Evaluation Initiatives:

US Activities in HLT Evaluation

Christopher Cieri
ccieri@ldc.upenn.edu

University of Pennsylvania
Linguistic Data Consortium and Department of Linguistics
3600 Market Street, Philadelphia, PA 19104 U.S.A.

www.ldc.upenn.edu

Recent Activities

- TIDES: Translingual Information Detection, Extraction and Summarization
  - Goal: build **technologies** to support a news understanding system that accepts queries in user’s language, finds relevant spoken or written documents in many other languages, extracts important information, translates and summarizes.
  - English, Mandarin Chinese, Modern Standard Arabic with *surprises* in Cebuano, Hindi
- EARS
  - Goal: improve speed, accuracy and readability/usability of STT technologies
  - MDE: tag SUs, disfluency for presentation, further processing
  - English, Mandarin Chinese, Egyptian and Levantine Colloquial Arabic

GALE

- Global Autonomous Language Exploitation
  - Technologies to absorb, analyze, interpret huge volumes of speech, text in multiple languages. Engines convert, distill data, delivering pertinent, consolidated information in easy-to-understand forms.
- Engines
  - Transcription: Multilingual Speech > English Text (includes translation)
  - Translation: Multilingual Text > English Text
  - Distillation: integrates information of interest to user from multiple sources and documents
- Data Types
  - newswire, newsgroups, blogs,
  - broadcast news, talk shows, conversations
- Annotations
  - Transcripts with some Arabic vowelized
  - Translations with some aligned at word level
  - Treebanks, Propbanks
  - “Distillation” Annotation
- New approach to evaluation post-facto, scenarios based
- New MT evaluation metric based on post-editing and edit distance
- NIST MT, RT evaluations continue independently, allow non-US sites.

