Karlsruhe Institute of Technology

Building Speech Recognition Systems with Low Resources

Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

Tanja Schultz

Cognitive Systems Lab, Institute for Anthropomatics, KIT

Cognitive Systems Lab

ASRU, Limited Resources Day, December 10th 2013, Olomouc, Czech Republik

What is a Low-resourced Language?

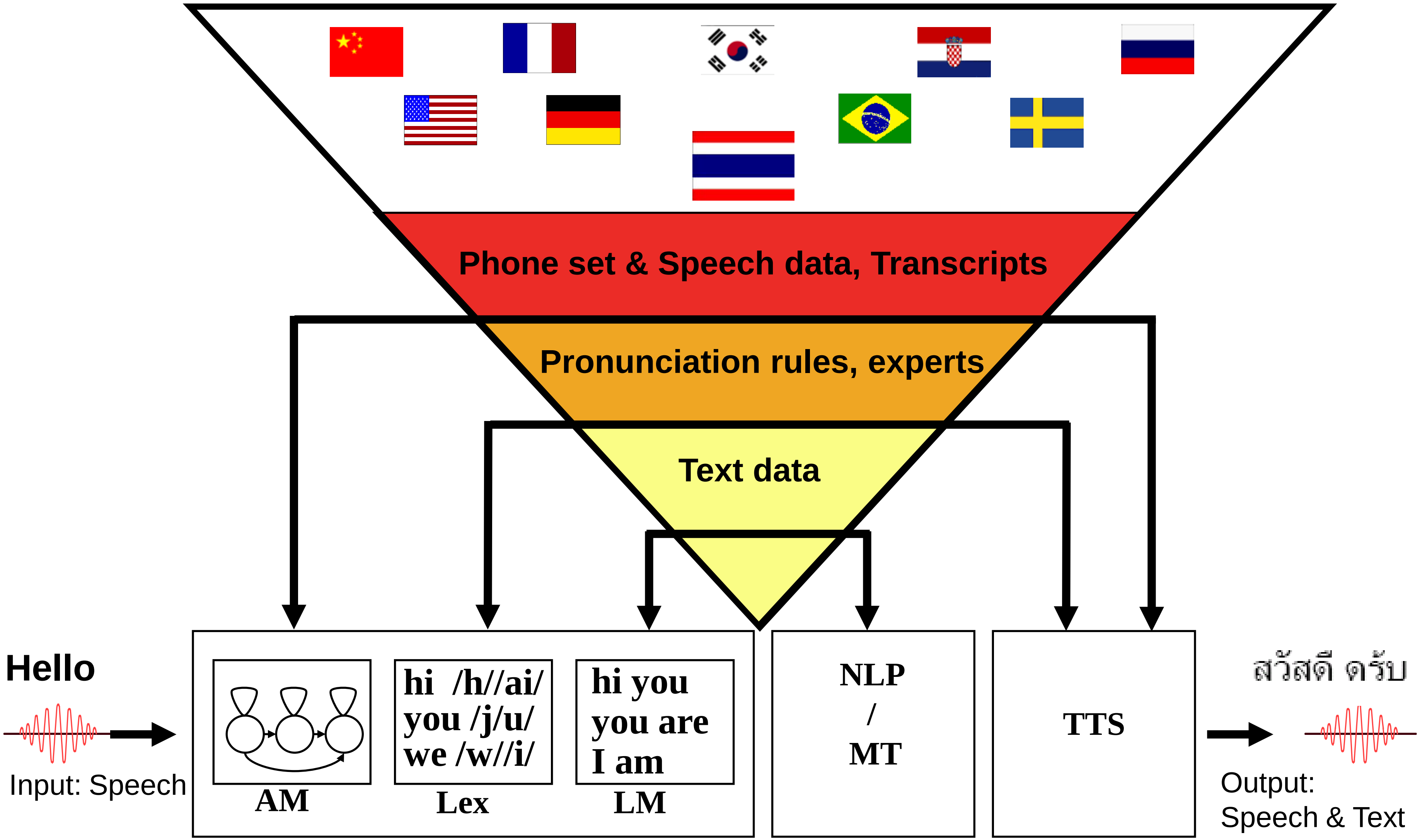
Cognitive Systems Lab

- Definition "under-resourced languages" (Krauwer 2003, Berment 2004) A language with some of (if not all) the following aspects:
  - Lack of **electronic resources** for speech and language processing,
  - Limited **presence on the web**,
  - Lack of a unique **writing system** or stable orthography,
  - Lack of **linguistic expertise**.
- Synonyms: low-density languages, resource-poor languages, low-data languages, less-resourced languages, low-resourced languages
- Low-resourced language ≠ minority language
  - Minority language is spoken by a minority of the population of a territory
  - Some under-resourced languages are official languages of their country and spoken by a very large population (e.g. Khymer)
  - Some minority languages are rather well-resourced (e.g. Catalan)
  - U-R lang. not necessarily endangered (while the opposite is usually true).

