Core Linguistic Resources for the World's Languages

Christopher Cieri, Mike Maxwell, Stepanie Strassel

{ccieri,maxwell,strassel}@ldc.upenn.edu

University of Pennsylvania

Linguistic Data Consortium and Department of Linguistics

3615 Market Street, Philadelphia, PA 19104-2608 U.S.A.

www.ldc.upenn.edu

Scoping the Problem

- 6700 Languages (according to Ethnologue)
- Assume international consortia create complete LRs for 50 languages/year at $700K/language
- Bottom Line: $4.7B and 134 years
- More importantly, the process of building LRs changes with the size of the language, its history of literacy, etc.
- E.g.: raw text acquisition; only 1500 languages written
  - Electronic harvest
  - Scanning/keyboarding of written text
  - Paying native speakers to create original works
  - Designing an orthography, interviewing native speakers and transcribing
- The motivation for building LRs also changes with language
  - Culture & Folk medicine versus International Markets
  - Understanding remote points of view

Proposal Features

- **Design Core Project - must be possible**
  - **Require <= 5 years**
  - **Budget should be conceivable given our previous collective experience**
- **Manageable set of core languages**
  - **many speakers worldwide, local experts & native-speaker annotators**
  - **raw resources available on web**
- **Manageable set of core resources**
  - **text, parallel text, translation lexicon, entity tagging**
  - **grammatical sketch, tokenizer, morph-analyzer**
- **Publish to encourage extension**
  - **Language resources & metadata describing them**
  - **Corpus specifications & tools**
- **Coordinate work on LRs to minimize duplication of effort**
- **Promote the plan to**
  - **international coordinating bodies, national governments, commercial sponsors**
  - **researchers**

Pre-History

- 1983: Penn Language Analysis Center founded; builds textbases, bilingual dictionaries in 35 languages
- 1992: LDC founded to distribute LRs for many languages
- 1995: CALLHOME corpora for Large Volume Continuous Speech Recognition
  - 200 telephone conversations of 20-30 minutes
  - Complete transcripts
  - Pronouncing lexicon
  - English, Spanish, Mandarin, Egyptian Arabic, German, Japanese
- 1996: CALLFRIEND corpora for Language Identification
  - 200 telephone conversations of 20-30 minutes
  - American English (Southern&Non-), Canadian French, Egyptian Arabic, Farsi, German, Hindi, Japanese, Korean, Mandarin Chinese (Mainland & Taiwan), Spanish (Caribbean & Non-), Tamil, Vietnamese

