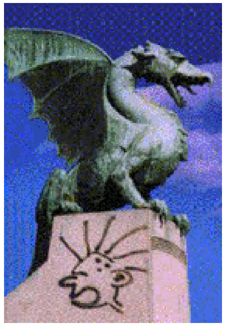
Rapid Deployment of Speech Processing Systems to New Languages and Domains

Tanja Schultz

University of Karlsruhe & Carnegie Mellon University
Cognitive Systems Laboratory

http://csl.ira.uka.de



IS-LTC 2008, *Thursday, October 16th 2008,* Ljubljana

Outline

- The World's Languages
  - 6900 languages – So what?
  - Language Extinction – What can the community do about it?
  - Do we need Speech Processing for all of them?
  - Is this really science – not just retraining on a new language?
- Language Characteristics
  - Written form, scripts, letter-to-sound relationship
  - Issues and Differences between languages
- Challenges for Multilingual Speech Processing
  - Lack of Resources (Money, Data, Technical Support)
  - Lack of Experts
- Solutions
  - SPICE: A Rapid Language Adaptation Server
  - Technologies: Leveraging off GlobalPhone & FestVox
  - Experiments and Results
- Conclusions and Future Work

Outline

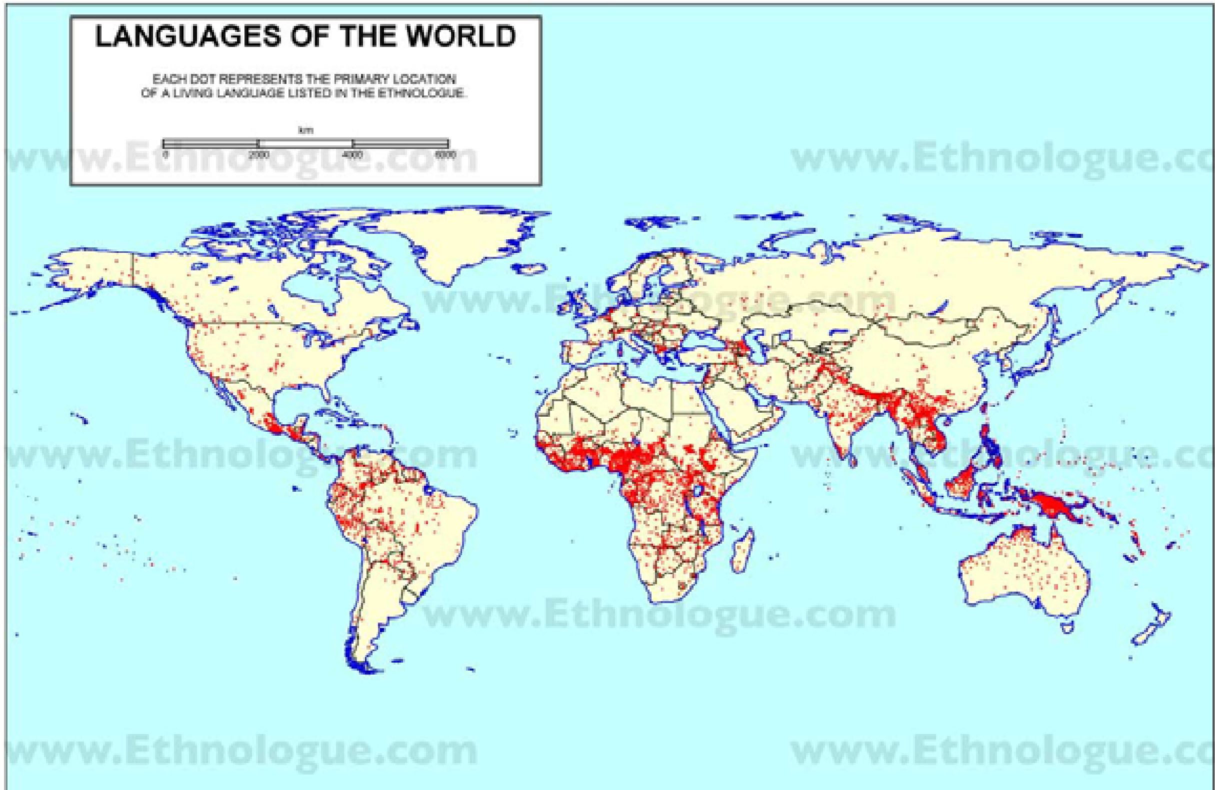
- The World's Languages
  - 6900 languages – So what?
  - Language Extinction – What can the community do about it?
  - Do we need Speech Processing for all of them?
  - Is this really science – not just retraining on a new language?
- Language Characteristics
  - Written form, scripts, letter-to-sound relationship
  - Issues and Differences between languages
- Challenges for Multilingual Speech Processing
  - Lack of Resources (Money, Data, Technical Support)
  - Lack of Experts
- Solutions
  - SPICE: A Rapid Language Adaptation Server
  - Technologies: Leveraging off GlobalPhone & FestVox
  - Experiments and Results
- Conclusions and Future Work

Everyone speaks English, why bother?

- Total number of Languages in the world: **6912**
- Language is not only a communication tool but fundamental to cultural identity and empowerment!
- Cultures, ideas, memories transmit ***through language***

- Intellectual issues (e.g. world history) Practical issues (medical practices) Literature, …
  Slovakian proverb: "with each newly learned language you acquire a new soul"

- Preserve linguistic diversity! Similar to eco systems (David Crystal)

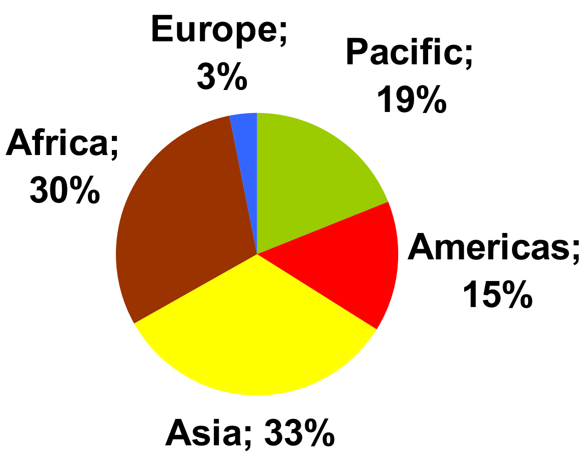


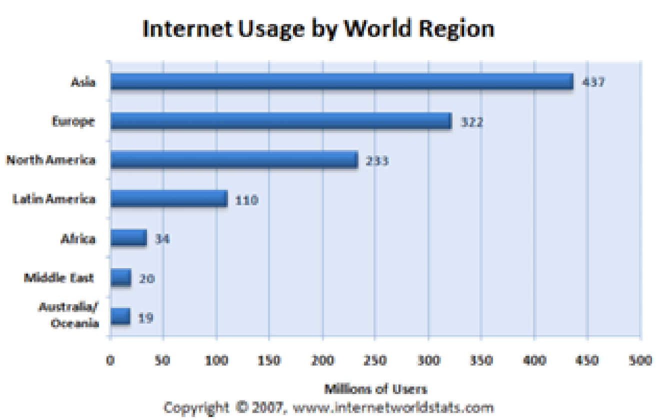
Each dot gives the geographic center of the 6,912 living languages, http://www.ethnologue.com (accessed Jul 2007)

Increasing Language Diversity in Web

Diversity of Languages in the Internet grows rapidly

- Top-10: 200%, All others: 440%
- Portuguese: 524%
- Arabic: 940%





Top Ten Languages Used in the Web
(Number of Internet Users by Language)

TOP TEN LANGUAGES IN THE INTERNET	% of all Internet Users	Internet Users by Language	Internet Penetration by Language	Internet Growth for Language (2000 - 2007)	2007 Estimate World Population for the Language
English	31.7 %	365,893,996	17.9 %	157.7 %	2,042,963,129
Chinese	31.7 %	166,001,513	12.3 %	413.9 %	1,351,737,925
Spanish	8.8 %	101,539,204	22.9 %	311.4 %	442,525,601
Japanese	7.5 %	86,300,000	67.1 %	83.3 %	128,646,345
German	5.1 %	58,981,592	61.1 %	112.9 %	96,488,326
French	5.1 %	58,456,702	15.1 %	379.2 %	387,820,873
Portuguese	4.1 %	47,326,760	20.2 %	524.7 %	234,099,347
Korean	3.0 %	34,120,000	45.6 %	79.2 %	74,811,368
Italian	2.7 %	31,481,928	52.9 %	138.5 %	59,546,696
Arabic	2.5 %	28,782,300	8.5 %	940.5 %	340,548,157
TOP TEN LANGUAGES	**84.8 %**	978,883,995	19.0 %	198.0 %	5,159,187,766
Rest of World Languages	**15.2 %**	175,474,783	12.4 %	440.3 %	1,415,478,651
WORLD TOTAL	**100.0 %**	1,154,358,778	17.6 %	219.8 %	6,574,666,417

(*) NOTES: (1) Internet Top Ten Languages Usage Stats were updated for June 30, 2007. (2) Internet Penetration is the ratio between the sum of Internet users speaking a language and the total population estimate that speaks that specific language. (3) The most recent Internet usage information comes from data published by Nielsen//NetRatings, International Telecommunications Union, Computer Industry Almanac, and other reliable sources. (4) World population information comes from the world gazetteer web site. (5) For definitions and navigation help, see the Site Surfing Guide. (6) Stats may be cited, stating the source and establishing an active link back to Internet World Stats.

Currently 6900 Languages, but …

- Extinction of languages on massive scale
  (David Crystal, Spotlight 3/2000)
- Half of all existing languages die out over next century
  ⇒ On Average: Every two weeks one language dies!
- Survey Feb 1999 from Summer Institute of Linguistics

51 languages with 1 speaker left

28 of those in Australia alone

500 languages with < 500 spks

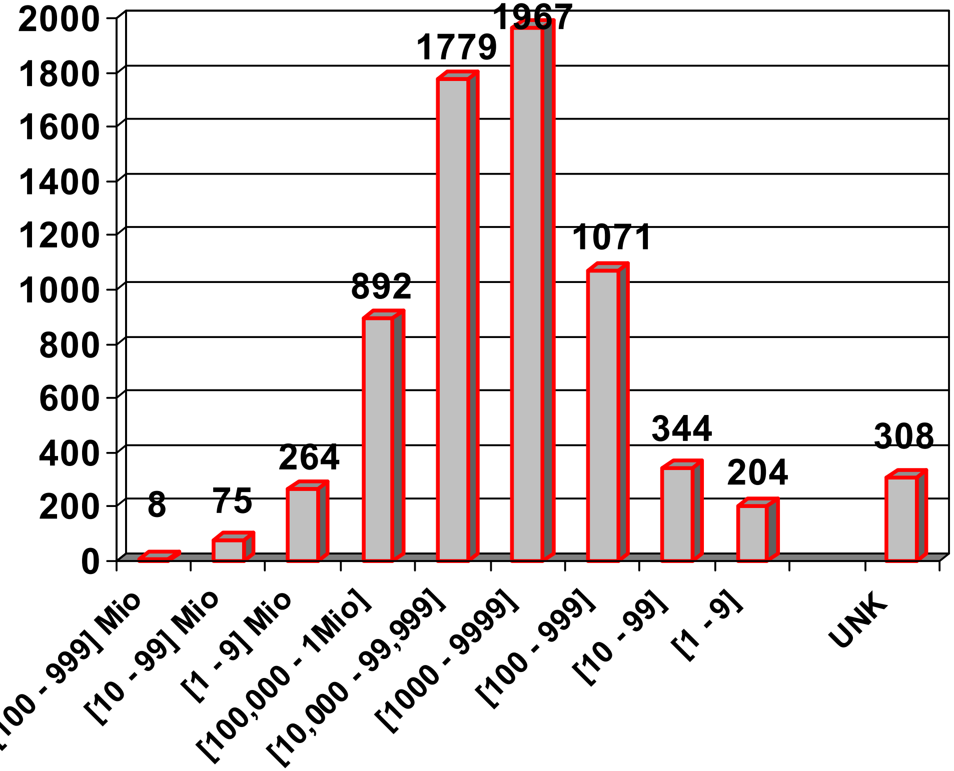
1500 languages with < 1000 spks

3000 languages with < 10.000

5000 languages with < 100.000

(not safe even for >100.000)

96% of world's languages are spoken by only 4% of its people



How to safe Languages?

Prerequisites and Costs:

- Community itself must want it, Surrounding culture must respect it
- Funding for courses, materials, and teachers, support the community
- Crystal estimates about $80.000 / year per language
- 3000 endangered languages is about $700Mio …
- Foundation of endangered languages (FEL), UNESCO project

How could our community contribute:

- Field Work and Community Outreach
  - Get the tools to the people, i.e. flexible, portable, easy to handle
  - Engage and actively involve native speakers
- Lower the overall costs for data acquisition
  - Automate the solicitation and data collection process
  - Identify language specific aspects and focus
- Reduce the data needs without sacrificing performance
  - Streamline techniques & approaches to perform on small data
  - Reuse language independent aspects of data and models

Why Speech Processing?

Language support is good but why *Speech*?

- Computerization: Speech is *the* key technology
  - ➔ Ubiquitous Information Access: on the go, phone-based
  - ➔ Mobile Devices: Too small and cumbersome for keyboards

- Globalization:
  - ➔ Cross-cultural Human-Human Interaction
    - ➔ Multilingual Communities: EU, South Africa, …
    - ➔ Humanitarian needs, disaster, health care
    - ➔ Military ops, communicate with local people
  - ➔ Human-Machine Interfaces
    - ➔ People expect speech-driven applications in their mother tongue

⇒ **<u>Speech Processing in multiple Languages</u>**

ML Speech Processing – A Research Issue?

Just retraining on foreign data? - No science!

- New language – new challenges
  - Writing system: different or no script, no vowelization, G-2-P
  - Word segmentation, morphology
  - Sound system: tonals, clicks
- Different Cultures – social factors
  - Trust, access, exposure, background
- Lack of Data and Resources
  - Audio, Transcripts, Pronunciations, Text, parallel bilingual data
- Lack of Experts
  - Technology experts without language expertise
  - Native language experts without technology expertise

If we can solve the research issues for some languages, we might help the others along the way!

Outline

- The World's Languages
  - 6900 languages – So what?
  - Language Extinction – What can the community do about it?
  - Do we need Speech Processing for all of them?
  - Is this really science – not just retraining on a new language?
- Language Characteristics
  - Written form, scripts, letter-to-sound relationship
  - Issues and Differences between languages
- Challenges for Multilingual Speech Processing
  - Lack of Resources (Money, Data, Technical Support)
  - Lack of Experts
- Solutions
  - SPICE: A Rapid Language Adaptation Server
  - Technologies: Leveraging off GlobalPhone & FestVox
  - Experiments and Results
- Conclusions and Future Work

Language Characteristics

- ➔ <u>Prosody, Tonality:</u> Stress, Pitch, Lenght pattern, Tonal contours
  (e.g. Mandarin 4, Cantonese 8, Thai & Vietnamese 5)
- ➔ <u>Sound system:</u> simple vs very complex sound systems
  (e.g. Hawaiian 5V+8C vs. German 17V+3D+22C)
- ➔ <u>Phonotactics:</u> simple syllable structure vs complex consonant clusters
  (e.g. Japanese Mora-syllables vs. German pf,st,ks)
- ➔ <u>Segmentation:</u> Written form separate words by white space?
  (NO: Chinese, Japanese, Thai, Vietnamese)
- ➔ <u>Morphology:</u> short units, compounds, agglutination

English: Natural segmentation into short units – great!