Laurent Besacier, Etienne Barnard, Alexey Karpov, Tanja Schultz, *Automatic Speech Recognition for Under-resourced Languages: A Survey*, Speech Communication, vol. 56, pp. 85-100, Jan 2014, http://dx.doi.org/10.1016/j.specom.2013.07.008

Karlsruhe Institute of Technology

The Ideal Case – Plenty of Resources



Tanja Schultz, Katrin Kirchhoff (2006): Multilingual Speech Processing. Elsevier, Academic Press, ISBN 13: 978-0-12-088501-5

Low Resources – Proposed Solutions

Cognitive Systems Lab

<u>Lack of data resources for speech processing</u>

- No Transcripts
  - MUT: Multilingual Unsupervised Transcription System
- No Pronunciation Dictionaries
  - G2P, Wiktionary, Keynounce

<u>Lack of a writing system</u>

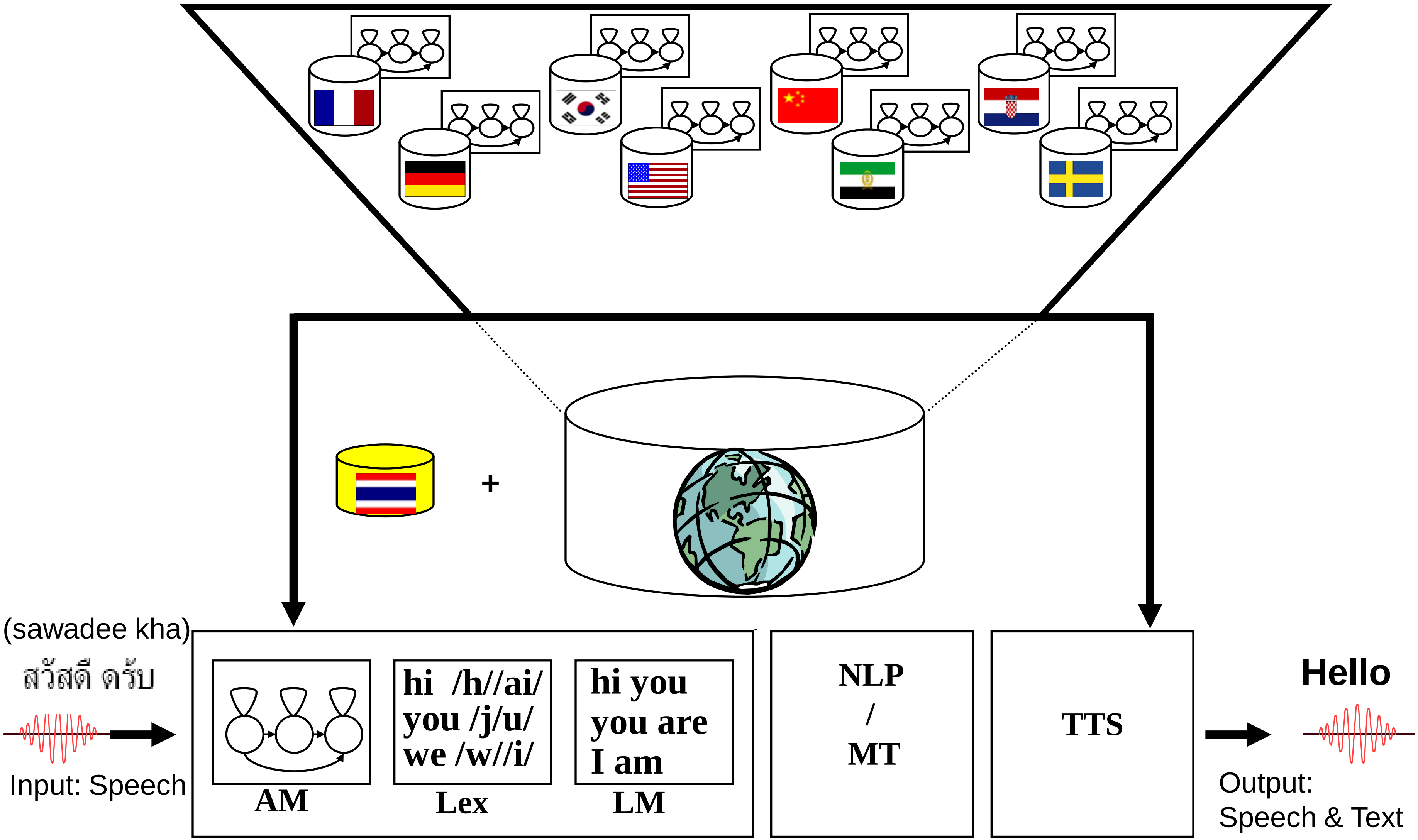
- No Transcripts and No Dictionaries (No writing system)
  - Cross-lingual Word-2-Phoneme alignments

<u>Lack of linguistic expertise</u>

- Web-based Tools RLAT and SPICE

<u>General Approach</u>: Leverage off existing knowledge and data resources from many languages

The Holy Grail – Rapid Adaptation



Tanja Schultz, Katrin Kirchhoff (2006): Multilingual Speech Processing. Elsevier, Academic Press, ISBN 13: 978-0-12-088501-5

GlobalPhone (Clean Speech, transcribed)