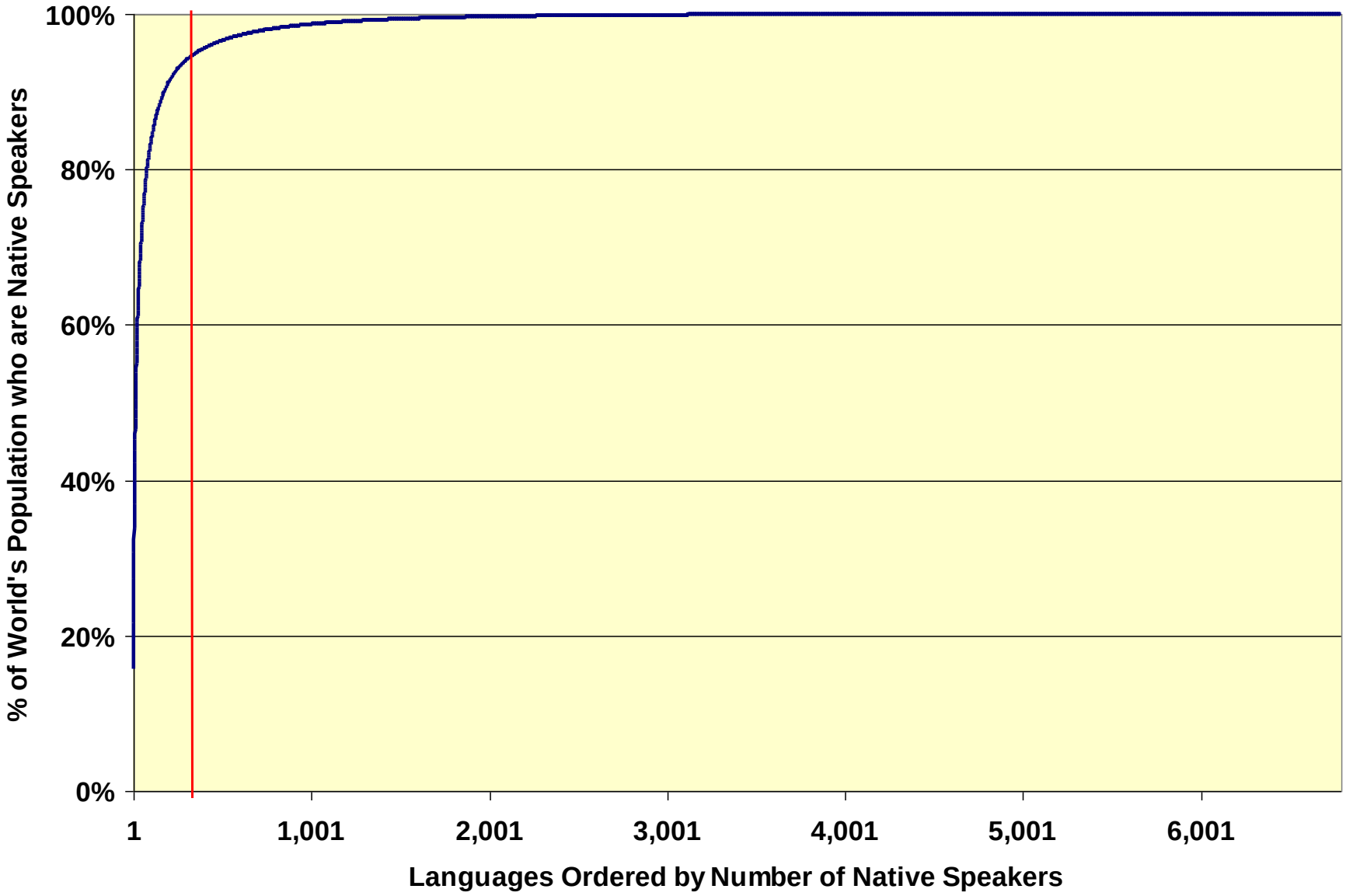
Recent History

- 1999: TIDES Planning begins
  - news understanding system for English speaking user
  - multilingual capabilities with rapid porting to new languages
- 1999: JHU Workshop on rapid development of statistical machine translation
- 2000: LDC completes 50 language TIDES VOA collection
- 2001: TIDES reorganized with 3 primary & 3 secondary languages
  - English, Mandarin, Arabic
  - Spanish, Japanese, Korean
- 2002: TIDES Surprise Language experiments announced; LDC begins resource survey in preparation
- 2002: ICWLR planning meeting
- 2003: Surprise Language experiments
  - Data collection dry run in Cebuano
  - Data collection, technology development and evaluation in Hindi

LR Survey

- **Preparation for TIDES Surprise Language Experiments**
  - **Given that LDC would have no prior knowledge of Surprise Language**
  - **And that, with the wrong choice, the experiment could become mired**
  - **LDC proposed the survey to inform program manager's choice**
  - **and to emphasize preparation over scramble**
  - **Survey avoids "gaming" experiment by permanently changing the landscape.**
- **Based upon Ethnologue**
- **Limited to languages with 1,000,000+ speakers**
- **Temporarily excluded "well studied" languages (Chinese, French)**
- **Excluded languages all of whose speakers also another language with greater number of speakers (Cajun English, Sicilian)**
- **Excluded languages that are not written.**
- **Performed triage on remaining languages**
  - **Developed decision tree where negative answers demote a language**
  - **Questions researched roughly in triage order**
- **Now have triage results for 150/320 languages**

