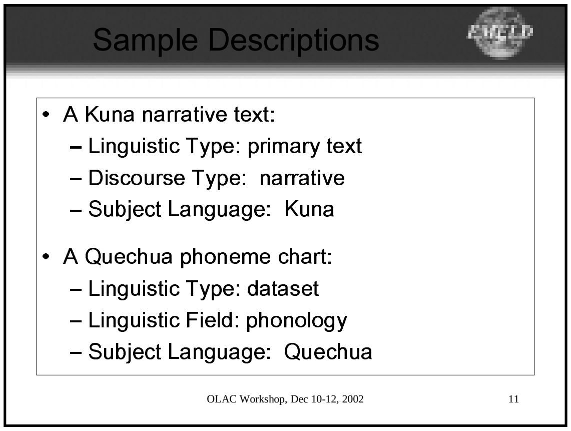
lexicon	text
lexicon/dictionary	text/narrative
lexicon/wordlist	text/oratory
lexicon/wordnet	text/dialogue
lexicon/thesaurus	text/singing
lexicon/terminology	text/drama
lexicon/proper-names	text/formulaic
lexicon/frequency	text/procedural
lexicon/bilingual	text/report
lexicon/etymological	text/ludic
lexicon/phonetic	text/unintelligible speech
lexicon/analytical	



Current Revision:

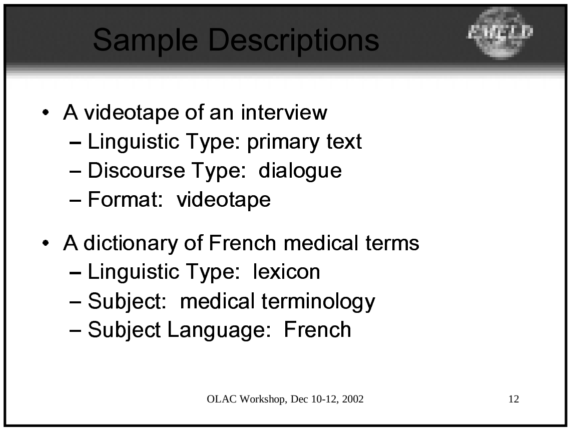
3 Different Vocabularies

- Linguistic Data Types: dataset, lexicon, description, primary text
- Discourse Types: narrative, oratory, dialogue, report, procedural, etc.
- Linguistic Fields: phonetics, syntax, phonology, morphology, etc.



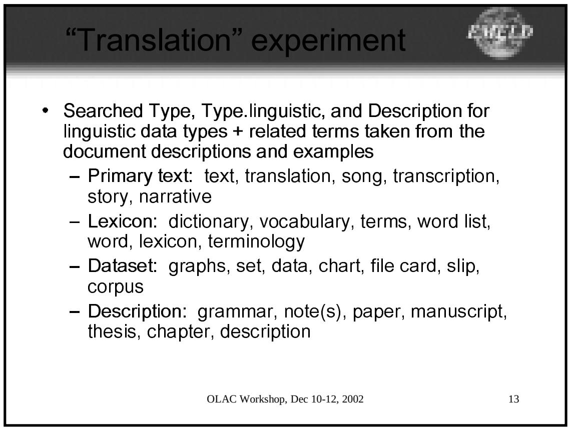
Sample Descriptions

- A Kuna narrative text:
  - Linguistic Type: primary text
  - Discourse Type: narrative
  - Subject Language: Kuna
- A Quechua phoneme chart:
  - Linguistic Type: dataset
  - Linguistic Field: phonology
  - Subject Language: Quechua



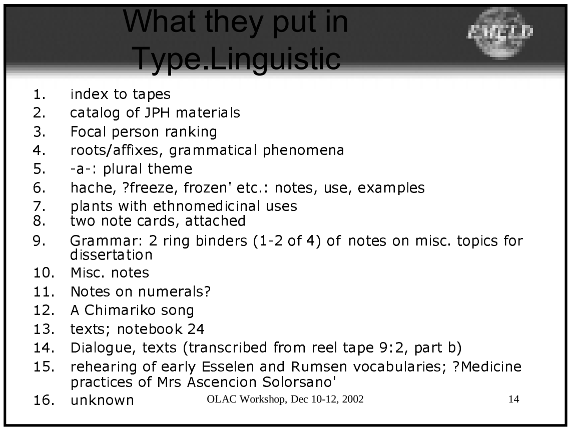
Sample Descriptions

- A videotape of an interview
  - Linguistic Type: primary text
  - Discourse Type: dialogue
  - Format: videotape
- A dictionary of French medical terms
  - Linguistic Type: lexicon
  - Subject: medical terminology
  - Subject Language: French