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

Multilingual Database

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

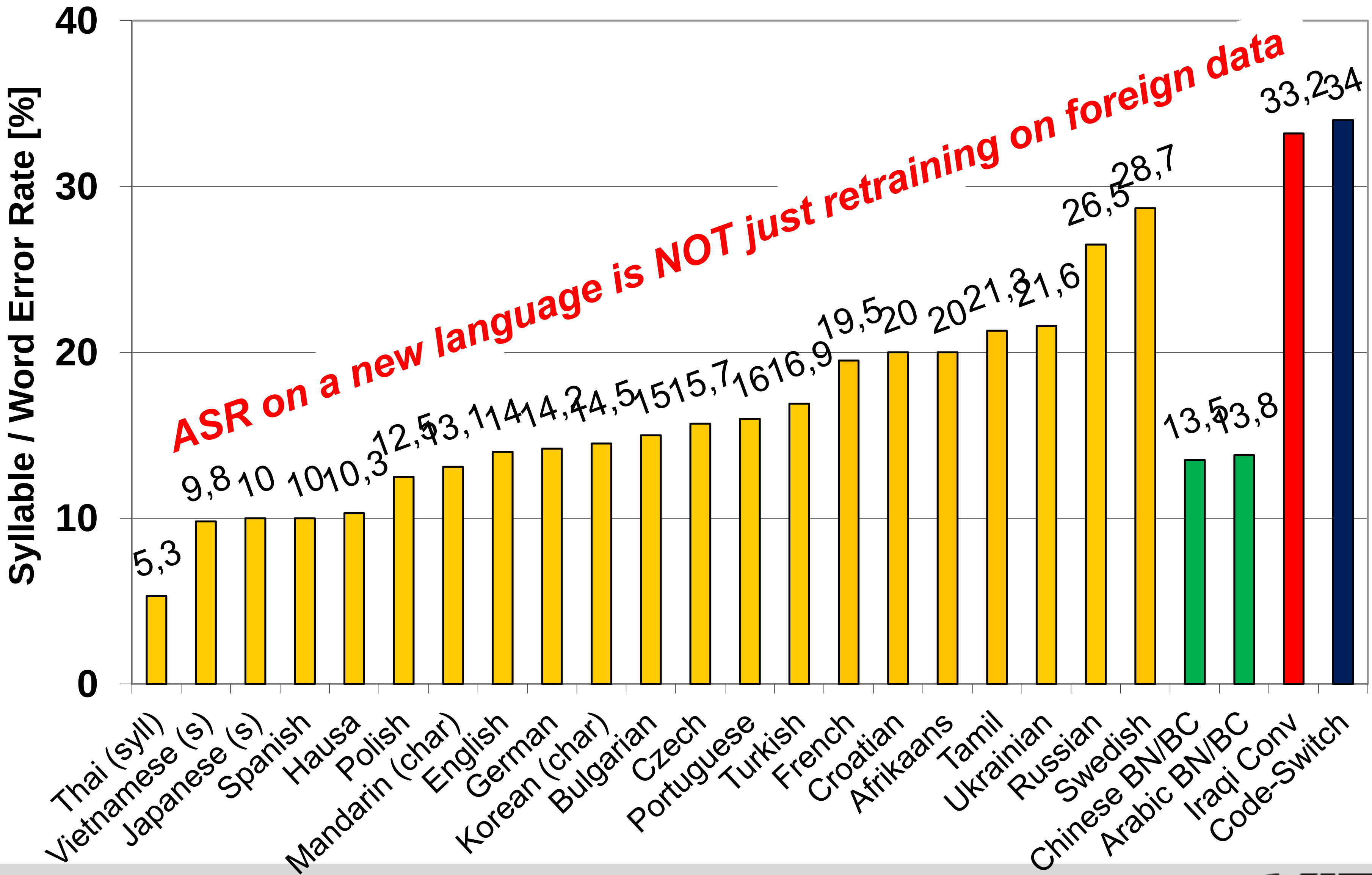
Corpus

- 21 Languages … counting
- ≥ 2000 native speakers
- ≥ 450 hrs Audio data
- Read Speech
- Filled pauses annotated

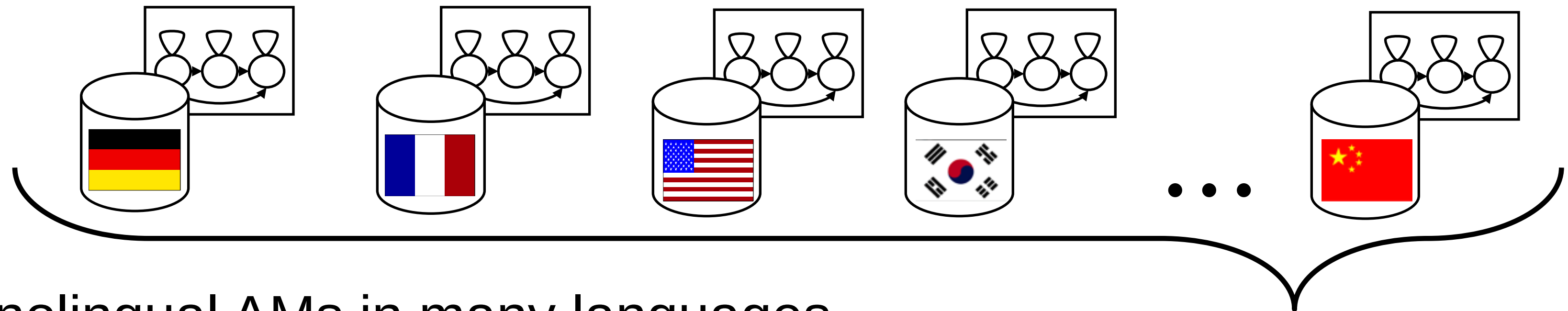
Available from ELRA, Appen

Tanja Schultz (2002): GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University, ICSLP Denver, CO