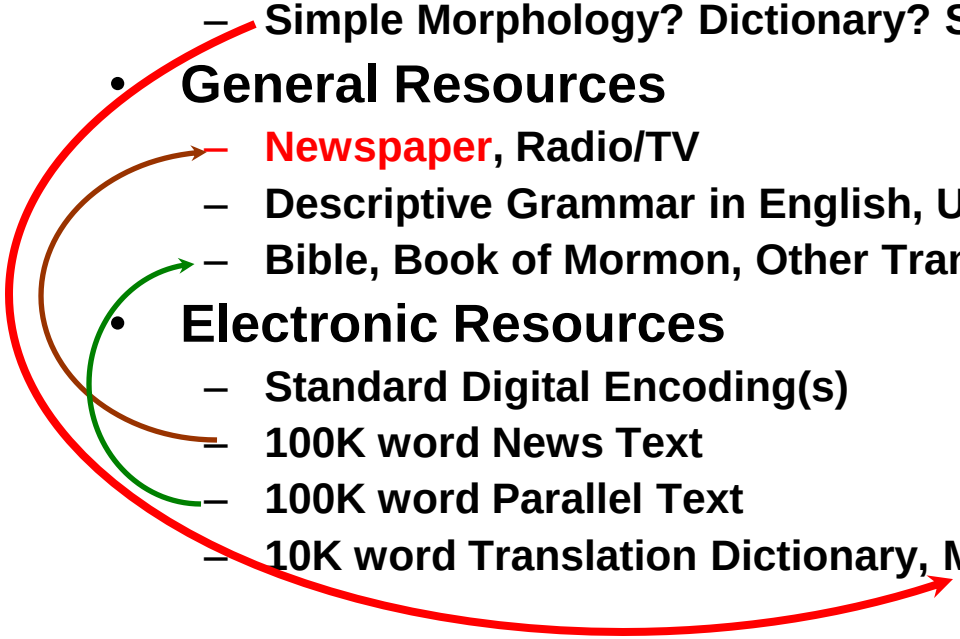
Languages/Speakers



- ELSNET, ENABLER, ICWLR 2003, Paris

Survey Questions

- Demographics
  - Language Name, SIL Code & Classification, Consider?
  - Primary Country, Other Countries where spoken
  - L1 Speakers Worldwide, % Who Speak Larger Language, Pivot
  - Speakers with Internet Access, Predicted Growth, Net Hosts
  - Is there a US Speaker Community? Literacy Rate? Students?
- Orthography
  - Language Written, Simple Orthography, Separate Sentences/Words
- Linguistic Structure
  - Simple Morphology? Dictionary? Special Considerations
- General Resources
  - Newspaper, Radio/TV
  - Descriptive Grammar in English, US Expert
  - Bible, Book of Mormon, Other Translations
- Electronic Resources
  - Standard Digital Encoding(s)
  - 100K word News Text
  - 100K word Parallel Text
  - 10K word Translation Dictionary, Morph Analyzer



Sample Summary

Summary contains decisions. Full report contains underlying data.

1	Language	Country	# Speakers	Written	Sentence_Punctuation	Words separated	News_Text	Newspaper	Parallel_Text	Bible	Xltn_Dictionary	Dictionary	Morphology	Morph_Analyzer	Sort order
11	FARSI_WESTERN	Iran	24,280,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
12	GREEK	Greece	12,000,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
13	HINDI	India	182,000,000	T	T	T	T	T	T	T	T	T	Q	T	1161105000
14	HUNGARIAN	Hungary	14,500,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
15	LATVIAN	Latvia	1,500,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
16	LITHUANIAN	Lithuania	4,000,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
17	ROMANIAN	Romania	26,000,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
18	RUSSIAN	Russia	170,000,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
19	TAMIL	India	63,075,000	T	T	T	T	T	T	T	T	T	F	T	1161105000
20	BENGALI	Bangladesh	189,000,000	T	T	T	T	T	T	T	T	T	Q	F	1161100000
21	HAUSA	Nigeria	24,200,000	T	T	T	T	T	T	T	T	T	F	F	1161100000
22	UKRAINIAN	Ukraine	41,000,000	T	T	T	T	T	T	T	T	T	F	F	1161100000
23	DUTCH	Netherlands	20,000,000	T	T	T	T	T	F	T	T	T	T	T	1161060000
24	INDONESIAN	Indonesia	17,050,000	T	T	T	T	T	Q	T	T	T	T	T	1161060000
25	MACEDONIAN	Macedonia	2,000,000	T	T	T	T	T	Q	T	T	T	T	F	1161060000
26	TAGALOG	Philippines	17,000,000	T	T	T	T	T	F	T	T	T	T	F	1161060000
27	VIETNAMESE	Viet Nam	67,662,000	T	T	T	T	T	Q	T	T	T	T	T	1161060000
28	ESTONIAN	Estonia	1,100,000	T	T	T	T	T	F	T	T	T	F	T	1161055000
29	HEBREW	Israel	4,612,000	T	T	T	T	T	F	T	T	T	F	T	1161055000
30	MARATHI	India	64,783,000	T	T	T	T	T	F	T	T	T	F	T	1161055000
31	SERBO_CROATIAN	Yugoslavia	21,000,000	T	T	T	T	T	F	T	T	T	F	T	1161055000
32	SWAHILI	Tanzania	5,000,000	T	T	T	T	T	F	T	T	T	F	T	1161055000
33	TURKISH	Turkey	59,000,000	T	T	T	T	T	F	T	T	T	F	T	1161055000
34	ARMENIAN	Armenia	6,836,000	T	T	T	T	T	F	T	T	T	F	F	1161050000
35	KURMANJI	Turkey	7,000,000	T	T	T	T		F	Q	T	T	T	F	1161010000