German: Compounds – not quite so good

Donau-dampf-schiffahrt**s**-gesellschaft**s**-kapitän**s**-mütze …

Turkish: Agglutination – looooong phrases

Osman-lı-laç-tır-ama-yabil-ecek-ler-imiz-den-mi**ş**-siniz

behaving as if you were of those whom we might consider not converting into Ottoman

Writing Systems

Writing systems – basic unit is a Grapheme:

Logographic: based on semantic units, grapheme represents meaning

Chinese: >10.000 *hanzi*; Japanese ~7000 *kanji,* Korean to some extend

Phonographic: based on sound units, grapheme represents sound

Segmental: grapheme roughly corresponds to phonemes

Latin (190), Cyrillic (65), Arabic (22) graphems

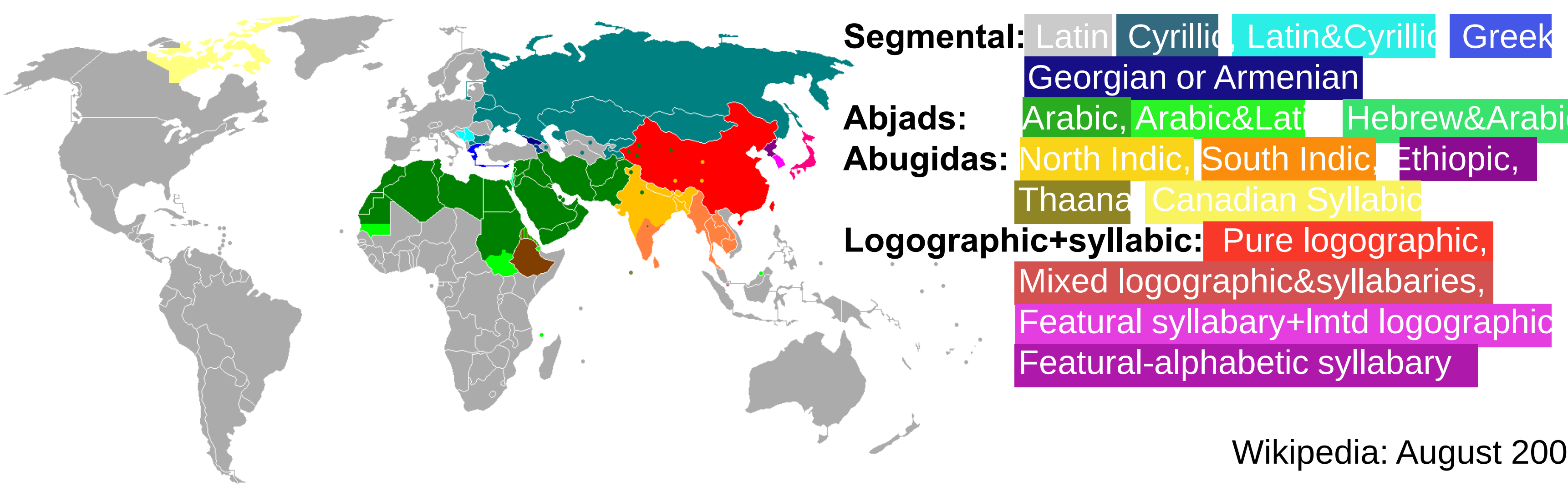
Abjads = consonantal segmental phonographic, e.g. Arabic

Syllabic: grapheme represents entire syllable, e.g. Japanese *kana*

Abugidas = mix of segmental and syllabic systems

Featural: elements smaller than phone, e.g. articulatory features

e.g. Korean: ~5600 *gulja*



Wikipedia: August 2007

Scripts – Some examples

Scripts of some languages: Arabic, Bulgarian, Catalan, Chinese, Croatian, Czech, English, Greek, Hebrew, Hindi, Italian, Japanese, Korean, Romanian, Serbian, Thai

<u>How many languages do have a written form?</u>

- Omniglot lists about 780 languages that have scripts
- True number might be closer to 1000

(Source Simon Ager, 2007, www.omniglot.com)

➔Logographic scripts, mostly 2 representatives:

- Chinese: ~ 10.000 hanzi,
- Japanese: ~7000 kanji (+ 3 other scripts ☺)

➔ Phonographic:

- Korean: ~5600 gulja,
- Arabic, Devanagari, Cyrillic, Roman: ~100 characters

Grapheme-to-Phoneme Relation

Grapheme-to-Phoneme (Letter-to-Sound) Relationship:

Logographic: NO relationship at all

concern for Chinese, Japanese, Korean

Phonographic: segmental: close – far – complicated

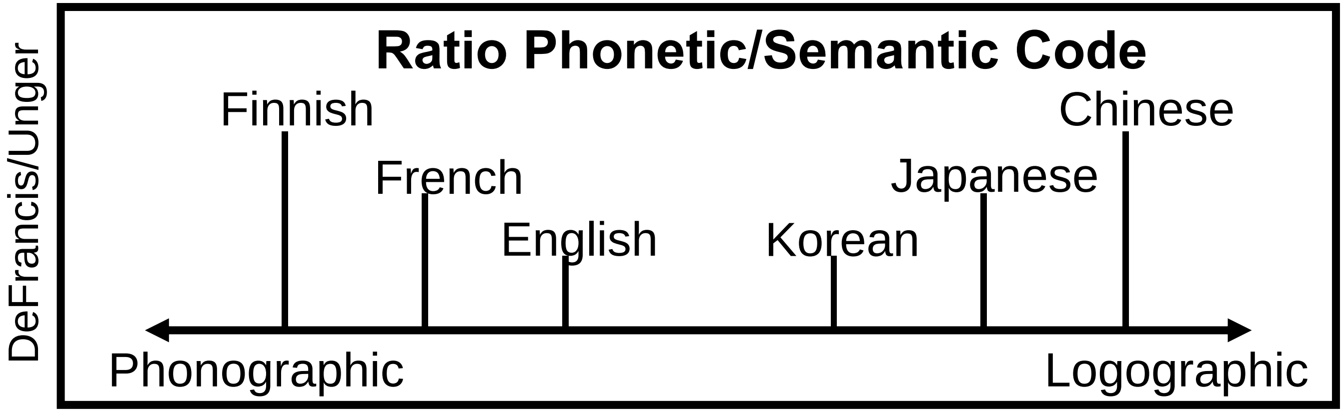
e.g. Finnish, Spanish: more or less 1:1, -- English: try „Phydough“

Phonographic: segmental – consonantal

e.g. Arabic: no short vowels written

Phonographic: syllabic

e.g. Thai, Devanagari: C-V flips



➔ Automatic Generation of Pronunciations might get complicated

Outline

- The World's Languages
  - 6900 languages – So what?
  - Language Extinction – What can the community do about it?
  - Do we need Speech Processing for all of them?
  - Is this really science – not just retraining on a new language?
- Language Characteristics
  - Written form, scripts, letter-to-sound relationship
  - Issues and Differences between languages
- Challenges for Multilingual Speech Processing
  - Lack of Resources (Money, Data, Technical Support)
  - Lack of Experts
- Solutions
  - SPICE: A Rapid Language Adaptation Server
  - Technologies: Leveraging off GlobalPhone & FestVox
  - Experiments and Results
- Conclusions and Future Work

Challenges of MLSP

- Lack of Resources: Stochastic approach needs **many** data
  - Hundreds of hours audio recordings and corresponding transcriptions
    Audio data ≤ 40 languages; Transcriptions take up to 40x real time
  - Pronunciation dictionaries for large vocabularies (>100.000 words)
    Large vocabulary pronunciation dictionaries ≤ 20 languages
  - Mono- and bilingual text corpora: few language pairs, pivot mostly English
- Algorithms are language independent – MLSP is not!
  - Other Languages bring unseen challenges (segmentation, G2P, etc.)
  - Have we already seen ALL or MOST of the language characteristics?
- Social and Cultural Aspects
  - Non-native speech and language, code switching
  - Combinatorical explosion (domain, speaking style, accent, dialect, ...)
  - Few native speakers at hand for minority (endangered) languages
- Lack of Language Experts
  - Bridge the gap between technology experts and language experts

One Solution: Learning Systems

⇒ Intelligent systems that learn a language from the user

- Efficient learning algorithms for speech processing
  - Learning:
    - Interactive learning with user in the loop
    - Statistical modeling approaches
  - Efficiency:
    - Reduce amount of data (save time and costs): at least by factor of 10
    - Speed up development cycles: days rather than months

  ⇒ Rapid Language Adaptation from universal models

- Bridge the gap between language and technology experts
    - Technology experts do not speak all languages in question
    - Native users are not in control of the technology

SPICE

Speech Processing:

Interactive Creation and Evaluation toolkit

- National Science Foundation, Grant 10/2004, 4 years
- Principle Investigator Tanja Schultz
- Bridge the gap between technology experts → language experts
  - Automatic Speech Recognition (ASR),
  - Machine Translation (MT),
  - Text-to-Speech (TTS)

- Develop web-based intelligent systems
  - Interactive Learning with user in the loop
  - Rapid Adaptation of universal models to unseen languages
- SPICE webpage *http://cmuspice.org*

SPICE - Goals

Three main goals:

- Lower the overall costs for data acquisition
  - Automate the solicitation and data collection process
  - Identify language specific aspects and focus
- Reduce the data needs without sacrificing performance
  - Streamline techniques to perform on small data
  - Reuse language independent aspects of data/models
- Field Work and Community Outreach
  - Get the tools to the people, i.e. flexible, portable, easy to handle
  - Engage and actively involve native speakers

CMU SPICE

Welcome to SPICE

Getting started

SPICE is a web-based system for building an end-to-end speech system (including Automatic Speech Recognition and Text-To-Speech) in your own language.

Existing Users

Login with your account:

Login tanja

Password ************

[Login]

New Users

Create a new account:

Login

Password

Re-type Password

Email

[Create new account]

CMU SPICE

Build Your System

- Text and prompt selection (help)
- Audio collection (help)
- Phoneme selection (help)
- Grapheme-to-phoneme rules (help)

build language model first

- Lexicon pronunciation creation (help)

build language model first

- Build acoustic model (help)
- Build language model (help)
- Create ASR system
- Create speech synthesis voice

User: **TanjaSchultz** Language: **Klingon** Project: **Interspeech2007** [Logout]

SPICE Project

You must do the following to build support for your language:

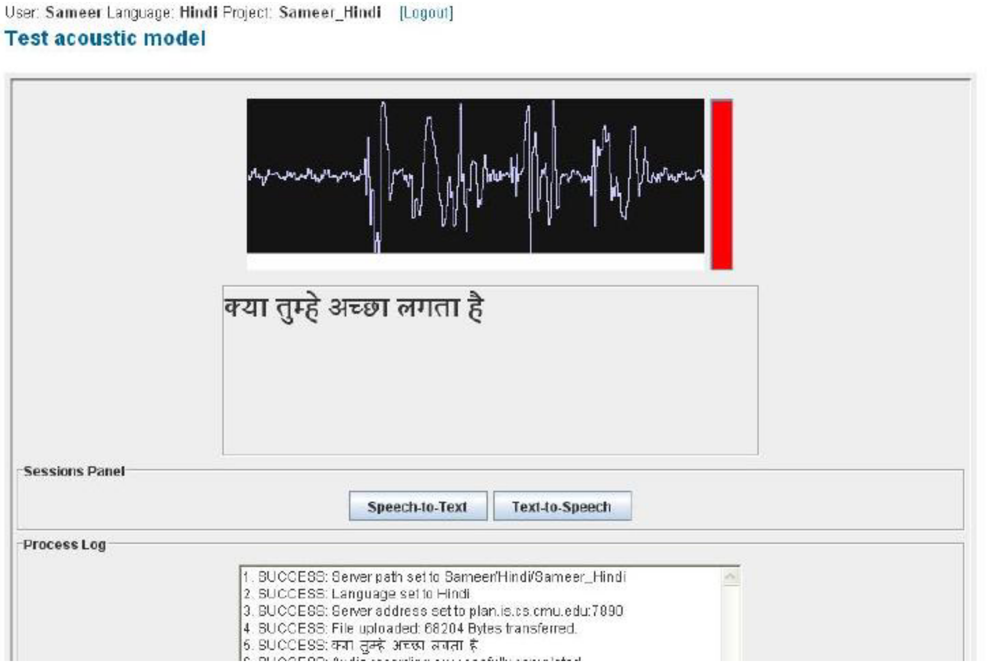
- Text collection and selection
- Audio collection
- Phoneme set specification
- Lexicon pronunciation creation
- Speech recognition acoustic model creation
- Speech recognition language model creation
- Speech synthesis voice creation

About Spice | Contact Us |

SPICE – System Functionalities

- SPICE Collects:
  - Appropriate text data
  - Appropriate audio data
- SPICE Defines and Refines:
  - Phoneme set
  - Rich prompt set
  - Lexical pronunciations
- SPICE Produces:
  - Vocabulary / Word lists (ASR, TTS, SMT)
  - Pronunciation model (ASR, TTS)
  - Acoustic model (ASR, TTS)
  - Language model (ASR, SMT)
  - Synthetic voices (TTS)
- SPICE Maintains:
  - Projects and users login
  - Data, Models, and Speech Processing Systems

Building Process



SPICE building process

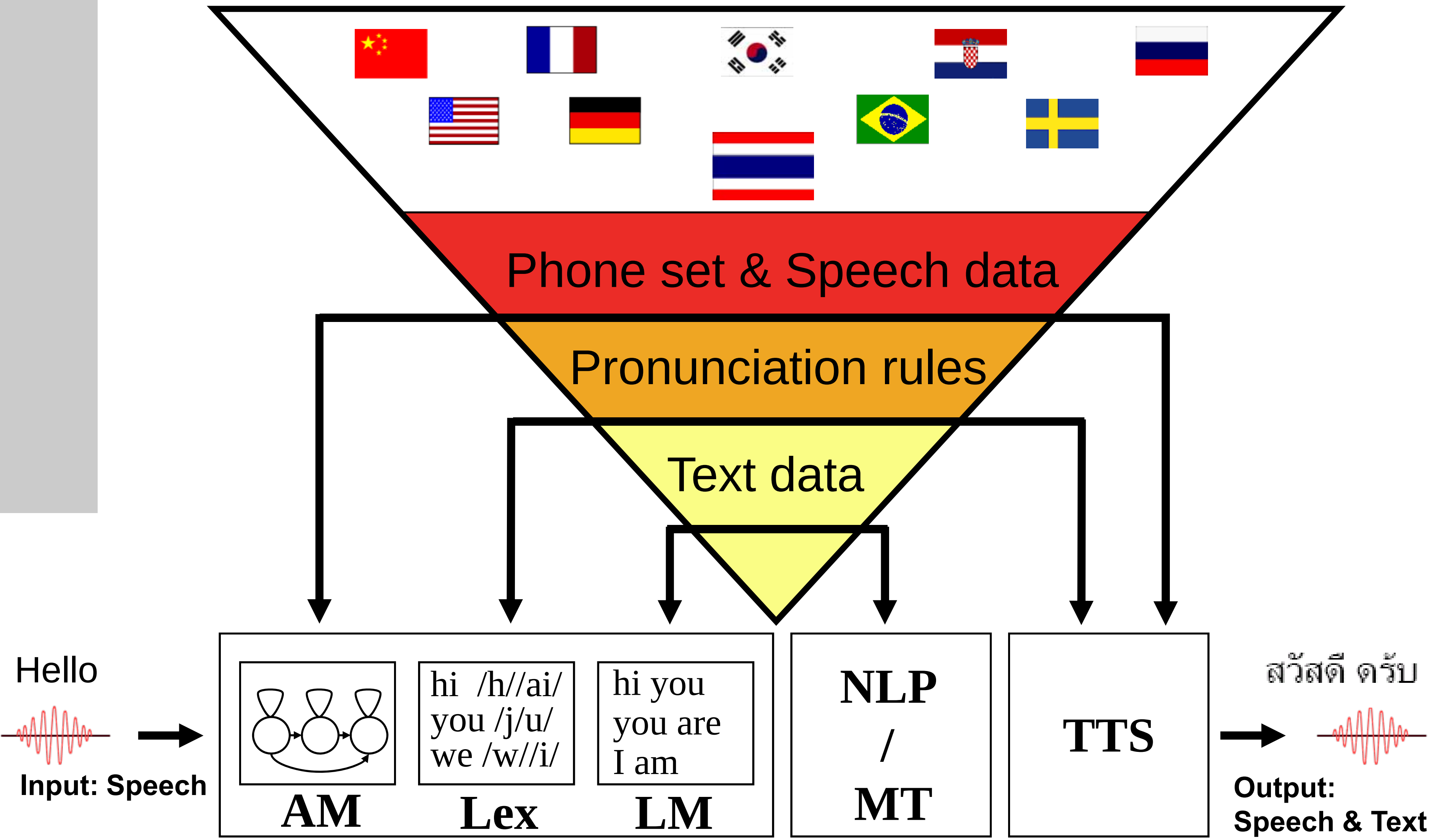
1. Collect a text corpus
2. Generate a 200-1000 ut
3. Record the prompt list from one or more native speakers
4. Define a phoneme set
5. Construct a lexicon and letter-to-sound rules
6. Build a language model from the text corpus
7. Build acoustic models for ASR
8. Build voice models for TTS
9. Evaluate ASR and TTS using “talk-back” function