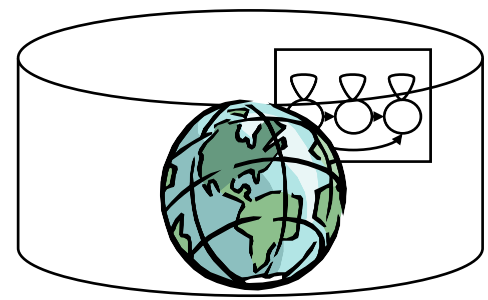
Speech Recognition in many Languages



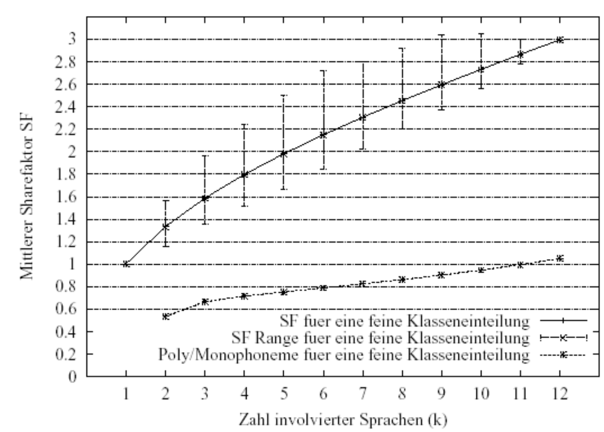
Multilingual Acoustic Modeling



Monolingual AMs in many languages
$\rightarrow$ Multilingual AM “language independent”



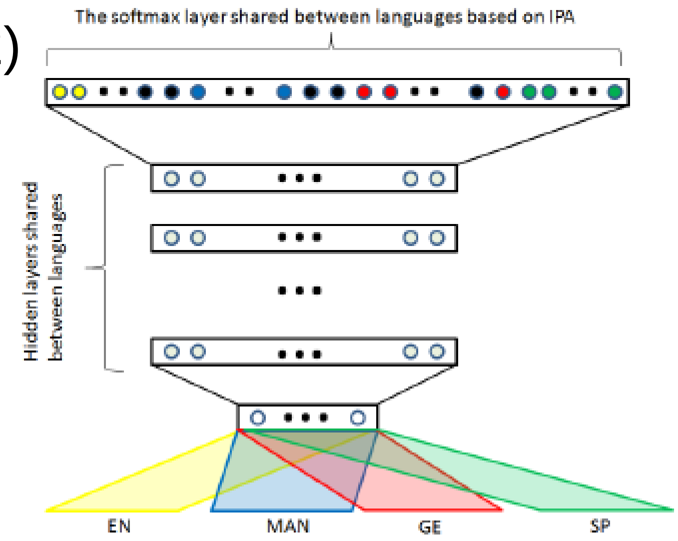
- IPA-based data sharing
  - Köhler 1997, Schultz 1997 (Context-independent)
  - On 12 languages: 485 $\rightarrow$ 162 (sharing factor ~3)
  - Context-dependent Ams, PDTS (Schultz, 1999)
  - Articulatory features (Stüker et al. 2003)
- Mono outperformed ML on training language
- BUT: ML gives benefits on unseen languages



Recent Approaches

- Multilayer Perceptrons (MLP) e.g. Bottle-Neck features
  - Several studies on multilingual and cross-lingual aspects
    E.g. A. Stolcke (2006), K. Livescu (2007), S. Thomas (2011)
  - Open target language MLP (Vu & Schultz 2012)
- Subspace GMMs (Burget, Povey et al., 2010)
- Cross-lingual NN features (Plahl et al., 2011)
- Hybrid HMMs using MLP posteriors (D. Imseng, 2011)
- Deep Neural Networks (Heigold et al., 2012)
- Vu/Imseng: ML DNN w/KL
  - 6 languages, (BG, EN, GE, JA, MA, SP)
    greedy layer-wise supervised (GL-ST)

Systems	CZ	HA	VN
DNN (GL-ST)	9.9	10.1	10.0
DNN-MUL-SEP	9.3	9.8	**8.6**
DNN-MUL-IPA	**9.2**	**9.5**	8.8



Proposed Solutions

Lack of data resources for speech processing

- No Transcripts
  - MUT: Multilingual Unsupervised Transcription System
- No Pronunciation Dictionaries
  - G2P, Wiktionary, Keynounce