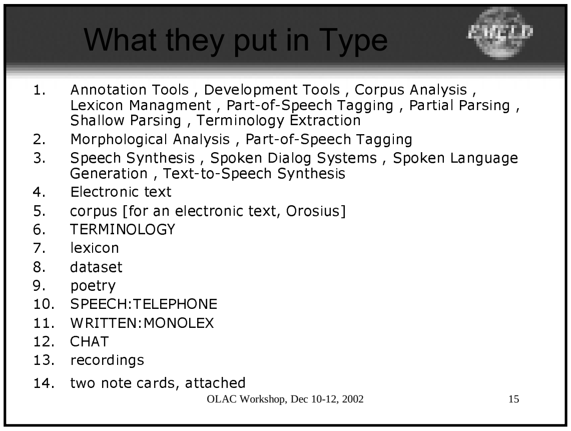
“Translation” experiment

- Searched Type, Type.linguistic, and Description for linguistic data types + related terms taken from the document descriptions and examples
  - Primary text: text, translation, song, transcription, story, narrative
  - Lexicon: dictionary, vocabulary, terms, word list, word, lexicon, terminology
  - Dataset: graphs, set, data, chart, file card, slip, corpus
  - Description: grammar, note(s), paper, manuscript, thesis, chapter, description



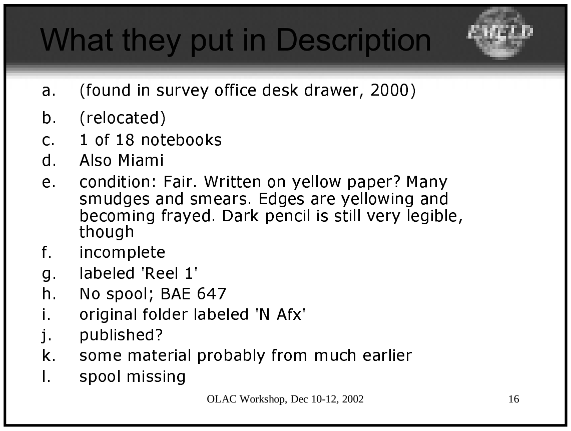
What they put in Type.Linguistic

1. index to tapes
2. catalog of JPH materials
3. Focal person ranking
4. roots/affixes, grammatical phenomena
5. -a-: plural theme
6. hache, ?freeze, frozen' etc.: notes, use, examples
7. plants with ethnomedicinal uses
8. two note cards, attached
9. Grammar: 2 ring binders (1-2 of 4) of notes on misc. topics for dissertation
10. Misc. notes
11. Notes on numerals?
12. A Chimariko song
13. texts; notebook 24
14. Dialogue, texts (transcribed from reel tape 9:2, part b)
15. rehearing of early Esselen and Rumsen vocabularies; ?Medicine practices of Mrs Ascencion Solorsano'
16. unknown



What they put in Type

1. Annotation Tools , Development Tools , Corpus Analysis , Lexicon Managment , Part-of-Speech Tagging , Partial Parsing , Shallow Parsing , Terminology Extraction
2. Morphological Analysis , Part-of-Speech Tagging
3. Speech Synthesis , Spoken Dialog Systems , Spoken Language Generation , Text-to-Speech Synthesis
4. Electronic text
5. corpus [for an electronic text, Orosius]
6. TERMINOLOGY
7. lexicon
8. dataset
9. poetry
10. SPEECH:TELEPHONE
11. WRITTEN:MONOLEX
12. CHAT
13. recordings
14. two note cards, attached



What they put in Description

a. (found in survey office desk drawer, 2000)

b. (relocated)

c. 1 of 18 notebooks

d. Also Miami

e. condition: Fair. Written on yellow paper? Many smudges and smears. Edges are yellowing and becoming frayed. Dark pencil is still very legible, though

f. incomplete

g. labeled 'Reel 1'