SPICE: Demo Tape

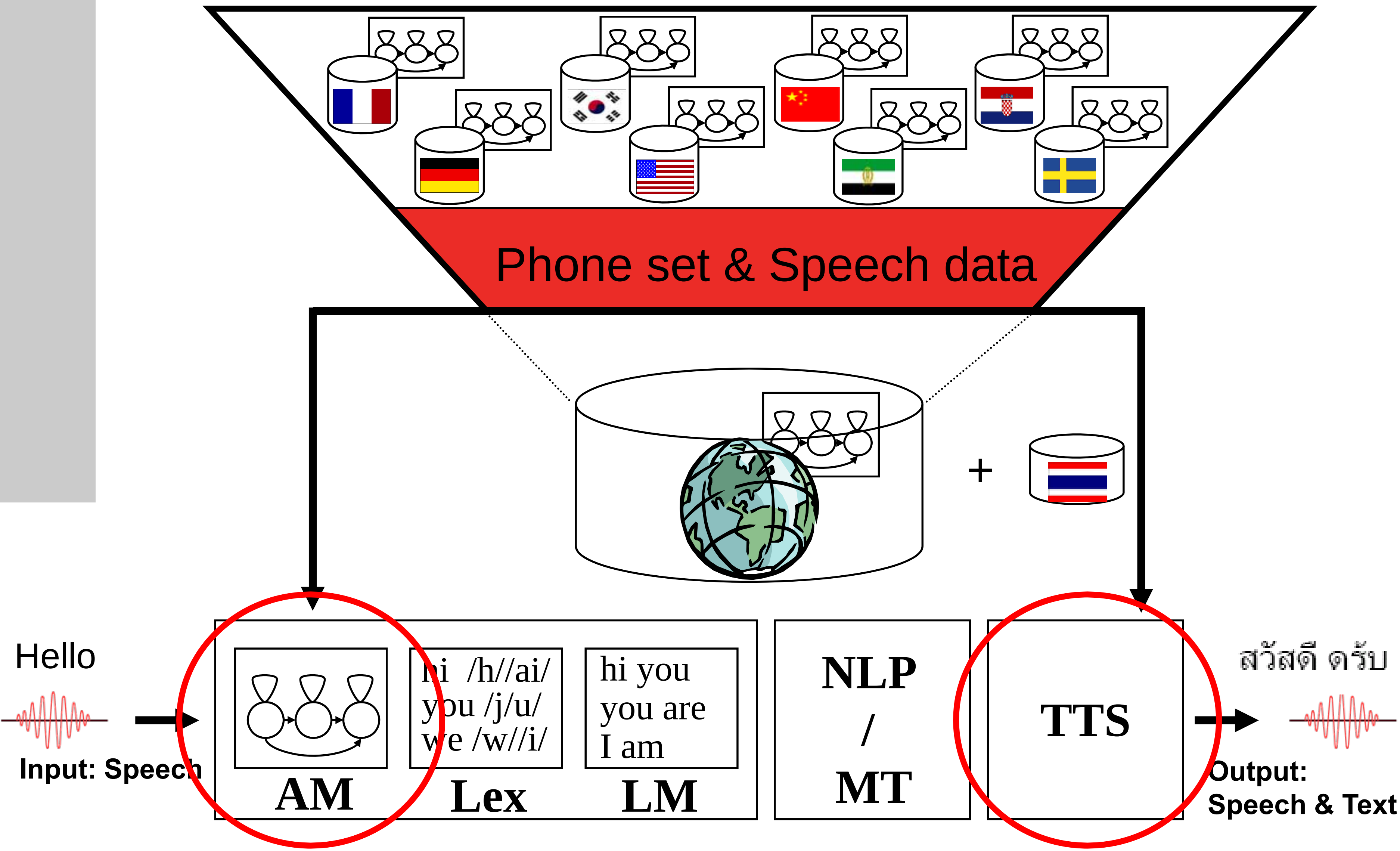
Outline

- The World's Languages
  - 6900 languages – So what?
  - Language Extinction – What can the community do about it?
  - Do we need Speech Processing for all of them?
  - Is this really science – not just retraining on a new language?
- Language Characteristics
  - Written form, scripts, letter-to-sound relationship
  - Issues and Differences between languages
- Challenges for Multilingual Speech Processing
  - Lack of Resources (Money, Data, Technical Support)
  - Lack of Experts
- Technical Solutions
  - SPICE: A Rapid Language Adaptation Server
  - Technologies: Leveraging off GlobalPhone & FestVox
  - Experiments and Results
- Conclusions and Future Work

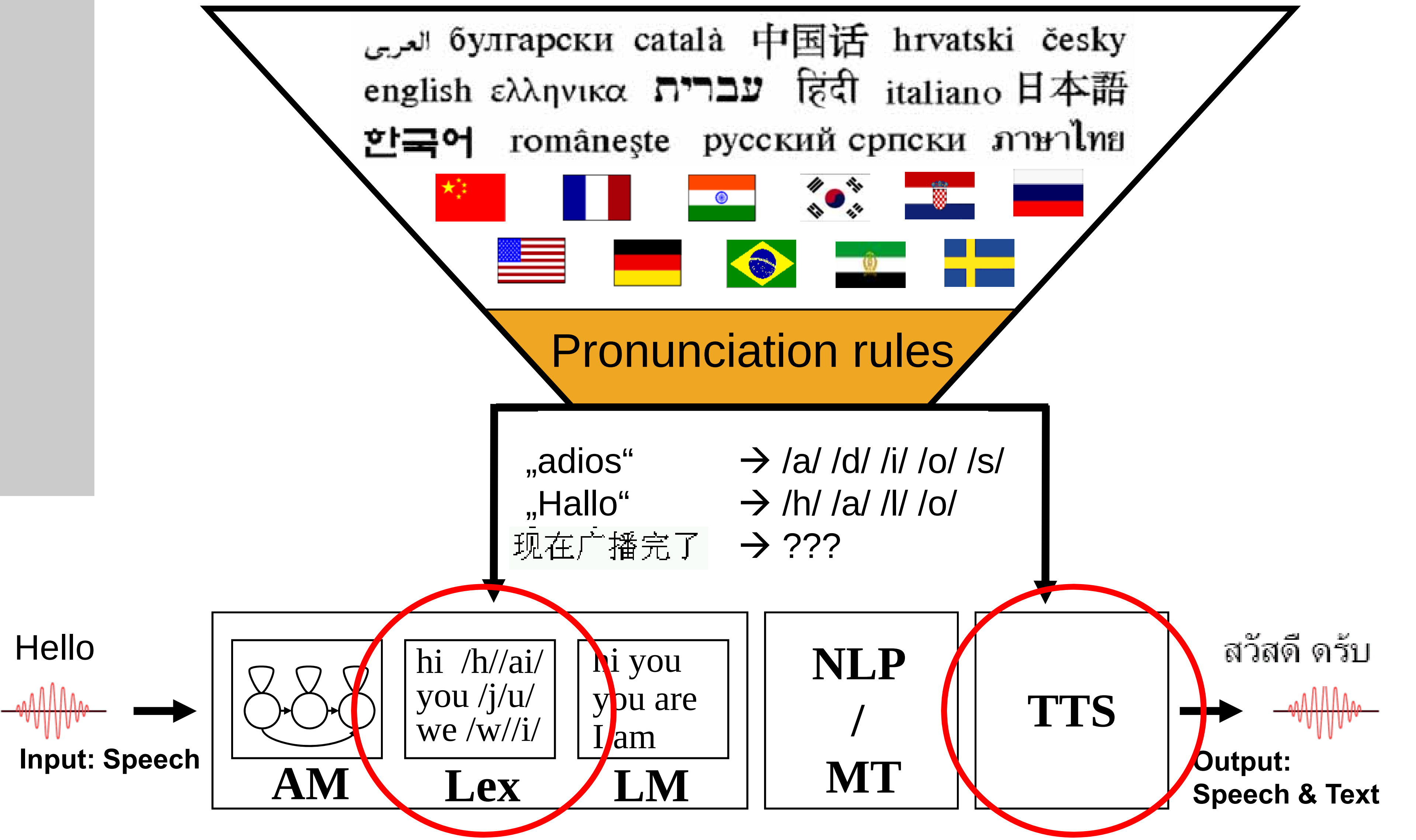
Speech Processing Systems



Rapid Portability: Acoustic Models

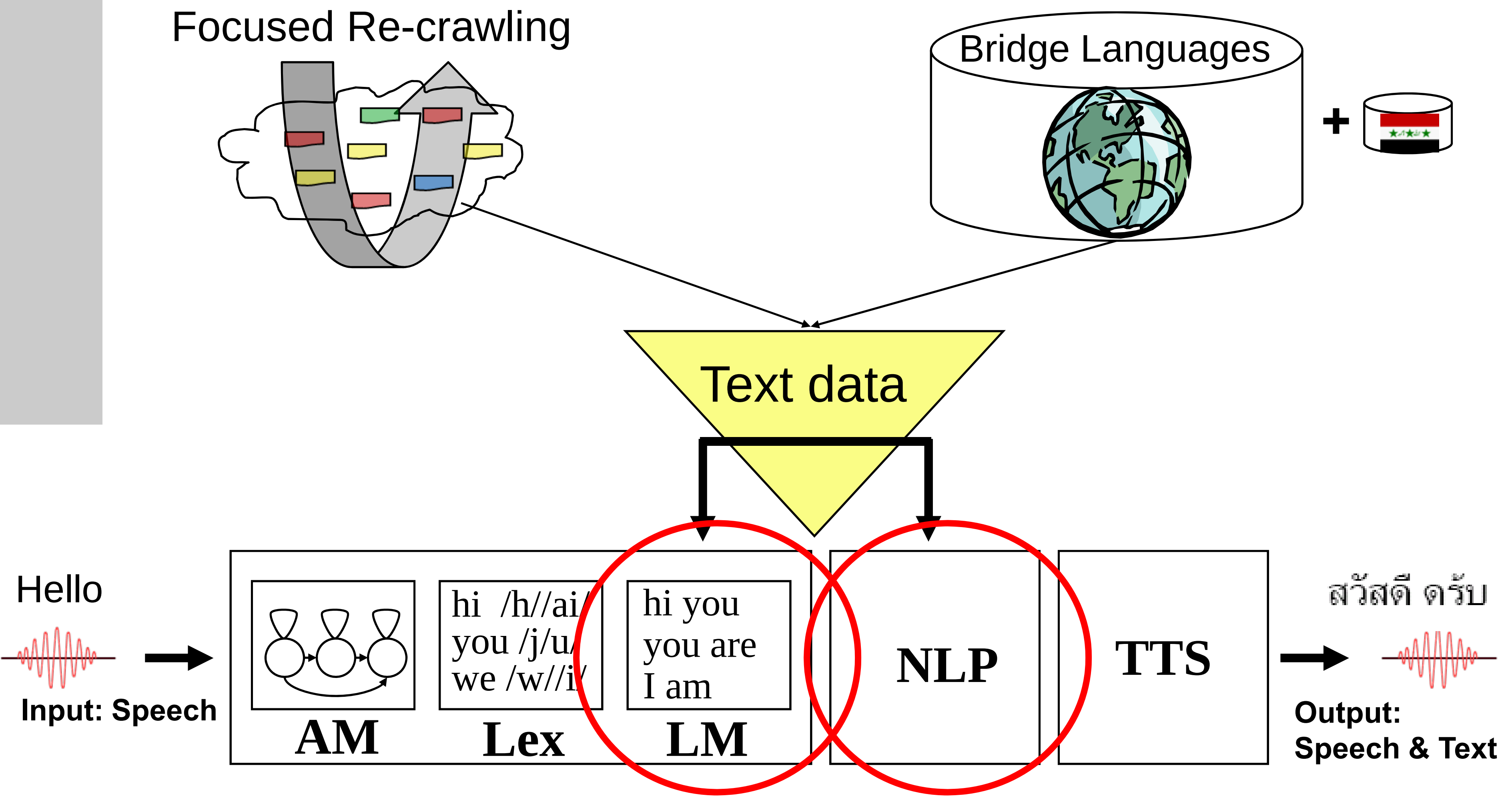


Rapid Portability: Pronunciation Dictionary

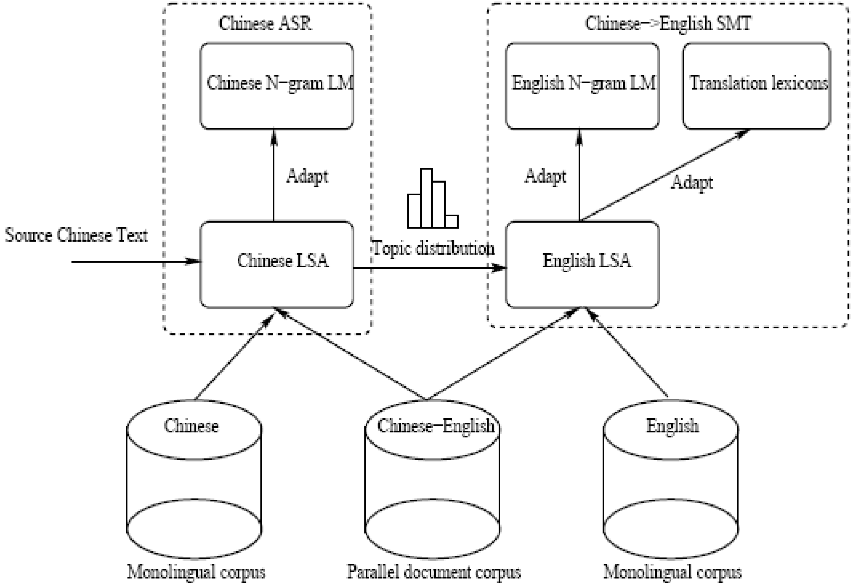


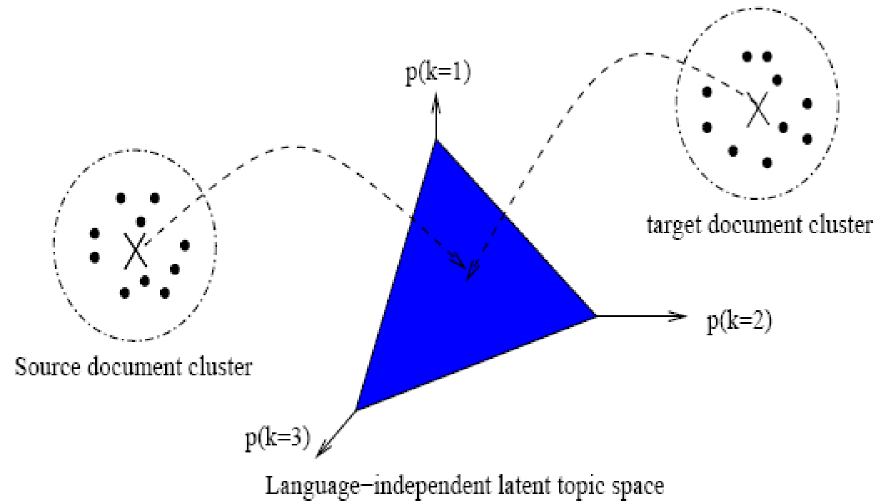
Rapid Portability: Language Modeling

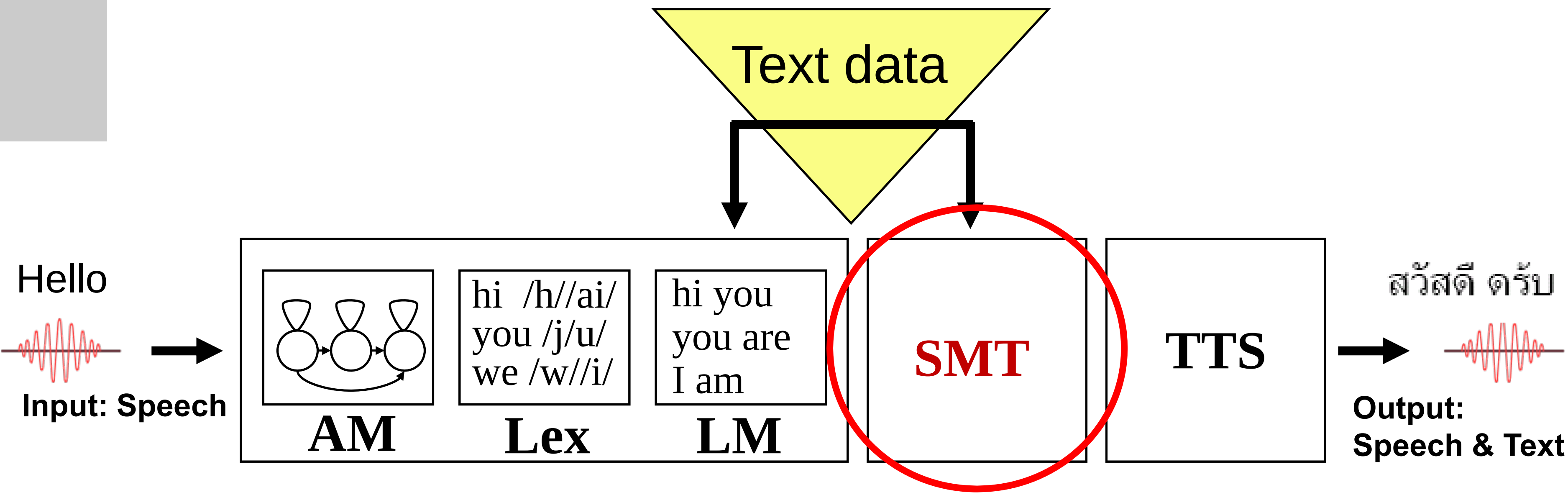
Resource rich languages ↔ Resource low languages