Lack of a writing system

- Cross-lingual Word-2-Phoneme alignments

Lack of linguistic expertise

- Web-based Tools RLAT and SPICE

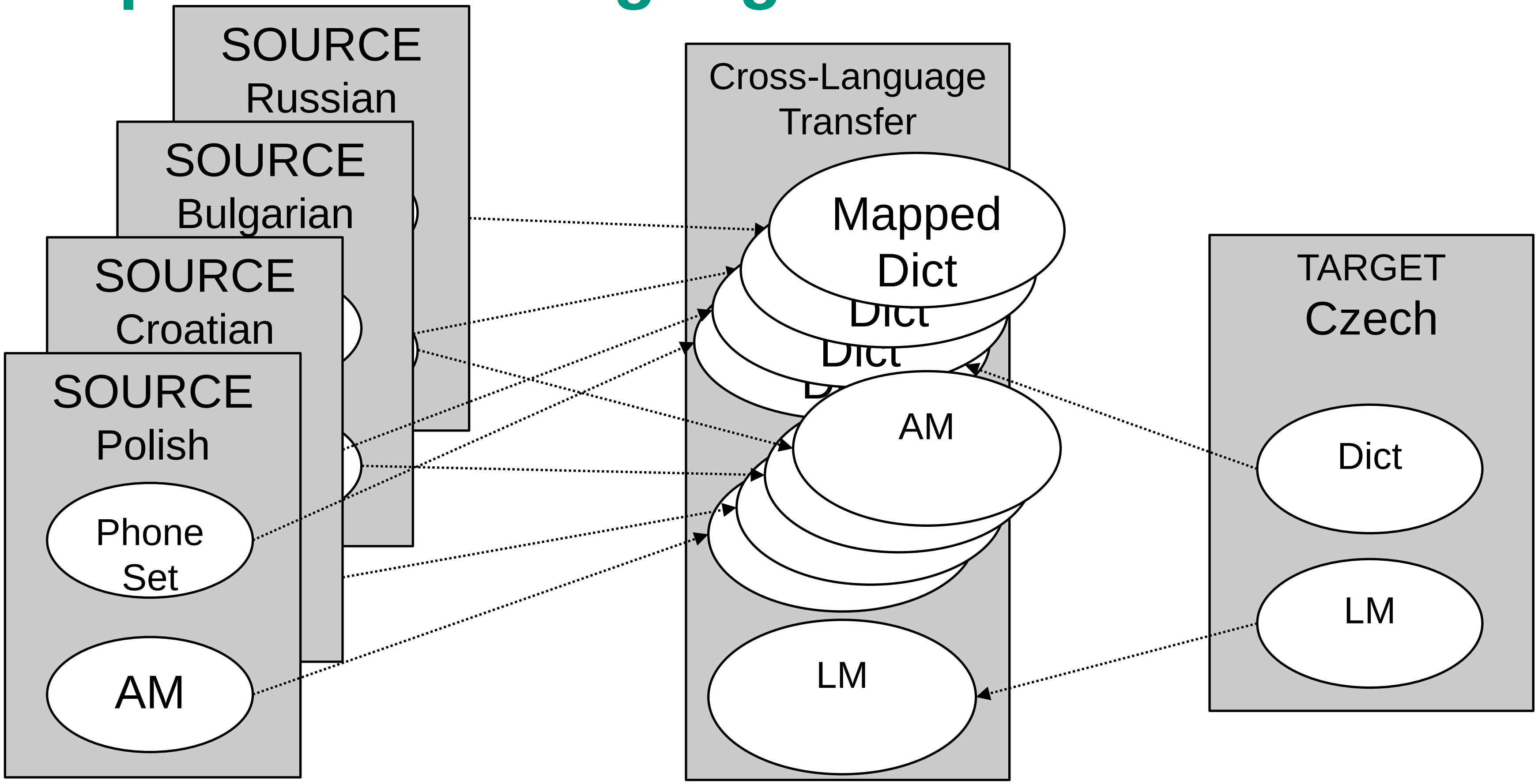
General Approach: Leverage off existing knowledge and data resources from many languages

Experimental Setup

- Wanted: ASR for Czech: (West-Slavic, 12M spks)
  - Assume ~20 hours of Speech, Dict, LM given but **no transcriptions**
- Solution: Leverage off knowledge from MANY languages
  - Given: Data, Transcripts, ASR for several languages (~20h each)
- ASR for 4 Slavic Languages (GlobalPhone)
  - Croatian (South-Slavic, 7M spks); Russian (East-Slavic, 165M spks)
  - Bulgarian (South-Slavic, 12M spks); Polish (West-Slavic, 56M spks)
- ASR for resource rich languages:
  - English,
  - French,
  - German,
  - Spanish

Language	WER	LM-Perplexity	OOV-rate	Vocab
BL	22.1%	543	1.3%	24 K
EN	15.4%	284	0.5%	64 K
FR	22.3%	352	2.4%	122 K
GE	13.2%	148	0.4%	39 K
HR	28.9%	813	3.6%	362 K
PL	18.9%	1373	4.1%	36 K
RU	35.2%	1684	2.8%	293 K
SP	23.3%	224	0.1%	31 K