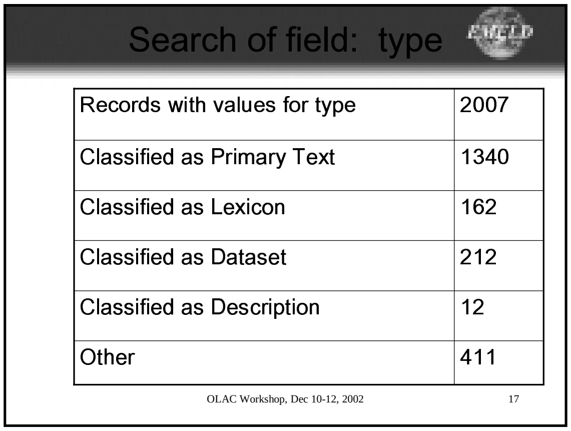
h. No spool; BAE 647

i. original folder labeled 'N Afx'

j. published?

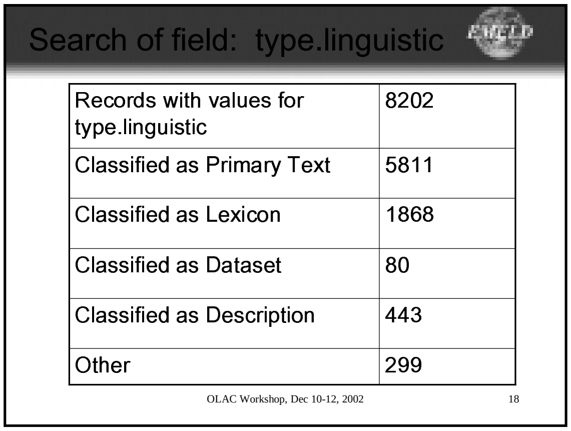
k. some material probably from much earlier

l. spool missing



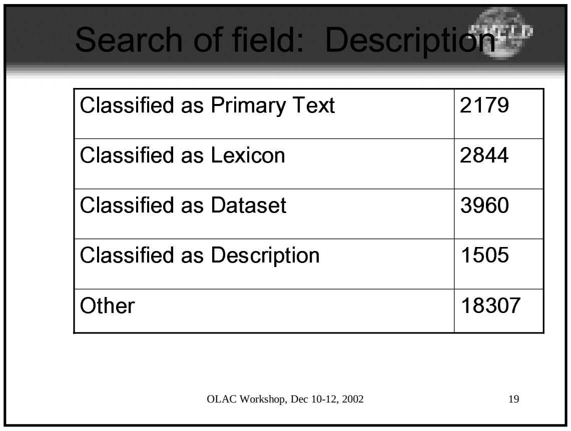
Search of field: type

Records with values for type	2007
Classified as Primary Text	1340
Classified as Lexicon	162
Classified as Dataset	212
Classified as Description	12
Other	411



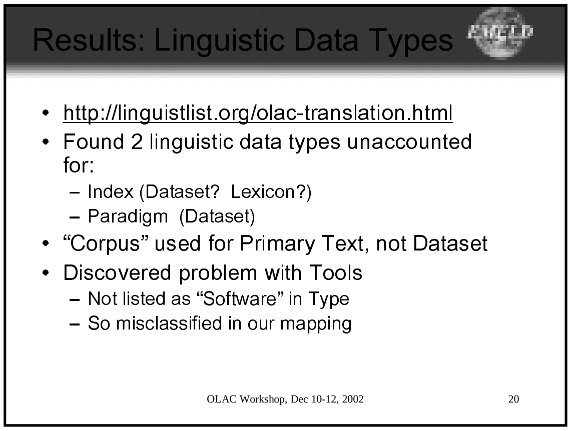
Search of field: type.linguistic

Records with values for type.linguistic	8202
Classified as Primary Text	5811
Classified as Lexicon	1868
Classified as Dataset	80
Classified as Description	443
Other	299



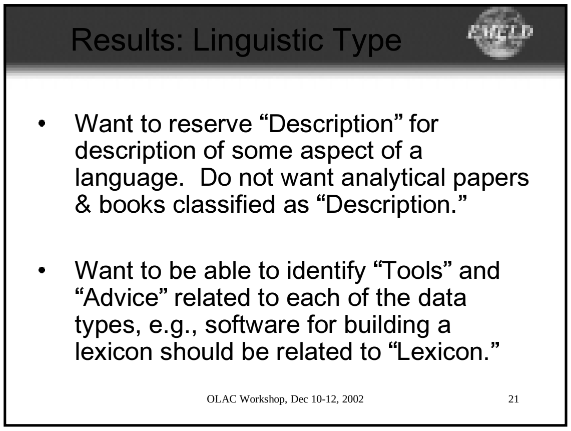
Search of field: Description

Classified as Primary Text	2179
Classified as Lexicon	2844
Classified as Dataset	3960
Classified as Description	1505
Other	18307