SL Dry Run

- **Planned Duration: 1 week beginning March 5; Multiple Sites**
  - **U. California at Berkeley, Carnegie-Mellon U., Johns Hopkins U., U. Maryland, MITRE, NYU, U. Pennsylvania/LDC, Sheffield U, USC/ ISI**
- **Philippine language Cebuano selected. Survey had identified:**
  - **Bible, small news text archive, several printed dictionaries and grammars**
- **8 hours into project, LDC had found**
  - **250,000 words of news texts, several other small monolingual and bilingual Cebuano texts, 4 computer-readable lexicons exceeding 24,000 entries in total**
  - **Considerable overlap among what different sites discovered**
- **Disparity between survey and experiment results**
  - **greater effort during the exercise**
  - **survey search methodology**
    - » **searches for "Cebuano" + "lexicon", "dictionary", "news." missed resources labeled with alternative names (Bisayan and Visayan)**
- **Issues**
  - **Overlap of effort inevitable**
  - **No mode of electronic communication fast enough; LDC staff sat together**
  - **Cebuano related closely to other Philippine languages, more distantly to other Malayo-Polynesian languages; difficult for non-speakers to distinguish Cebuano**
    - » **Identified unique Cebuano worlds without inflectional morphology**
    - » **Cebuano speakers checked the texts**

SL Formal Evaluation

- Locate or build resources, develop & evaluate systems
- Language
  - Hindi; Results significantly different
  - Orders of magnitude more text on web; problem shifted to processing
  - Within few hours basic resources located
  - “large resource conspiracy” developed
- Encoding
  - Hindi written in Devanagari
  - Character Encodings Standards such as UNICODE & ISCII not commonly used.
  - Every website had proprietary encodings; several sites had more than one
- Results
  - All texts converted to Unicode (UTF-8) even though underspecified
  - Team created finer encoding specification
  - Texts also delivered in original form and ITRANS romanization
  - Although character conversion took several weeks, integration of LRs and system development were accomplished in 1 month
  - Hindi systems compared favorably in Topic Detection and Tracking, Cross Language IR, Content Extraction, Summarization and MT
- Recommendation from sites
  - The surprise language experiment was tremendous success!
  - Let’s NOT do it again.

Current & Forthcoming

- **LDC has NSF funds to extend resource finding, building efforts to 6 languages working in collaboration with University of Maryland at Baltimore and Johns Hopkins University**
  - **languages with >1,000,000 native speakers**
  - **high probability of basic resources available electronically**
  - **wide variety of morpho-syntactic features**
  - **wide variety of geographical regions**
  - **at least two closely related language to support transfer experiments**
  - **not likely to include European languages, Arabic, Chinese**
  - **likely to include Dravidian, Indo-Aryan, Ingush, Malayo-Polynesian, Semitic, Turkic languages**
  - **All data will be published**
  - **metadata will be catalogued in OLAC as well as LDC Catalog**

- **TIDES community**
  - **will fund continuation of the survey**
  - **wants to extend the set of resources available for the 6 languages**
  - **Specifically wants annotations to support information detection extraction, summarization and translations**

Proposal

- LDC obligated to current path for at least the next year.

- SuperConsortium (e.g. of ICWLR, COCOSDA, ELSNET, ENABLER Network, LDC, ELRA, Korterm/Kaist, GSK, LDCIL & Talkbank and other partners) promote a minimum specification of core languages, core LRs, survey questions; define extended set of languages and resources on longer term
- LDC makes LR survey available to sites who submit complete survey answers for one new language
- SuperConsortium promotes the plan to EC, NSF, national funding agencies & commercial sponsors
- In many cases resources already exist but need to be identified and published. Resources collected & created are distributed through LDC, ELDA.
- Metadata for resources is published in OLAC and IMDI compliant forms and union catalogs
- Corpus specifications and annotation tools, including AGTK and tools created by Talkbank, are shared with other researchers, research groups to extend the LR catalog to new languages and for new data types.