Step 1: Cross-Language Transfer

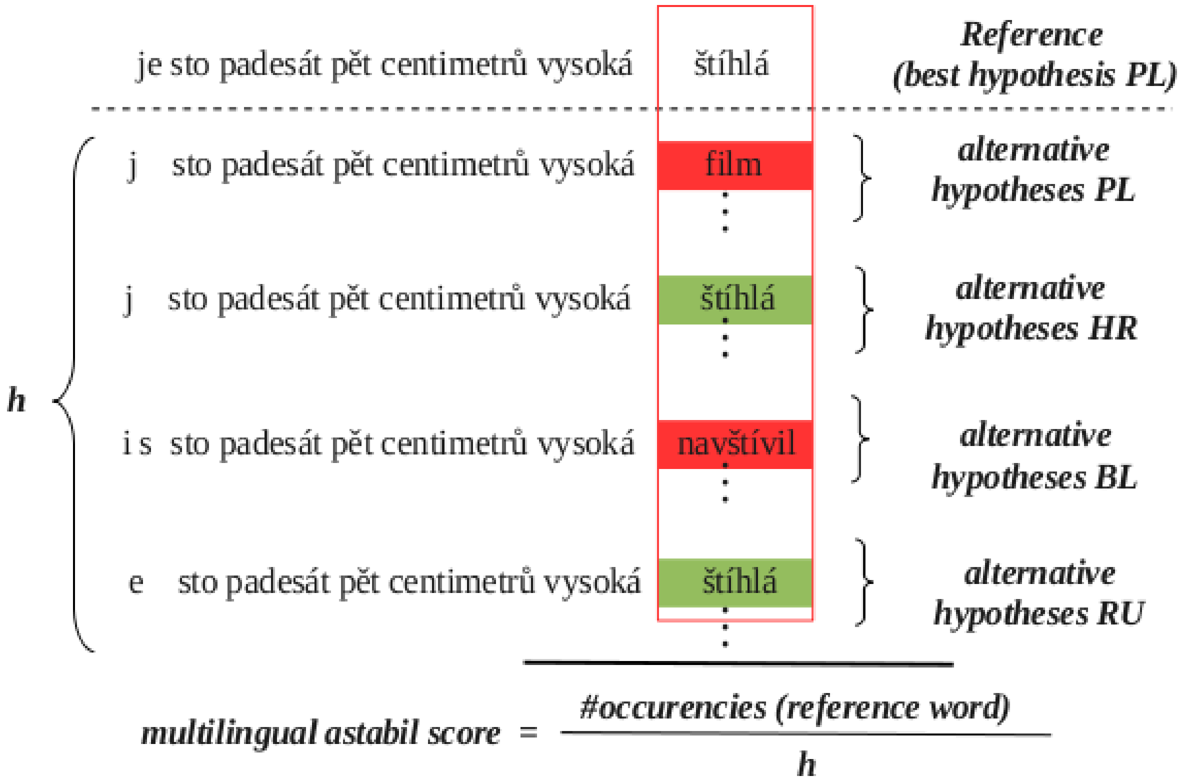


- Modify target language dictionary (phones from source language)
- Apply source AM to decode Czech speech data
  Unsupervised Training, Zavaliagkos&Colthurst ‘98, Kemp ‘99, Lamel ‘00

N.T. Vu, F. Kraus, T. Schultz. Cross-language bootstrapping based on completely unsupervised training. ICASSP, 2011.