Bilingual LSA for Speech Translation

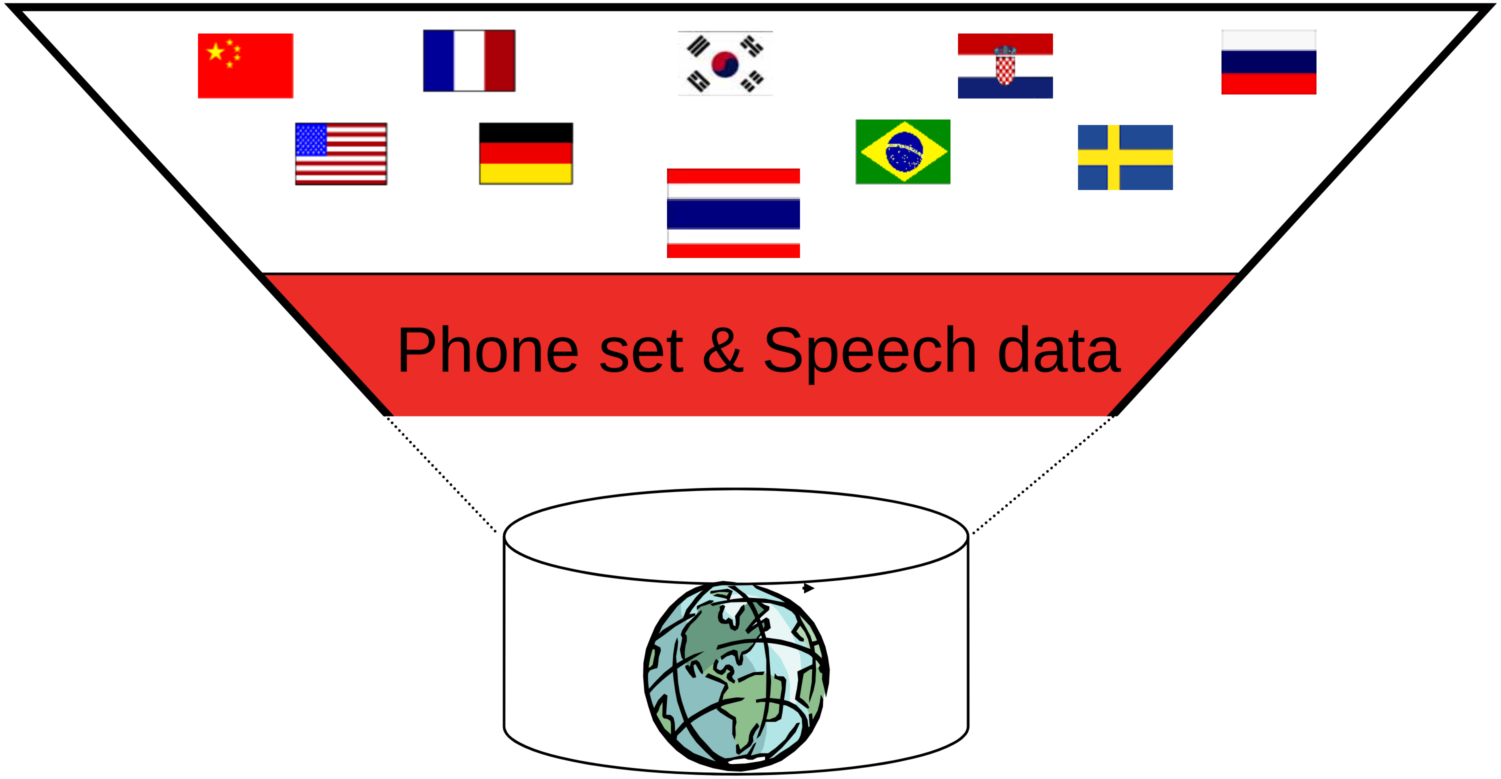






Yik-Cheung Tam, Tanja Schultz, Bilingual-LSA Based Translation Lexicon Adaptation for Spoken Language Translation, IS2007

Multilingual Text and Speech Database



First step for studies on Multilingual Speech Processing
and language dependencies:
Collect large amounts of data in many languages
Project GlobalPhone (since 1995)

GlobalPhone

Arabic	Croatian	Turkish
Ch-Mandarin	Czech	+ Thai
Ch-Shanghai	Portuguese	+ Creole
German	Russian	+ Polish
French	Spanish	+ Bulgarian
Japanese	Swedish	+ ... ???
Korean	Tamil	

Multilingual Database

- Widespread languages
- Native Speakers
- Uniform Data
- Broad Domain
- Large Text Resources
  - ➔ Internet, Newspaper

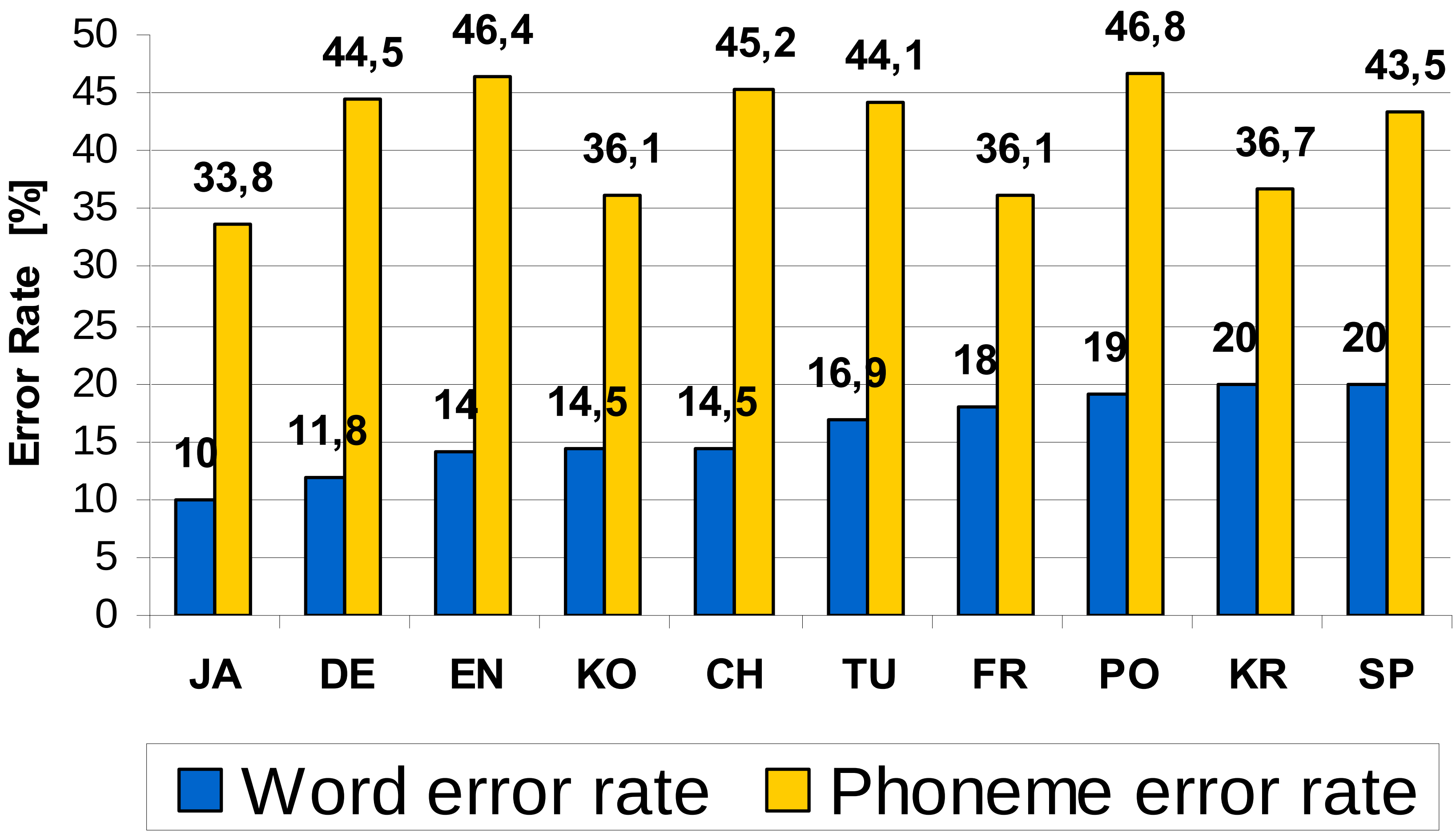
Corpus

- 19 Languages … counting
- ≥ 1800 native speakers
- ≥ 400 hrs Audio data
- Read Speech
- Filled pauses annotated

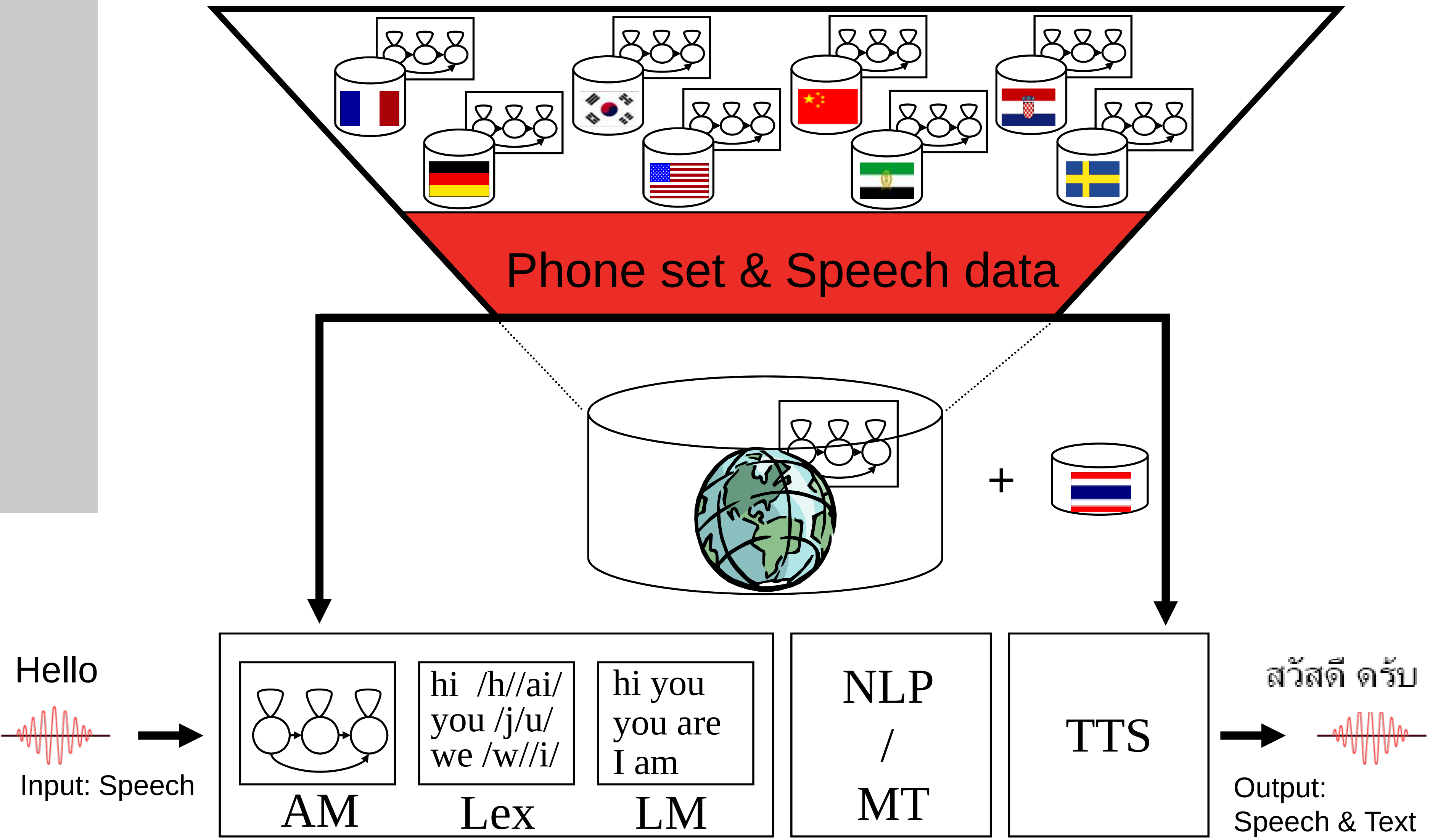
Available from ELRA

http://www.cs.cmu.edu/~tanja/GlobalPhone

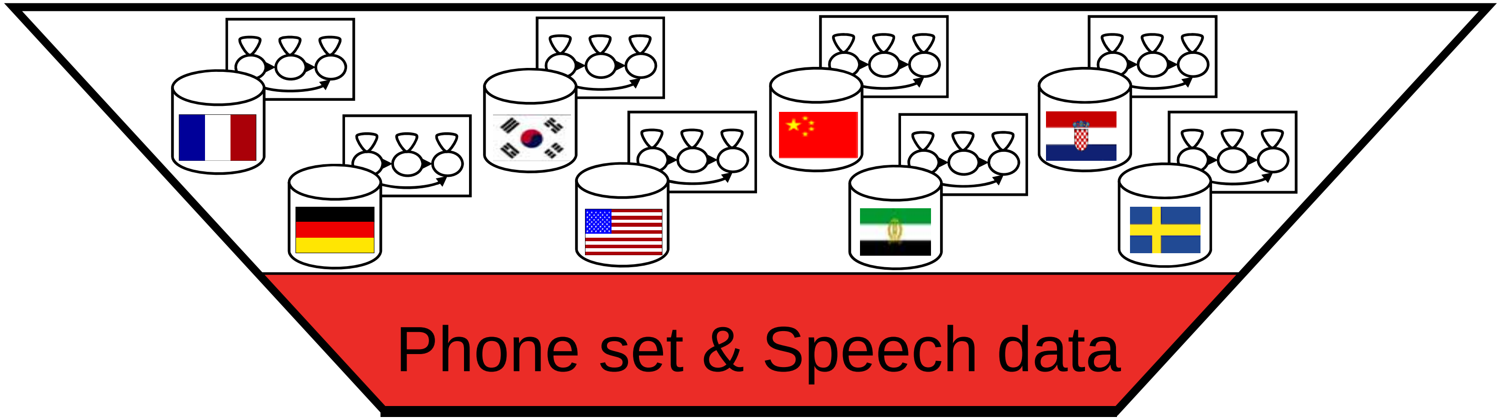
GlobalPhone Recognizers in 10 Languages



Rapid Portability: Acoustic Models



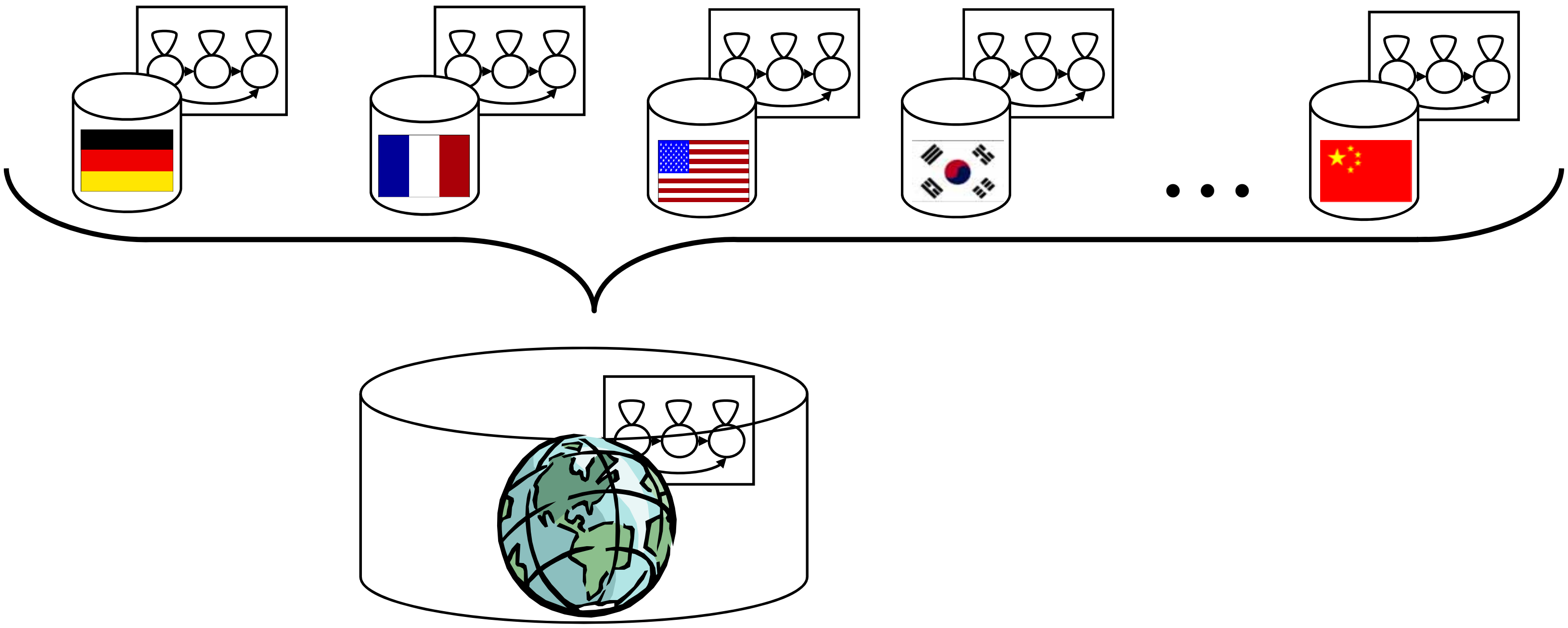
Rapid Portability: Data



Step 1:

- Uniform multilingual database (GlobalPhone)
- Build Monolingual acoustic models in many languages

Multilingual Acoustic Modeling



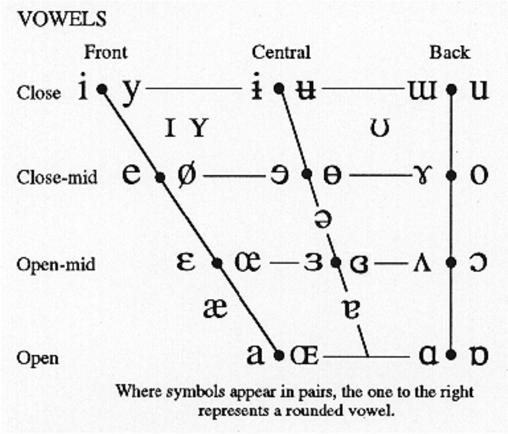
<u>Step 2:</u>

- Combine monolingual acoustic models to a set of multilingual “language independent” acoustic model

Universal Sound Inventory

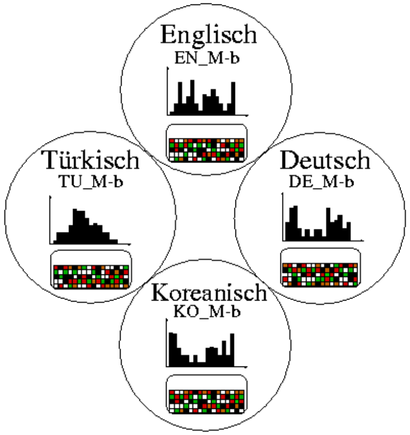
Speech Production is independent from Language $\Rightarrow$ IPA

1) IPA-based Universal Sound Inventory

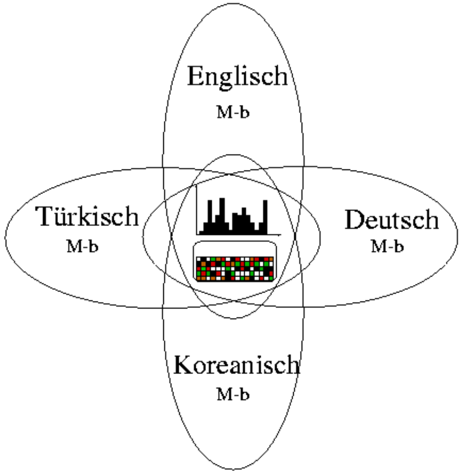


2) Each sound class is trained by data sharing

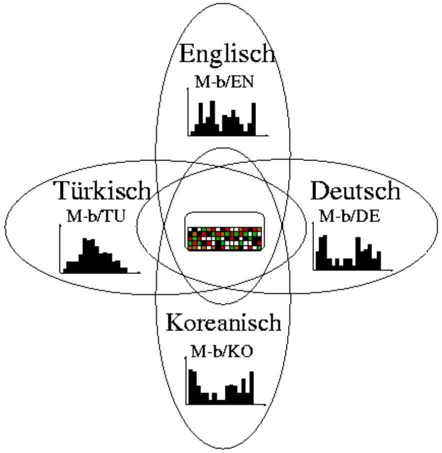
- Reduction from 485 to 162 sound classes
- *m,n,s,l* appear in all 12 languages
- *p,b,t,d,k,g,f* and *i,u,e,a,o* in almost all



ML-Sep



ML-Mix

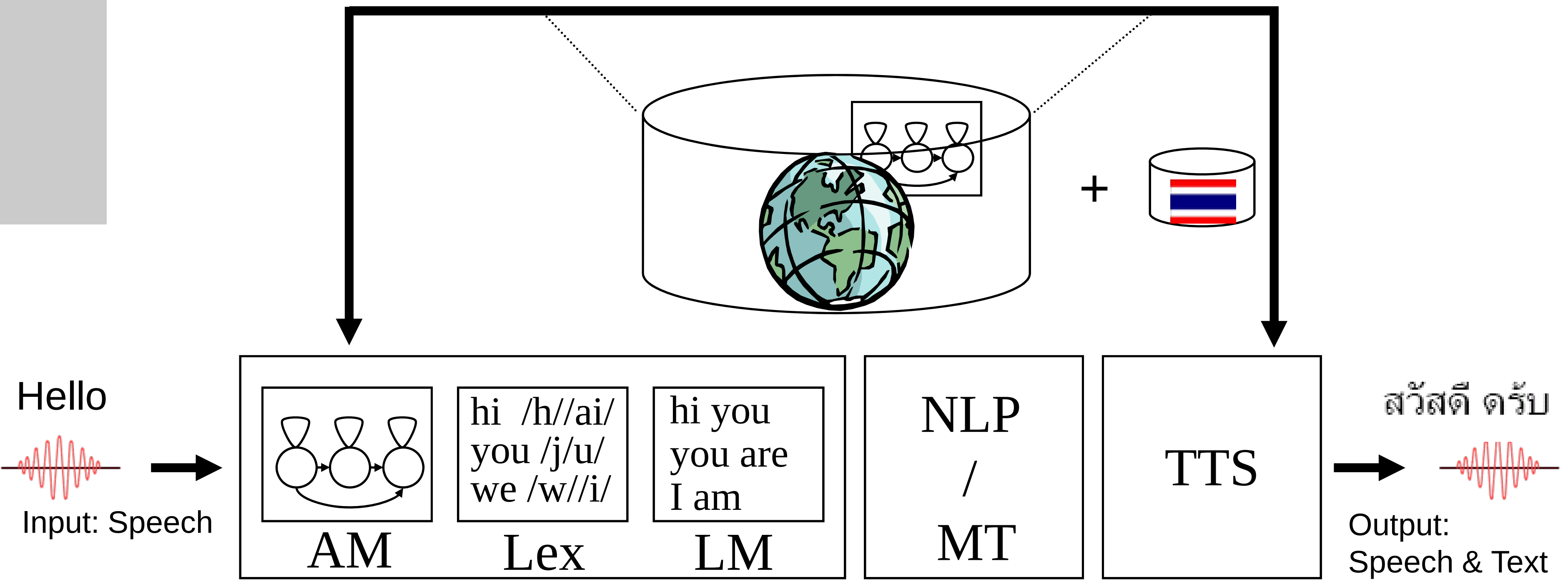


ML-Tag

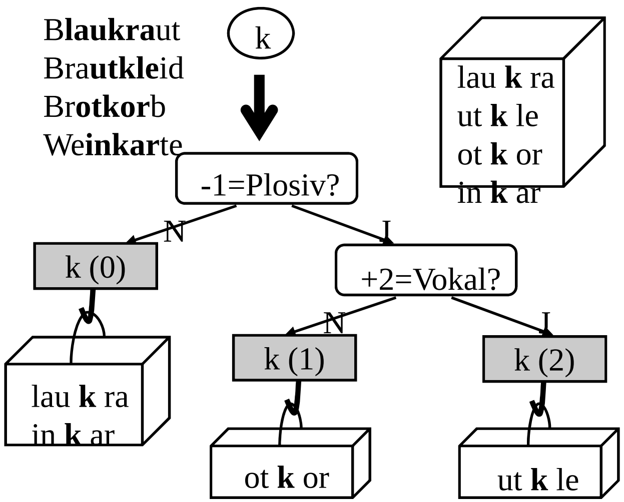
Rapid Portability: Acoustic Models

Step 3:

- Define mapping between ML set and new language
- Bootstrap acoustic model of unseen language



Polyphone Decision Tree Adaptation



Problem:

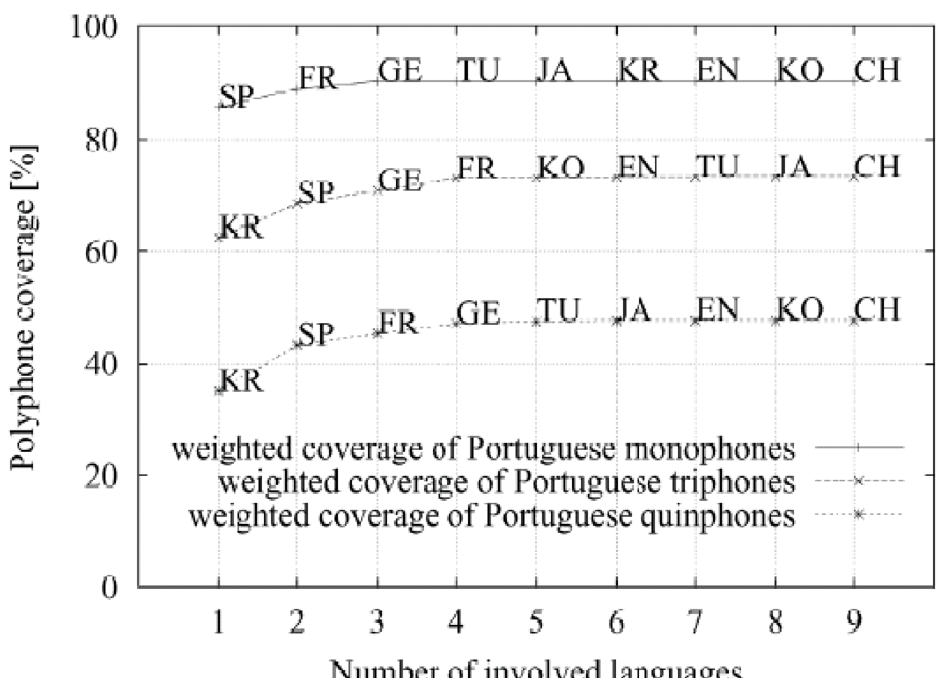
Context of sounds are language specific

How to train context dependent models for new languages?

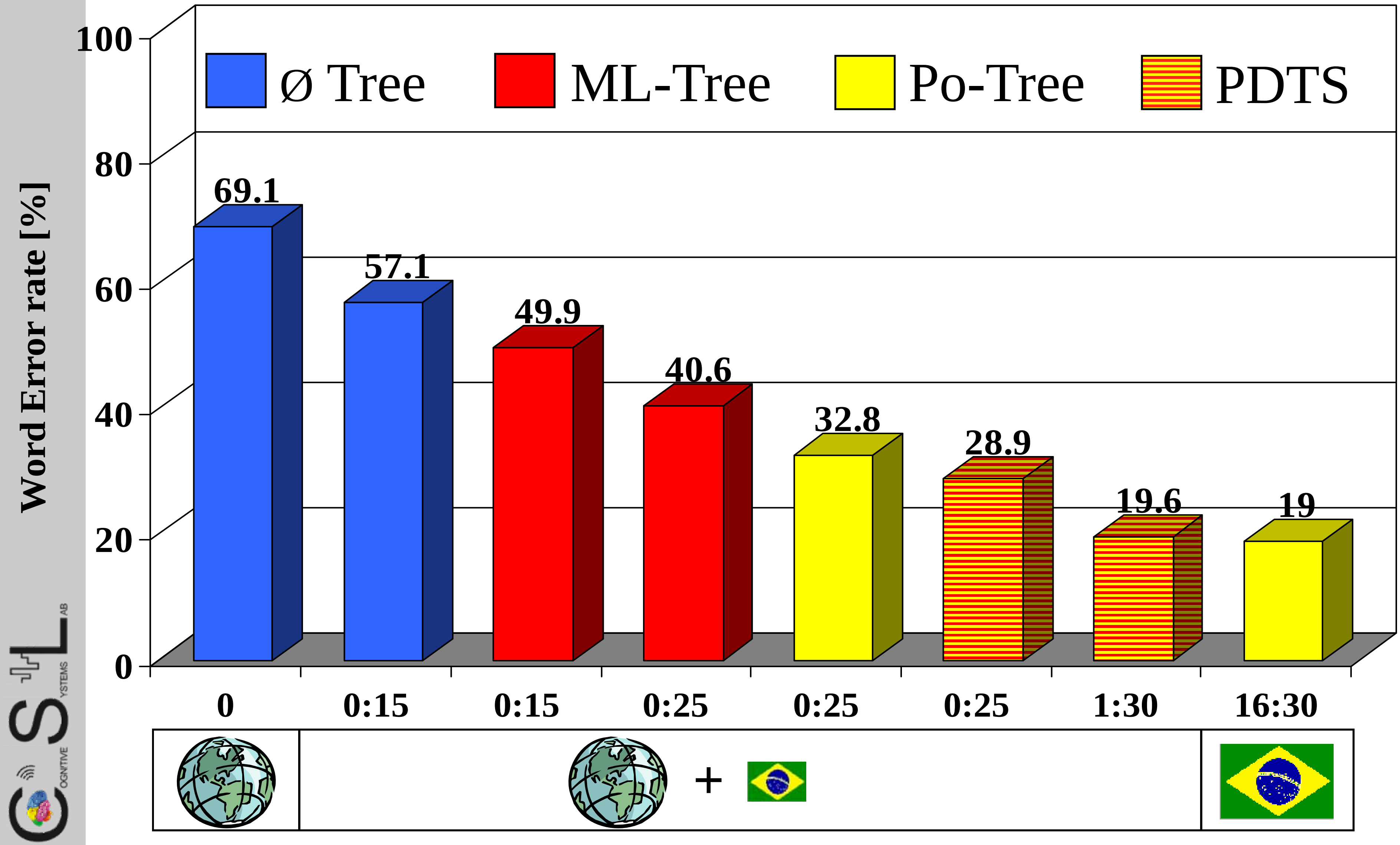
Solution:

1) Multilingual Decision Context Trees

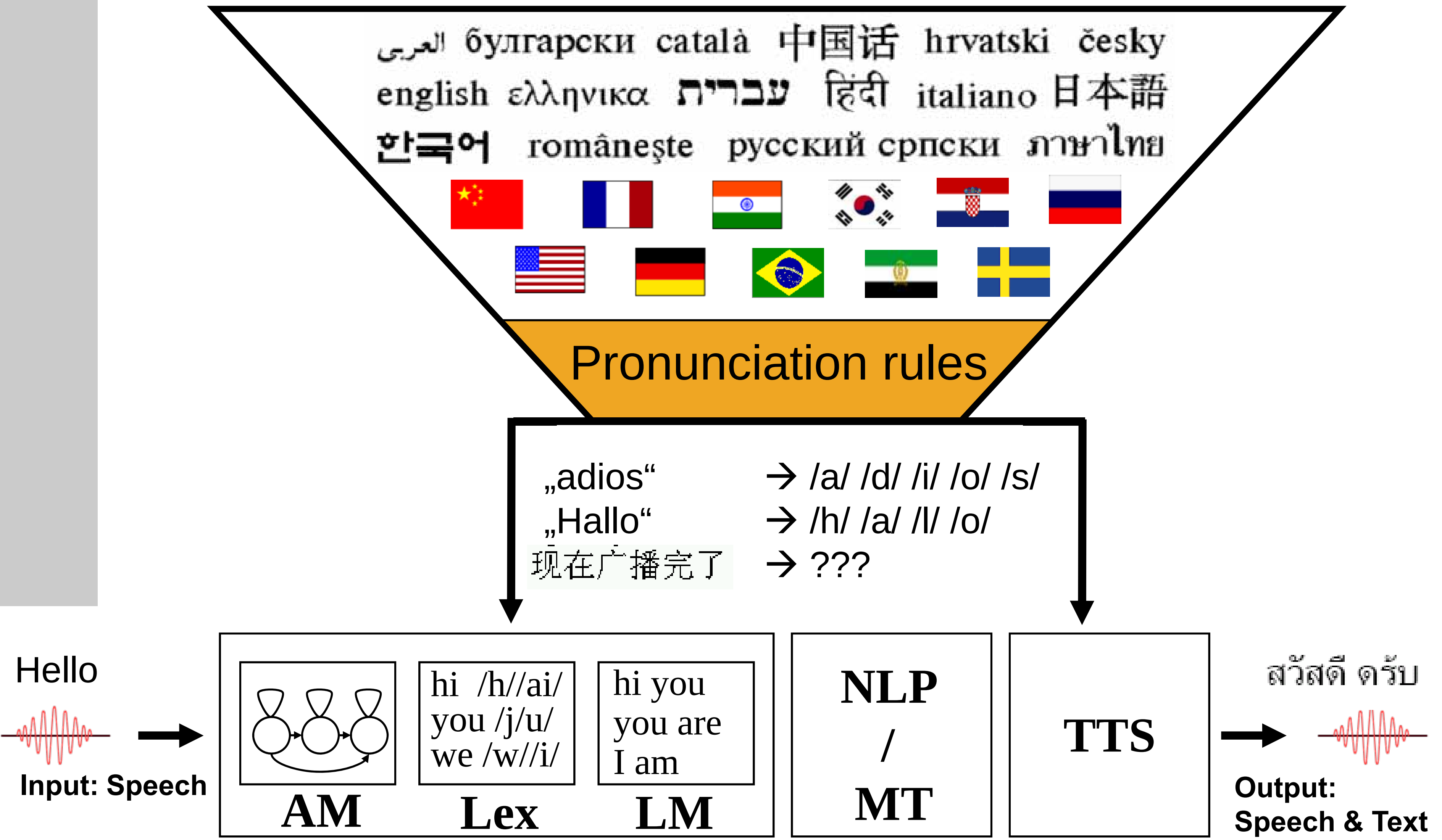
2) Specialize decision tree by Adaptation



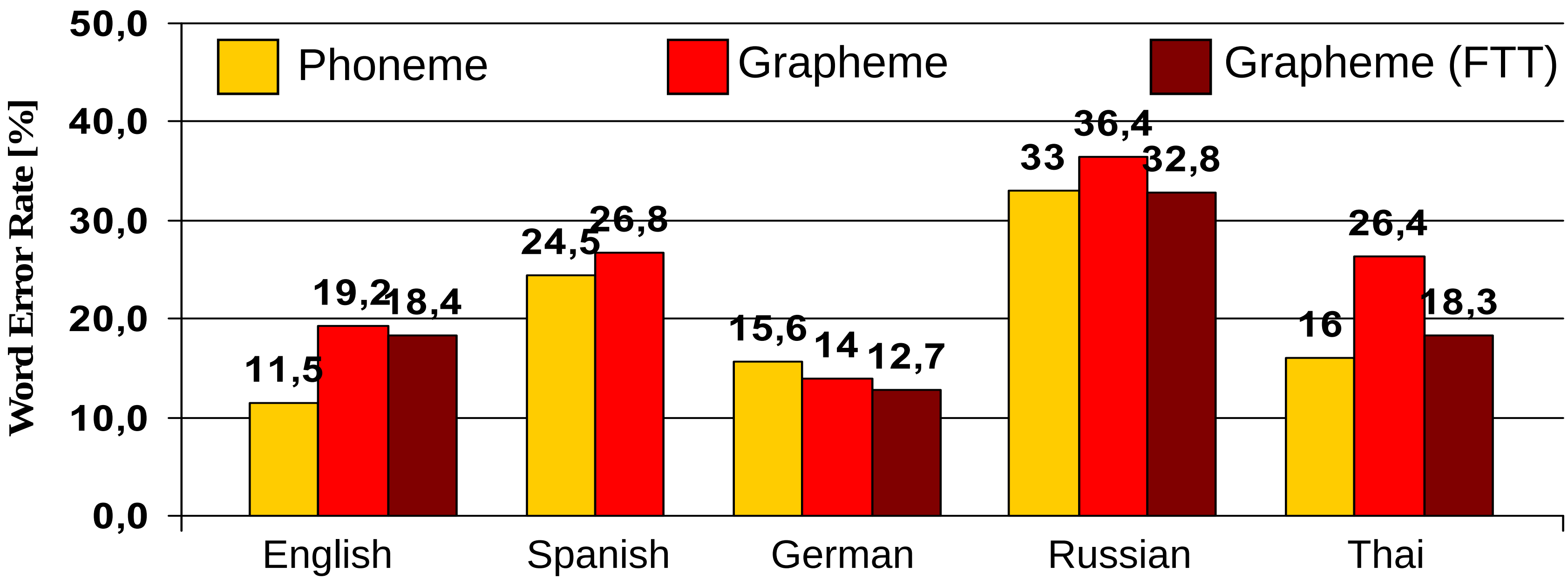
Rapid Portability: Acoustic Model



Rapid Portability: Pronunciation Dictionary



Phoneme- vs Grapheme based ASR



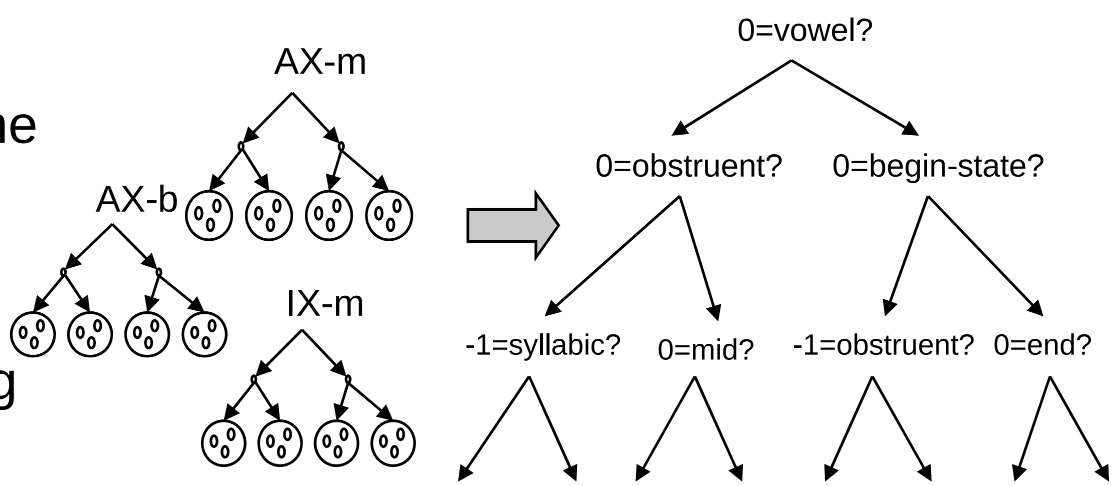
Problem:

- 1 Grapheme ≠ 1 Phoneme

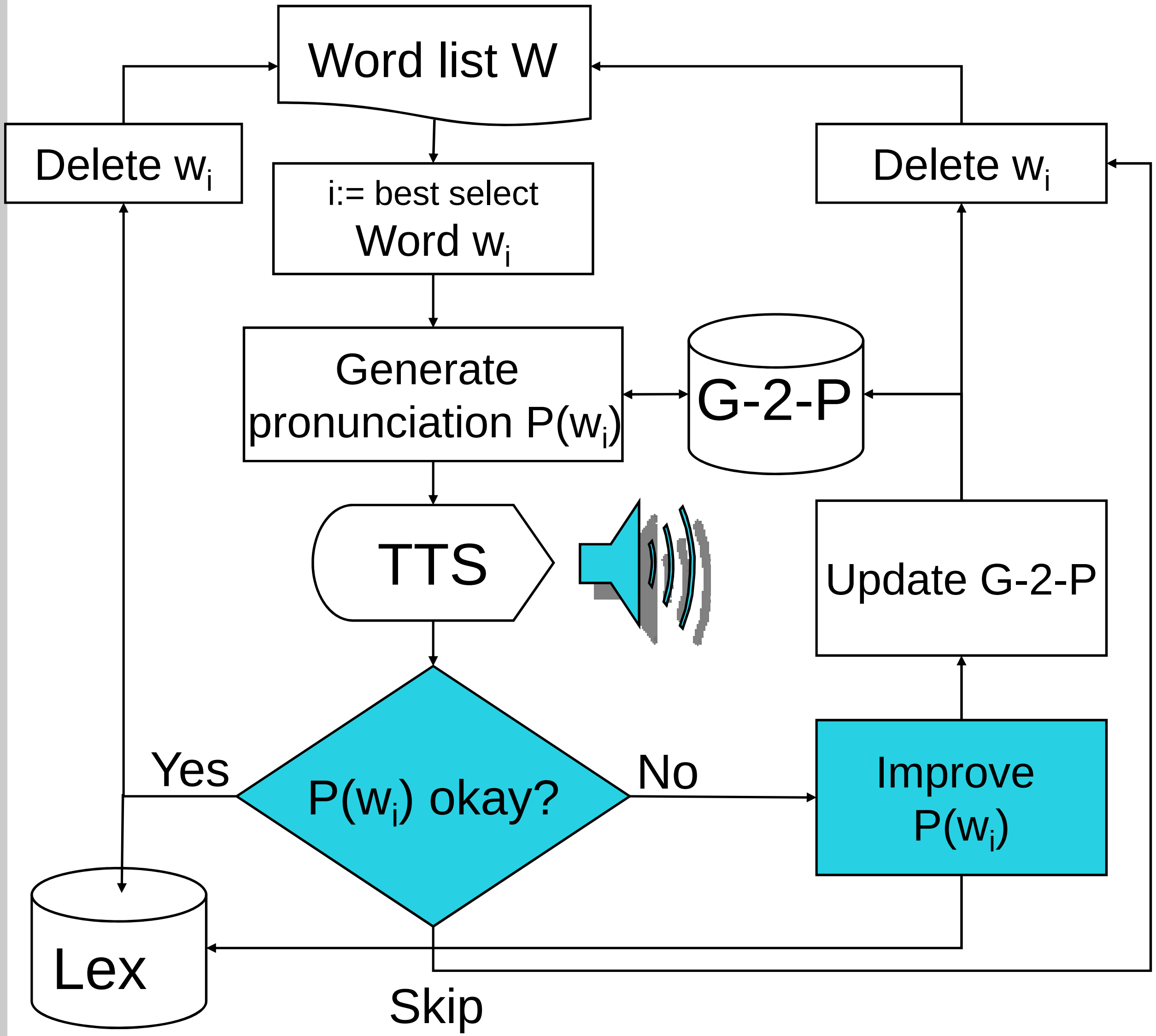
Flexible Tree Tying (FTT):

One decision tree

- Improved parameter tying
- Less over specification
- Fewer inconsistencies



Dictionary: Interactive Learning



* Follow the work of Davel&Barnard

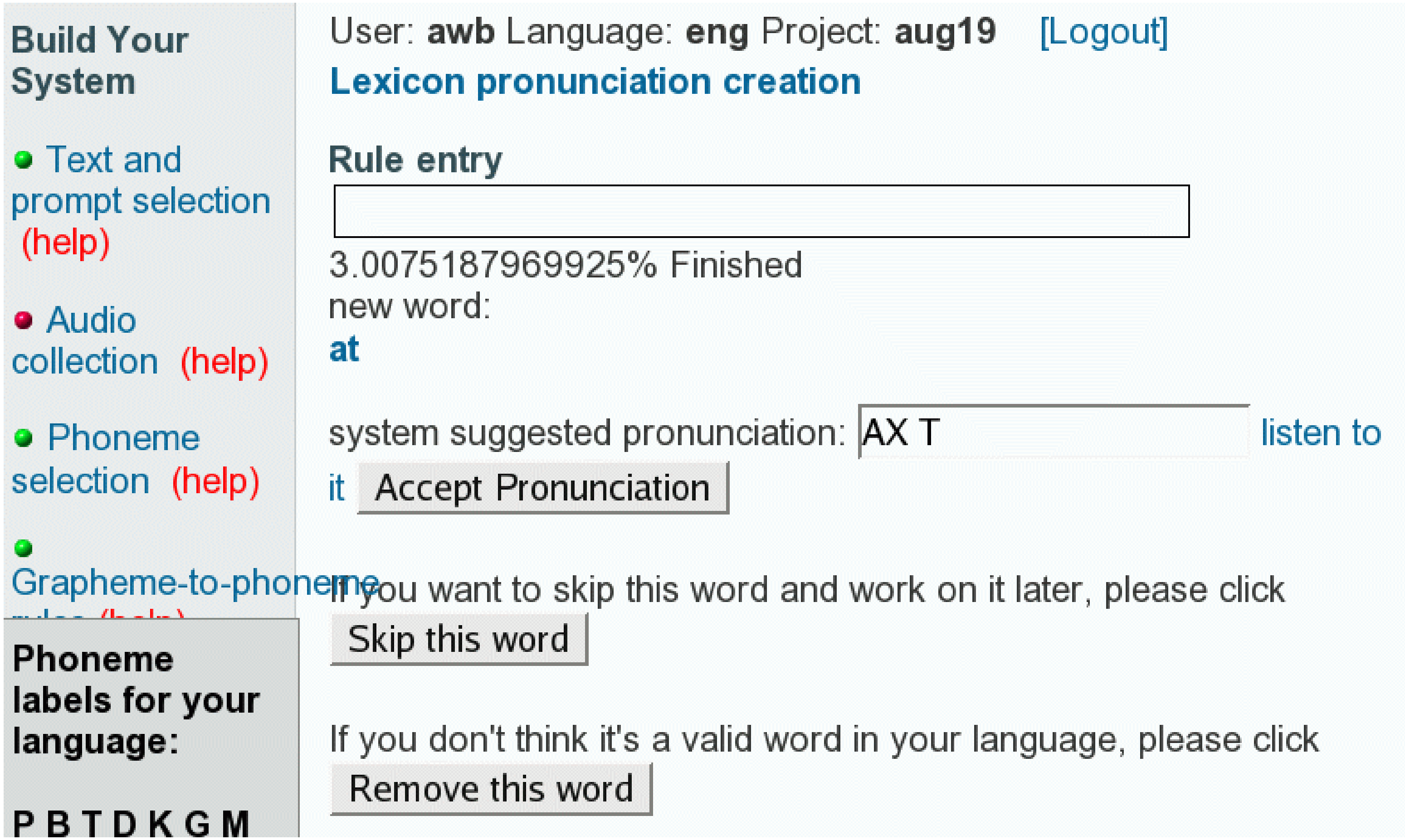
* Word list: extract from text

* G-2-P
- explicit map rules
- neural networks
- decision trees
- instance learning (grapheme context)