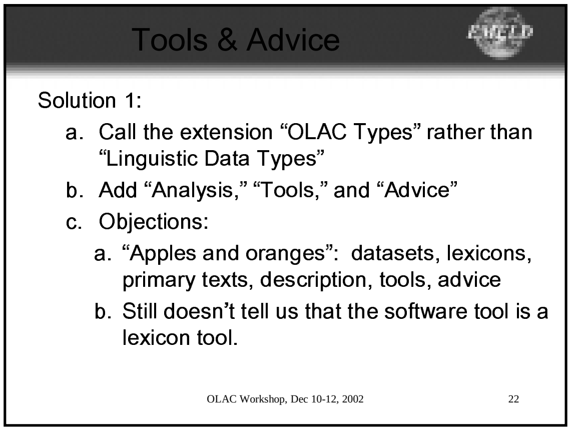
Results: Linguistic Data Types

- http://linguistlist.org/olac-translation.html
- Found 2 linguistic data types unaccounted for:
  - Index (Dataset? Lexicon?)
  - Paradigm (Dataset)
- "Corpus" used for Primary Text, not Dataset
- Discovered problem with Tools
  - Not listed as "Software" in Type
  - So misclassified in our mapping



Results: Linguistic Type

- Want to reserve "Description" for description of some aspect of a language. Do not want analytical papers & books classified as "Description."
- Want to be able to identify "Tools" and "Advice" related to each of the data types, e.g., software for building a lexicon should be related to "Lexicon."



Tools & Advice

Solution 1:

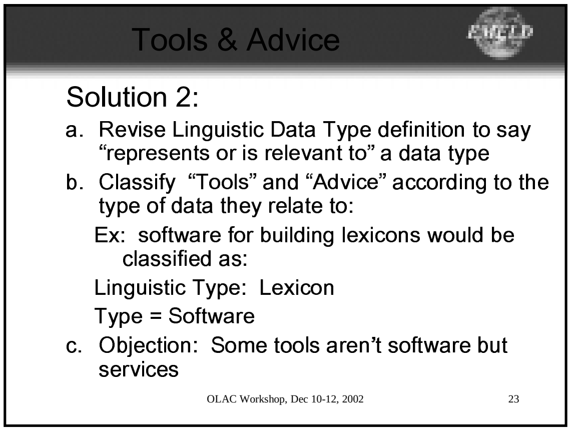
a. Call the extension "OLAC Types" rather than "Linguistic Data Types"

b. Add "Analysis," "Tools," and "Advice"

c. Objections:

   a. "Apples and oranges": datasets, lexicons, primary texts, description, tools, advice

   b. Still doesn't tell us that the software tool is a lexicon tool.



Tools & Advice

Solution 2:

a. Revise Linguistic Data Type definition to say "represents or is relevant to" a data type

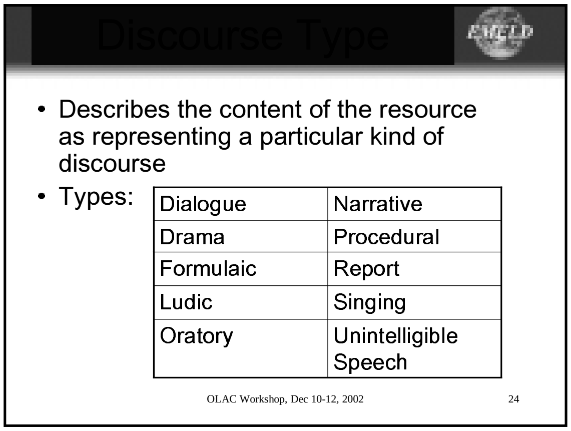
b. Classify "Tools" and "Advice" according to the type of data they relate to:

   Ex: software for building lexicons would be classified as:

   Linguistic Type: Lexicon

   Type = Software

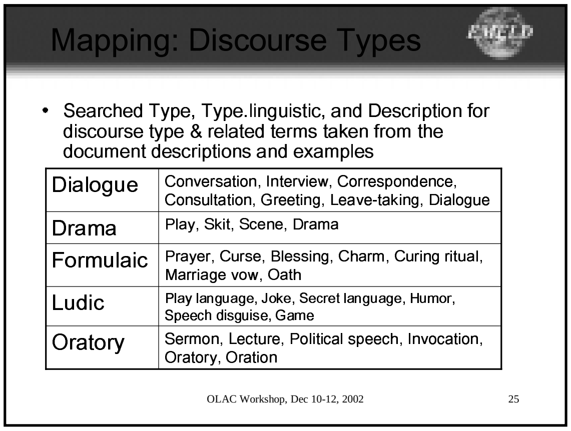
c. Objection: Some tools aren't software but services



Discourse Type

- Describes the content of the resource as representing a particular kind of discourse
- Types:

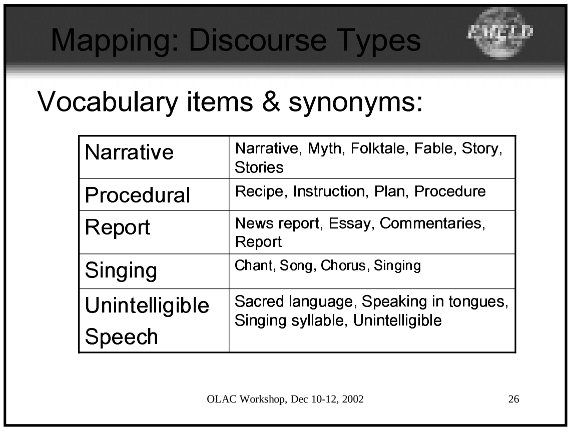
Dialogue	Narrative
Drama	Procedural
Formulaic	Report
Ludic	Singing
Oratory	Unintelligible Speech



Mapping: Discourse Types

- Searched Type, Type.linguistic, and Description for discourse type & related terms taken from the document descriptions and examples

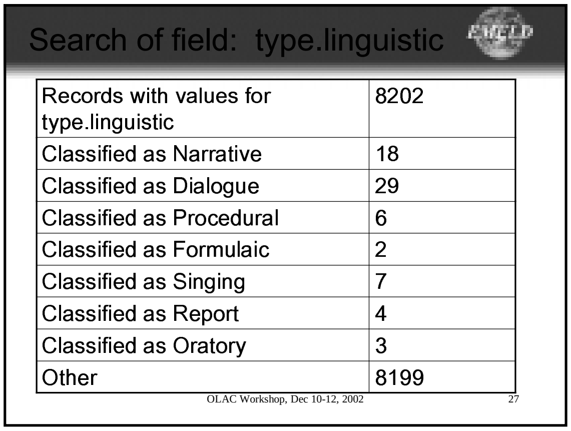
Dialogue	Conversation, Interview, Correspondence, Consultation, Greeting, Leave-taking, Dialogue
Drama	Play, Skit, Scene, Drama
Formulaic	Prayer, Curse, Blessing, Charm, Curing ritual, Marriage vow, Oath
Ludic	Play language, Joke, Secret language, Humor, Speech disguise, Game
Oratory	Sermon, Lecture, Political speech, Invocation, Oratory, Oration



Mapping: Discourse Types

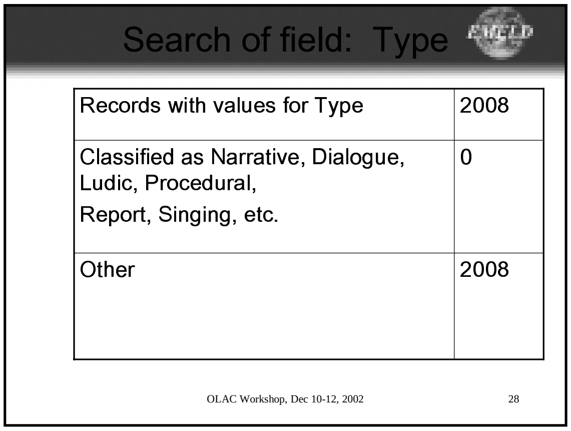
Vocabulary items & synonyms:

Narrative	Narrative, Myth, Folktale, Fable, Story, Stories
Procedural	Recipe, Instruction, Plan, Procedure
Report	News report, Essay, Commentaries, Report
Singing	Chant, Song, Chorus, Singing
Unintelligible Speech	Sacred language, Speaking in tongues, Singing syllable, Unintelligible