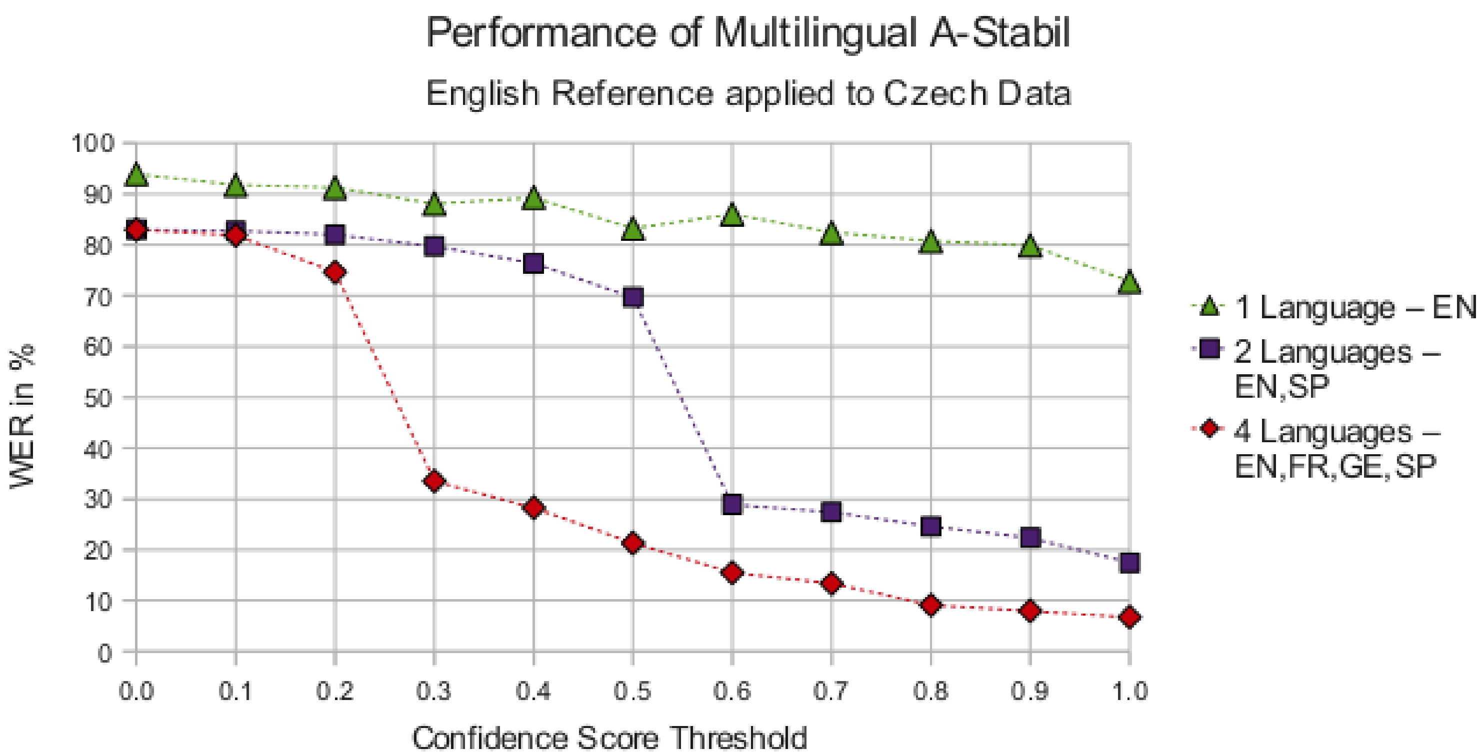
Step 2: Multilingual A-Stabil

- Word-based confidence measure based on word lattices (Kemp, Schaaf 1999)
- A-stabil = *acoustic stability:* frequency of a word over several hypotheses
- Apply to multilingual setting – hypotheses from different languages
- Languages agree on the same word → higher probability that it is correct



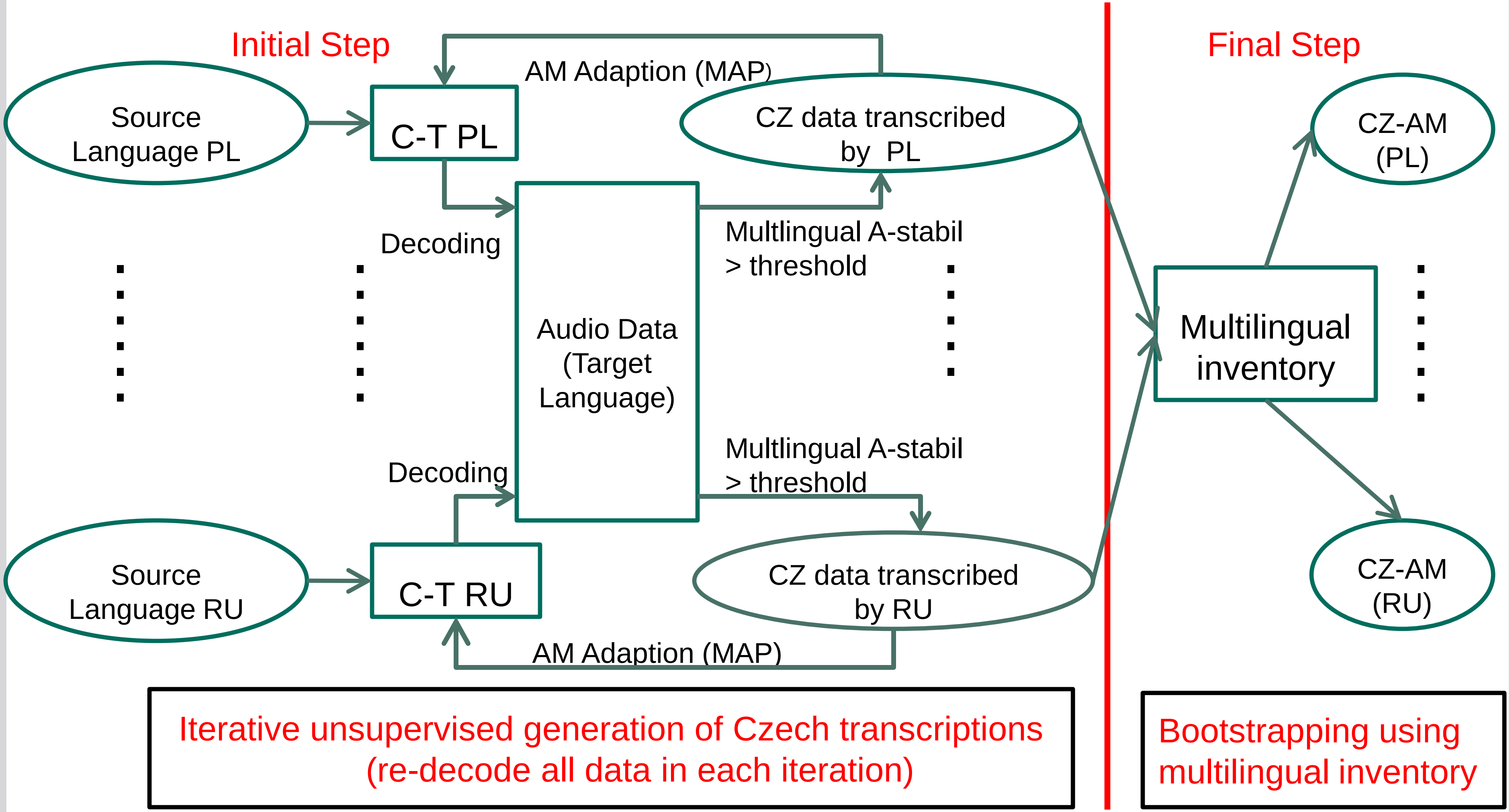
N.T. Vu, F. Kraus, T. Schultz. Multilingual A-stabil: A new confidence score for multilingual unsupervised training. SLT 2010.