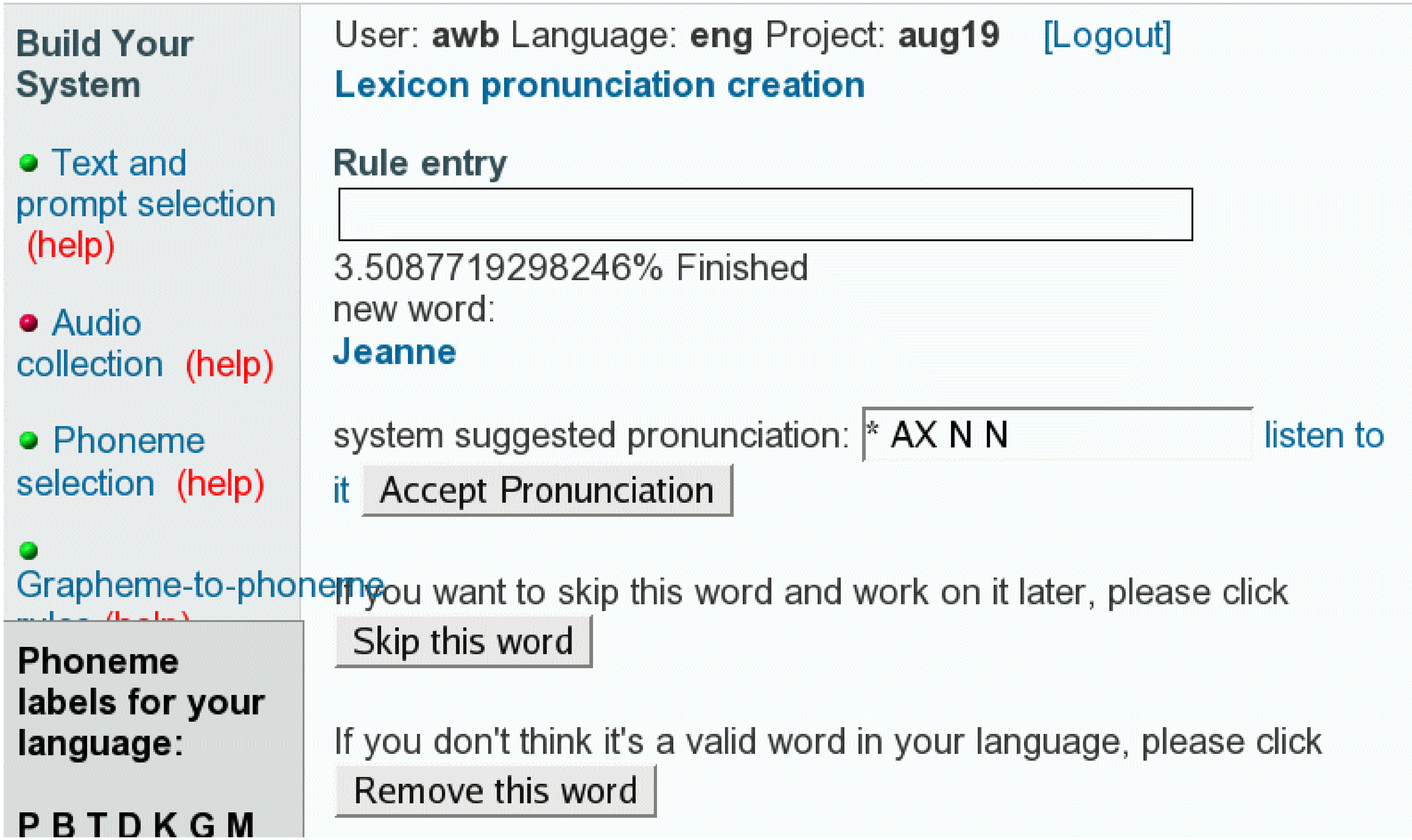
* Update after each w_i
→ effective training

User

Lex Learner


Build Your System
Text and prompt selection (help)
Audio collection (help)
Phoneme selection (help)
Grapheme-to-phoneme
Phoneme labels for your language:
P B T D K G M
User: awb Language: eng Project: aug19 [Logout]
Lexicon pronunciation creation
Rule entry
3.0075187969925% Finished
new word:
at
system suggested pronunciation: AX T
listen to it
Accept Pronunciation
If you want to skip this word and work on it later, please click
Skip this word
If you don't think it's a valid word in your language, please click
Remove this word

Lex Learner


Build Your System
Text and prompt selection (help)
Audio collection (help)
Phoneme selection (help)
Grapheme-to-phoneme
Phoneme labels for your language:
P B T D K G M
User: awb Language: eng Project: aug19 [Logout]
Lexicon pronunciation creation
Rule entry
3.5087719298246% Finished
new word:
Jeanne
system suggested pronunciation:
* AX N N
listen to it
Accept Pronunciation
If you want to skip this word and work on it later, please click
Skip this word
If you don't think it's a valid word in your language, please click
Remove this word

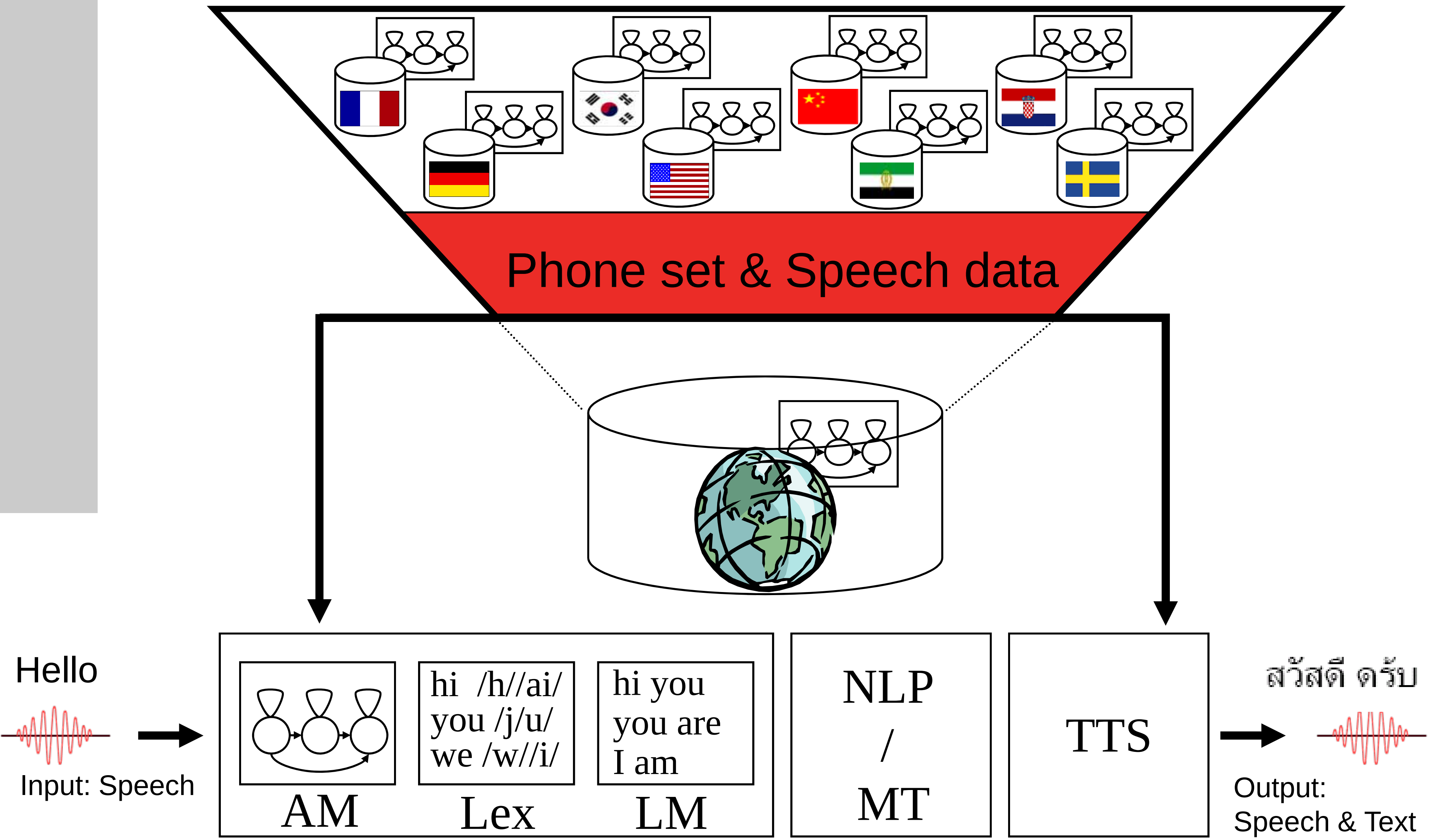
Issues and Challenges

- How to make best use of the human?
  - Definition of successful completion
  - Which words to present in what order
  - How to be robust against mistakes
  - Feedback that keeps users motivated to continue

- How many words to be solicited?
  - G2P complexity depends on the language (SP easy, EN hard)
  - 80% coverage
    hundred (SP) to thousands (EN)
  - G2P rule system perplexity

Language	Perplexity
English	50.11
Dutch	16.80
German	16.70
Afrikaans	11.48
Italian	3.52
Spanish	1.21

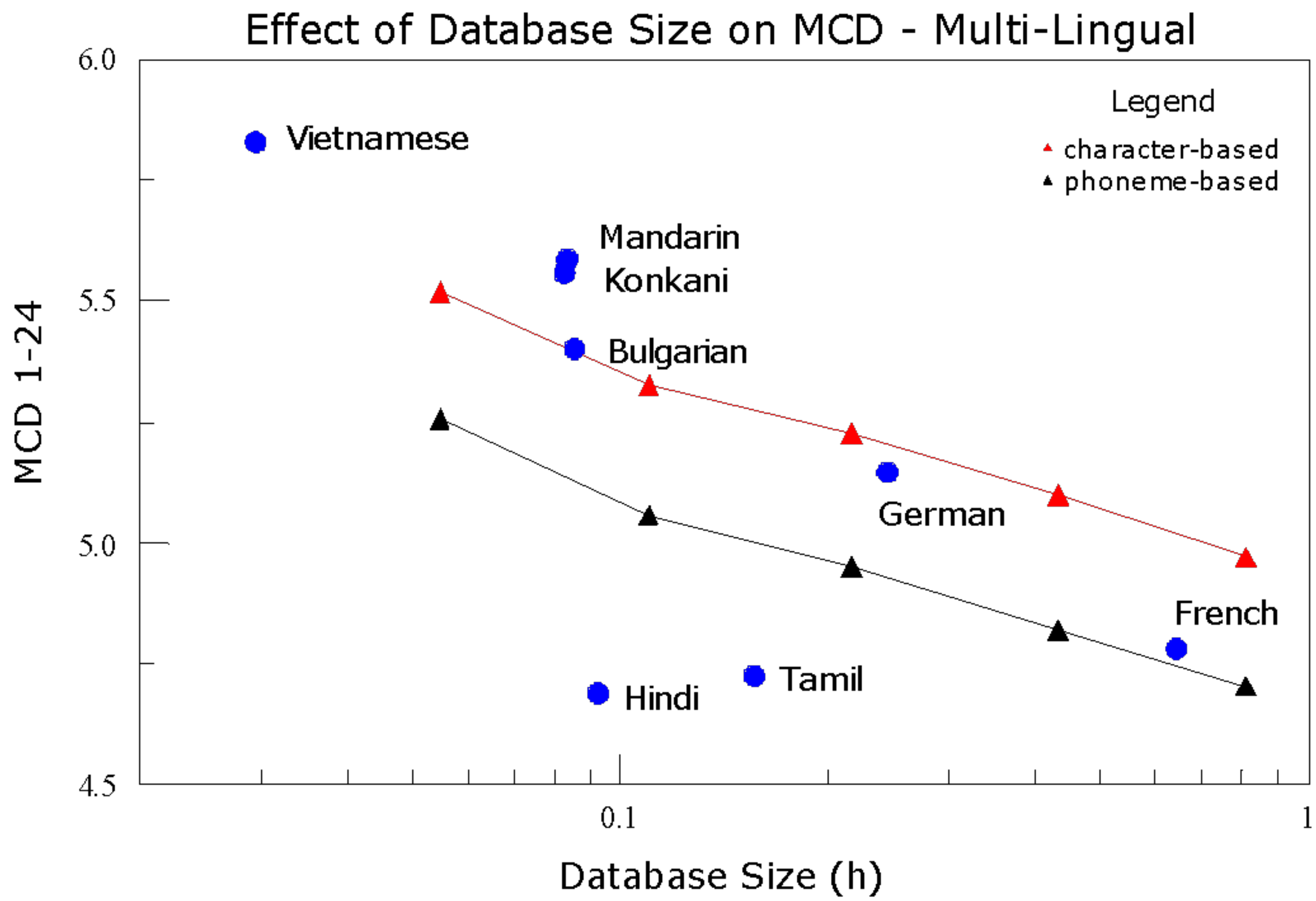
Rapid Portability: TTS



Statistical Parametric TTS

- Text-to-speech for Applications, Common technologies:
  - Diphone: too hard to record and label
  - Unit selection: too much to record and label
- Statistical Parametric Synthesis: "just right"
  - "HMM synthesis": ***clustergen*** trajectory synthesis
  - Clusters representing context-dependent allophones
  - Works robustly with as little as 10min speech data
  - But … Signal may sound "buzzy", can lack varied prosody
- Voice Building Process
  - Collect 300-500 utterances from single speaker, rich prompt set
  - Lexical coverage (from Lex Learner)
  - Automatic labeling from acoustic models
  - Automatic: spectral and prosodic models
- http://festvox.org [Black and Lenzo 2000]
  - Documentation, Tools, Scripts, Examples

TTS with very litte Data



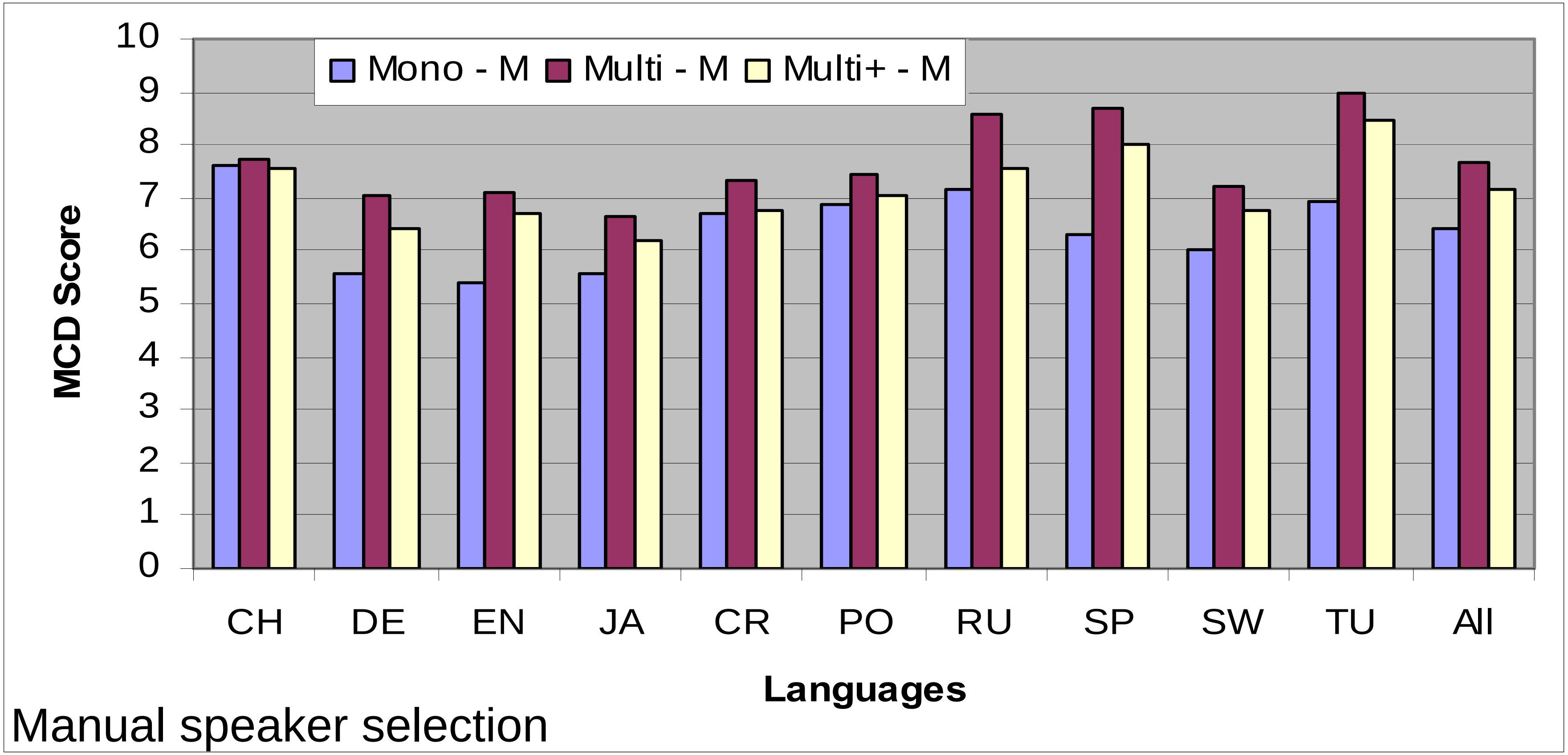
Rule of Thumb for getting the best gain per amount of labor