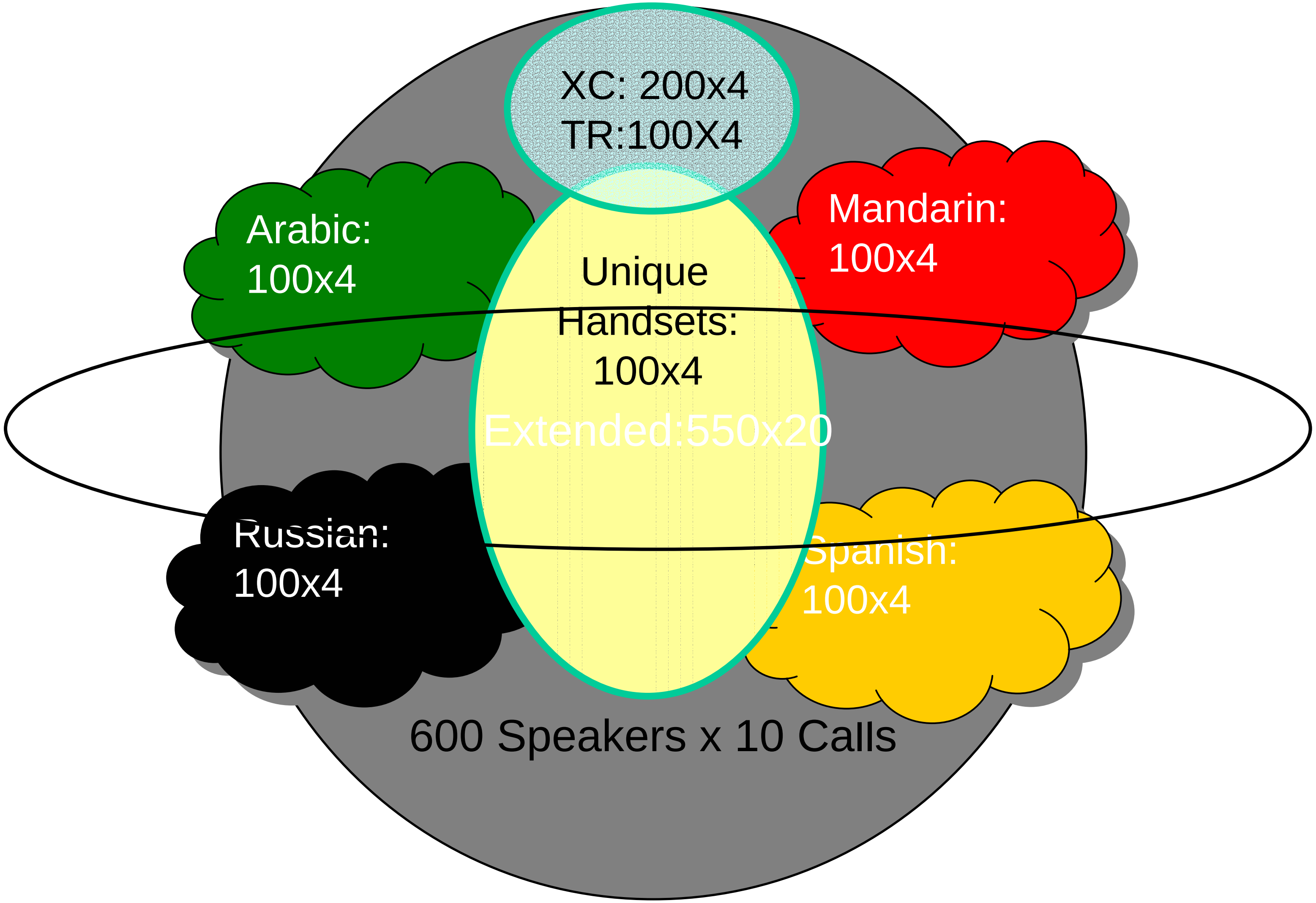
DOI/REFLEX

- Research on English & Foreign Language Processing
  - Machine Translation, Content Extraction (ACE)
  - Less Commonly Taught Languages
- LCTL => 1M native speakers, scarce resources
  - Bengali, Punjabi, Hungarian, Tagalog, Tamil, Thai, Yoruba, Urdu, Uzbek
- Resources
  - Monolingual News Text
  - LCTL -> Eng Parallel Text
  - Eng -> LCTL News, Elicitation, Conversation
  - Lexicon
  - Encoding Converter
  - Word, Sentence Segmenters
  - POS Tagset, Tagger, Tagged Text
  - Morphological Analyzer, Morphologically Analyzed Text
  - Named Entity Tagged Text, Tagger
  - Name Transliterator
  - Grammatical Sketch

LVD ID

- Supports Language ID Research, Technology Development & Evaluation
- Language Variety and Dialect ID
  - English: British, Indian, Australian/New Zealand and American English
  - Chinese: Standard Mandarin, Shanghai Wu, Southern Min/Taiwanese Cantonese (Guangdong Hua/Yue Yu)
  - Aceh, Amharic, Bengali, Burmese, Chechen, Farsi, Georgian, Guarani, Hindi, Italian, Japanese, Khmer, Korean, Lao, Punjabi, Russian, Spanish, Tagalog, Tamil, Thai, Tigrigna, Urdu, Uzbek, Vietnamese
- Resource Sharing with Speaker Recognition
- Speaker Recognition
  - Increasing impact of forensic applications
  - Shift toward multilingual, and multichannel tasks
  - Mixer Corpora

Mixer Corpora Phases I & II



Mixer Corpora Yields

	Subjects	
	Targeted	Recorded
Base (x10 Calls)	650	1150
Arabic (x4 Calls)	100	129
Mandarin (x4 Calls)	100	115
Russian (x4 Calls)	100	107
Spanish (x4 Calls)	100	100
Extended (x20 Calls)	550	611
Super-Extended (x30 Calls)	0	382
Cross Channel (x4 Calls)	200	201
Transcript Reading	100	100