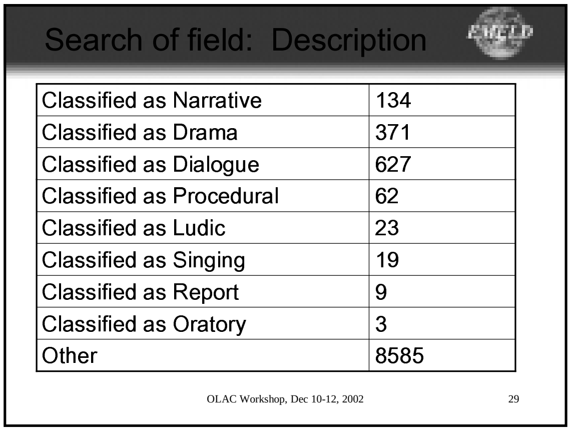
Search of field: type.linguistic

Records with values for type.linguistic	8202
Classified as Narrative	18
Classified as Dialogue	29
Classified as Procedural	6
Classified as Formulaic	2
Classified as Singing	7
Classified as Report	4
Classified as Oratory	3
Other	8199



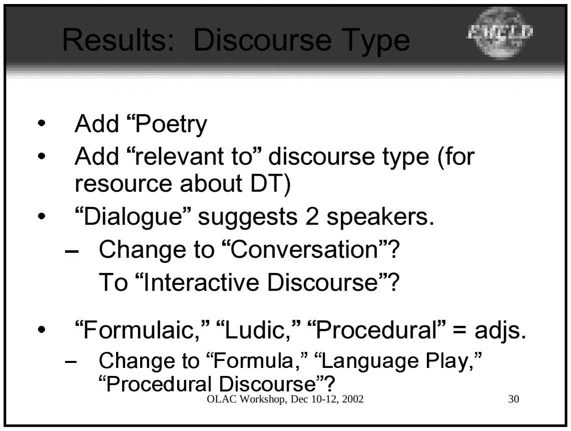
Search of field: Type

Records with values for Type	2008
Classified as Narrative, Dialogue, Ludic, Procedural, Report, Singing, etc.	0
Other	2008



Search of field: Description

Classified as Narrative	134
Classified as Drama	371
Classified as Dialogue	627
Classified as Procedural	62
Classified as Ludic	23
Classified as Singing	19
Classified as Report	9
Classified as Oratory	3
Other	8585



Results: Discourse Type

- Add “Poetry
- Add “relevant to” discourse type (for resource about DT)
- “Dialogue” suggests 2 speakers.
  - Change to “Conversation”?
    To “Interactive Discourse”?
- “Formulaic,” “Ludic,” “Procedural” = adjs.
  - Change to “Formula,” “Language Play,” “Procedural Discourse”?

Linguistic Field

- Describes the resource as relevant to a particular subfield of linguistic science
- Fields:
  - anthropological linguistics
  - applied linguistics
  - cognitive science
  - computational linguistics
  - discourse analysis
  - general linguistics
  - historical linguistics
  - history of linguistics

Linguistic Field

- Fields (cont):
  - Language Description
  - Lexicography
  - Linguistics and literature
  - Linguistic theories
  - Morphology
  - Neurolinguistics
  - Philosophy of science
  - Phonetics
  - Phonology
  - Pragmatics

Linguistic Field

- Fields (cont):
  - Psycholinguistics
  - Semantics
  - Sociolinguistics
  - Syntax
  - Text and corpus linguistics
  - Translation
  - Typology
  - Writing systems

Results:The the The if the Linguistic Field

- Add “Language Acquisition”?
  - Definition: The study of the process of acquiring human language.
  - Comment: Language Acquisition may be used to describe materials relating to either adult or child language acquisition, and to either first or later language acquisition. However, if the materials deal specifically with language teaching, or with the process of language learning from a pedagogical point of view, they may be best classified as Applied Linguistics.
  - Examples: Studies of first language acquisition, audio or video tapes of language acquisition experiments, and guides to experimental techniques in eliciting acquisition data.

Problems w/ Linguistic Field

- Add “Forensic Linguistics”?
  - Definition: Applications of linguistic science to the domain of law
  - Comment: Forensic linguistics refers to the use of linguistic methodology to make legal determinations. Analyses of courtroom language are best classified as Discourse Analysis.
  - Examples: Papers on issues in dispute in court cases, e.g., authorship identification, assessment of ambiguity in texts, voice attribution.

Search for Linguistic Fields

Demo page:

http://linguistlist.org/olac/search-demo.html