≤ 30-60min speech: collect additional data

> 60min speech: improve lexicon

Kominek, J., Schultz, T., Black, A. *Synthesizer Voice Quality of New Languages Calibrated with Mean Mel Cepstral Distortion, SLTU-2008 Workshop, Hanoi, Vietnam.*

Mono vs. Multilingual Models



⇒ For all languages monolingual TTS performs best

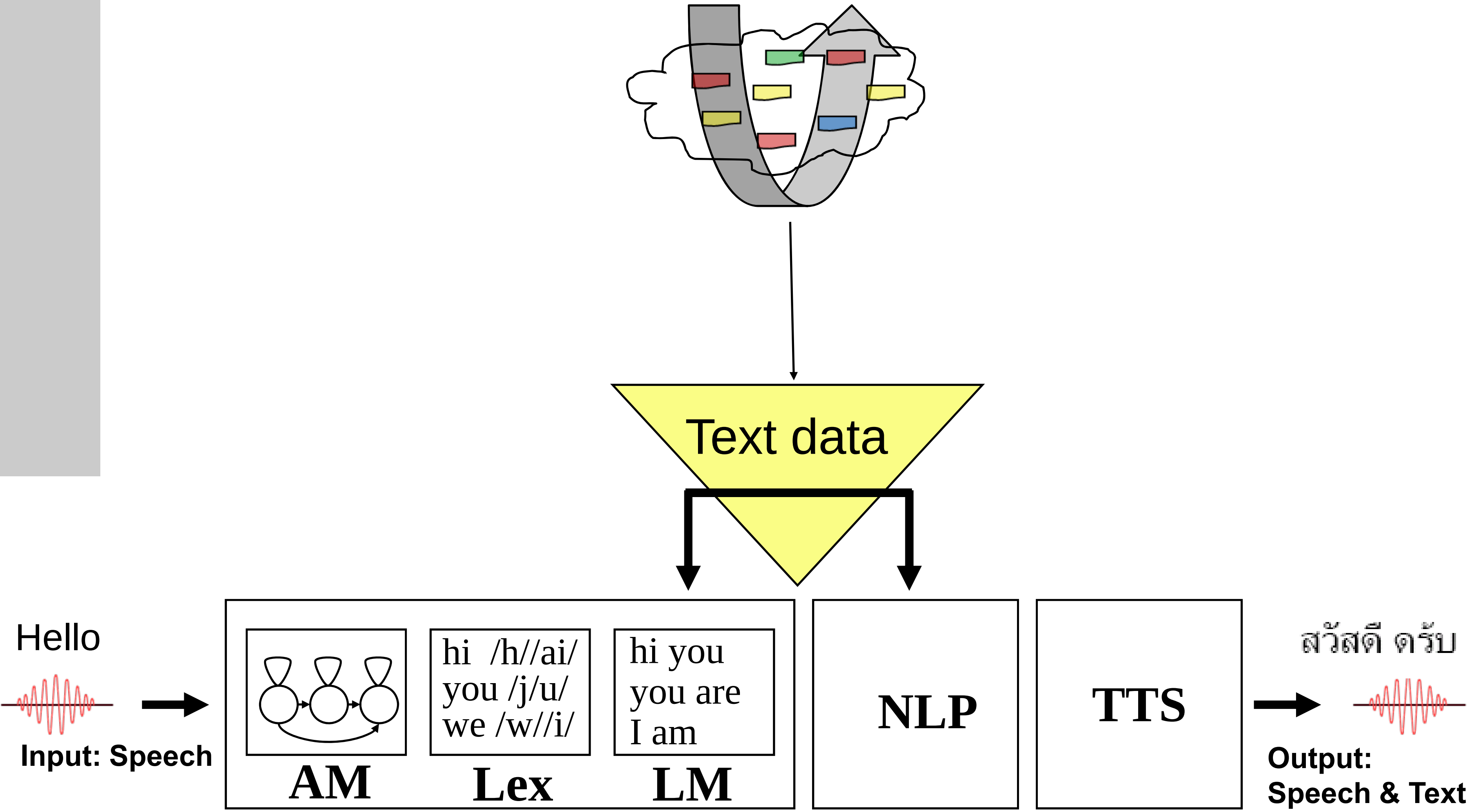
⇒ Multilingual Models perform well …

… only if knowledge about language is preserved (Multi+)

(only small amount of sharing actually happens)

Rapid Portability: Language Modeling

Focused Re-crawling



Language Model Building

Goal: Get as much relevant text data as possible

- Use the retrieved text data for
  - Generating recording prompts
  - Generating vocabulary lists
  - Build Language Models for ASR

Approach

1. User provides an URL or Text or Vocab list
2. Crawler retrieves N documents (web-pages)
3. Compute the statistics (TF-IDF) from the N documents
4. Terms with highest TF-IDF score form query terms
5. User may check terms for in/exclusion
6. Search engine (Google) gets URLs for the query terms
7. Crawl the top K URLs for the data

Case Study with very small data - Hindi

- Targeted Domain in Hindi: Cooking recipes
- Data: 192 sentences, 1,523 words = 13min speech, 1 spk
- Use speech to adapt multi-lingual acoustic models
- Use transcripts to build bigram LM1
- LM2: Expanded by focused re-crawling to 159,995 words
- LM3: Expanded to 360,395 words

- Three evaluation sets (spoken by same speaker)

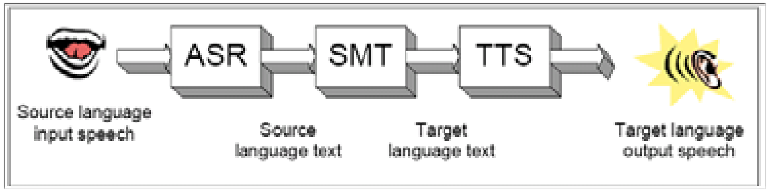
⇒ Focused recrawling significantly reduces the OOV rate and thus WER

LM	*word count*	*Word Error Rate (WER) (%) perplexity / OOV rate (%)*			
		split 1	*split 2*	*split 3*	*ave.*
1	1523	95.88 5.2/68.7	97.92 6.9/57.9	84.93 7.8/50.0	92.91 6.6/58.9
2	159995	55.15 177/16.8	56.25 93.4/27.4	51.81 165/13.4	54.41 145/19.2
3	360395	54.12 214/15.0	52.08 113/25.0	50.60 187/11.3	52.27 171/17.1

John Kominek, Sameer Badaskar, Tanja Schultz, Alan W Black, IMPROVING SPEECH SYSTEMS BUILT FROM VERY LITTLE DATA, Interspeech 2008, Brisbane

SPICE 2005: Afrikaans – English

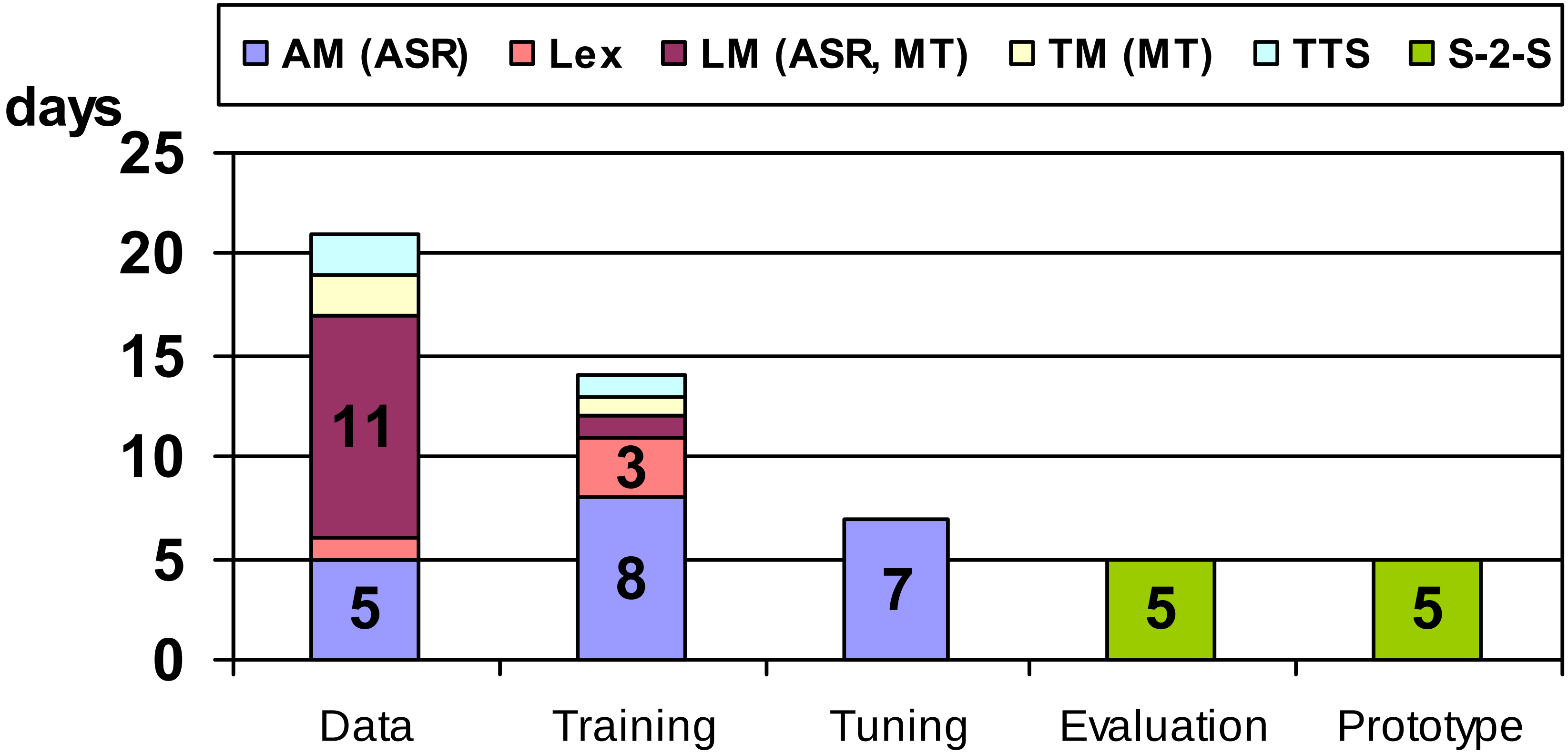
- Goal: Build Afrikaans – English Speech Translation System with SPICE
  - Cooperation with University Stellenbosch and ARMSCOR
  - Bilingual PhD visited CMU for 3 month
  - Afrikaans: Related to Dutch and English,
    g-2-p very close, regular grammar, simple morphology
- SPICE, all components apply statistical modeling paradigm
  - ASR: HMMs, N-gram LM (JRTk-ISL)
  - MT: Statistical MT (SMT-ISL)
  - TTS: Unit-Selection (Festival)
  - Dictionary: G-2-P rules using CART decision trees



- Text: 39 hansards; 680k words; 43k bilingual aligned sentence pairs;
  Audio: 6 hours read speech; 10k utterances, telephone speech (AST)

Time Effort

- Good results: ASR 20% WER; MT A-E (E-A) Bleu 34.1 (34.7), Nist 7.6 (7.9)
- Shared pronunciation dictionaries (for ASR+TTS) and LM (for ASR+MT)
- Most time consuming process: data preparation → reduce amount of data!
- Still too much expert knowledge required (e.g. ASR parameter tuning!)



Herman Engelbrecht, Tanja Schultz, Rapid Development of an Afrikaans-English Speech-to-Speech Translator, IWSLT 2005, Pittsburgh, PA, October 2005

SPICE 2007: Field Experiments

- Now targeting *more* languages in a *shorter* time frame
- 6-weeks Hands-on Course at CMU in Spring 2007
  - Adopt native languages of participating students as targets
  - Added up to 10 different languages: Bulgarian, English, French, German, Hindi, Konkani, Mandarin, Telugu, Turkish, Vietnamese
- Teams of two students with different native language
- Course goal was to build a simple S-2-S system and use this to communicate with each other in their mother tongue
  - Solely rely on SPICE tools
  - Build speech recognition components in two languages
  - Build simple SMT component in two directions
  - Build speech synthesis components in two languages
  - Report back on problems and system shortfalls

Schultz, T., Black, A., Badaskar, S., Hornyak, M., Kominek, J., *SPICE: Web-based Tools for Rapid Language Adaptation in Speech Processing Systems, Interspeech 2007, Antwerp.*

Field Experiments (2)

- The 10 languages cover broad range of peculiarities
- Writing system:
  - Logographic Hanzi (Mandarin);
  - Cyrillic (Bulgarian);
  - Roman (German, French and English);
  - phonographic segmental (Telugu and Hindi);
  - phonographic featural (Vietnamese)
  - No script: Konkani
- Segmentation: No segmentation (Chinese); Segmentation white spaces do not necessarily indicate word (Vietnamese)
- Morphology: simple, low inflecting (English), compounding (German), agglutinating (Turkish) …
- Sound System: tonal (Mandarin and Vietnamese), stress (Bulgarian)
- G-2-P: straightforward (Turkish), challenging (Hindi), difficult (English), no relationship (Chinese), invented (Konkani)

Lessons Learned

- It is possible to create speech processing components for 10 languages in 6-weeks using SPICE
- Each language brings new challenges
- Many SPICE features turned out to be very helpful, e.g. only ONE speaker of Konkani in Pittsburgh, web recorder allowed remote collection of more speakers

- Log: time spent in SPICE interface
- Improve interface using breakdown
- Use feedback
- Interface allows for collaborative work

Task	Time Spent [hh:mm]
Text Collection	8:35
Audio Collection	10:07
Phoneme Selection	4:05
LM building	1:25
G-2-P specs	1:30

SPICE 2008: Cross-continental Course

- SPICE-based course between CMU and UKA
  - Students at Carnegie Mellon University, PA
  - Students at Karlsruhe University, Germany
  - Linked by weekly meeting over VC
- Similar to 2007 BUT distributed collaboration
  - Students create ASR & TTS in their native language
  - Bonus for the ambitious: train SMT systems and create a speech-to-speech translation system
- Evaluation includes
  - Time to complete
  - Task difficulties
  - ASR word error rate
  - TTS voice quality
- Fall 2008 course already in progress

Outline

- The World's Languages
  - 6900 languages – So what?
  - Language Extinction – What can the community do about it?
  - Do we need Speech Processing for all of them?
  - Is this really science – not just retraining on a new language?
- Language Characteristics
  - Written form, scripts, letter-to-sound relationship
  - Issues and Differences between languages
- Challenges for Multilingual Speech Processing
  - Lack of Resources (Money, Data, Technical Support)
  - Lack of Experts
- Solutions
  - SPICE: A Rapid Language Adaptation Server
  - Technologies: Leveraging off GlobalPhone & FestVox
  - Experiments and Results
- Conclusions and Future Work

Conclusions

- **Challenges in Multilingual Speech Processing**
  - Well defined build processes: ASR, MT, TTS … BUT:
  - Every new language brings unseen challenges
  - Current (statistical) approaches require lots of data
  - … and native language expert and technology expertise
  - How to bridge the gap between language and tech expert?
- **Proposed solution: SPICE**
  - Learning by interaction from a cooperative (but naïve) user
  - Rapid adaptation from language universal models
  - Knowledge sharing across components
  - Development cycle: Days rather than weeks

Next Steps

- **Continuous Server Support**
  - Improve Interface based on user feedback and lessons learned
  - Improve Language Robustness: font encoding, …
  - Software Engineering, Scaling
- **Collaboration**
  - Multiple people working on the same project
  - Leverage from archived projects
- **Cross-confirmation**
  - Multiple views for within and across project confirmation
  - Confidence measure to find appropriate combination
- **Error-blaming**
  - End-to-end system Evaluation vs Component Evaluation
  - Automatic Generation of Recommendations to improve systems

Try This At Home

- System is online at http://cmuspice.org
- Use system for your own project
  - Create new login/passwd and project
- Preloaded Hindi Example
  - Login as
    - Login: demo
    - Passwd: demo
  - Chose project # (your birth day)
- Book on ML Speech Processing
  Elsevier, Academic Press, 2006