Multilingual A-Stabil – Performance



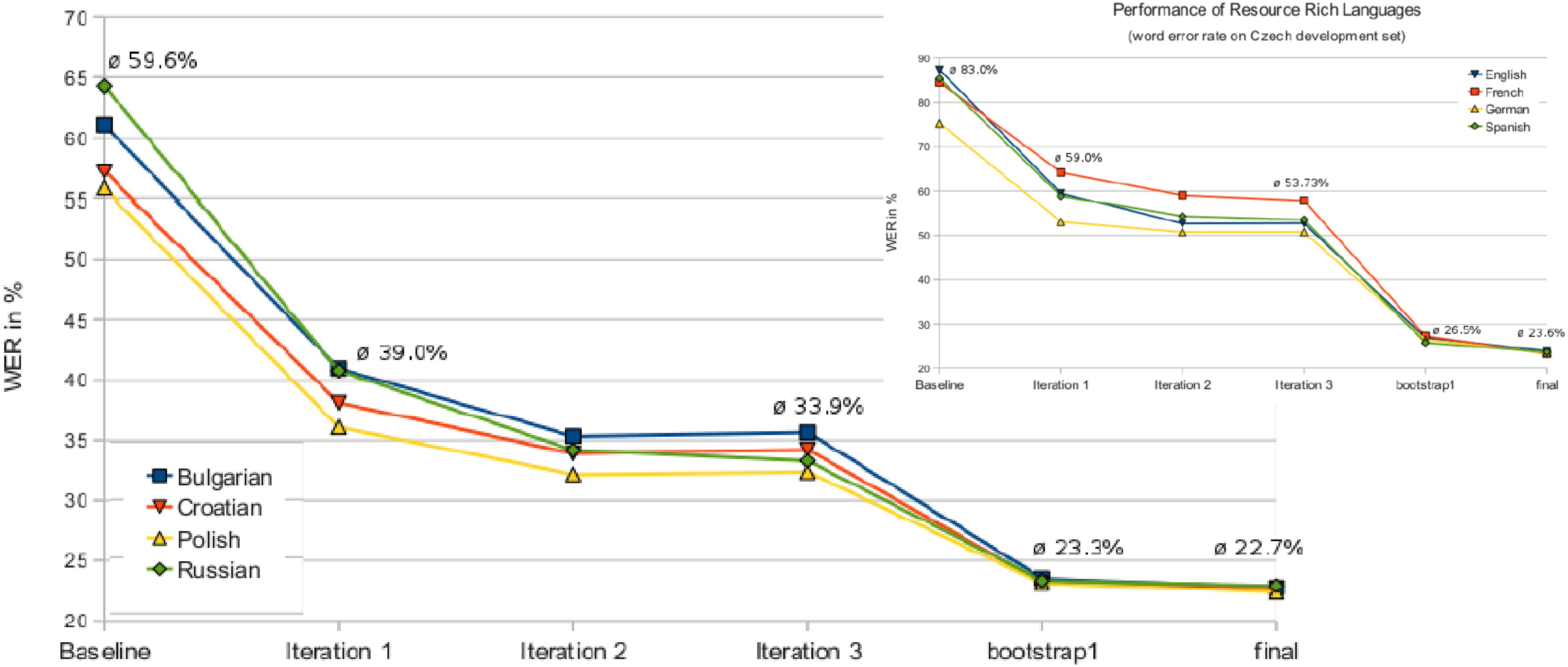
- More languages agree → higher quality (Word Error Rate)
- Multilingual effect: if at least 2 languages agree, chance of correctness is sufficiently high - threshold ≈ 1/N (for N languages)

Step 3: Multilingual Unsupervised Training



N.T. Vu, F. Kraus, T. Schultz.
Rapid building of an ASR system for Under-Resourced Languages based on multilingual unsupervised training. Interspeech 2011.

Results MUT



Extracted 80% / 73% of training data with14.5 / 14.6% WER

Same language family [WER]: Best **22.4**% RU; Min, Max [22.4, 22.9]

Resource-rich languages: Best **23.3**% FR; Min, Max [23.3, 23.9]

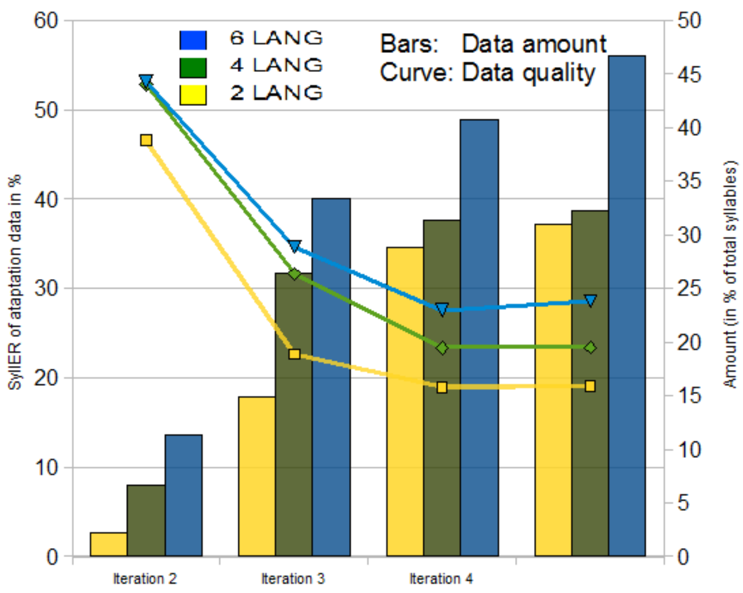
Czech baseline (supervised): **21.8**% WER

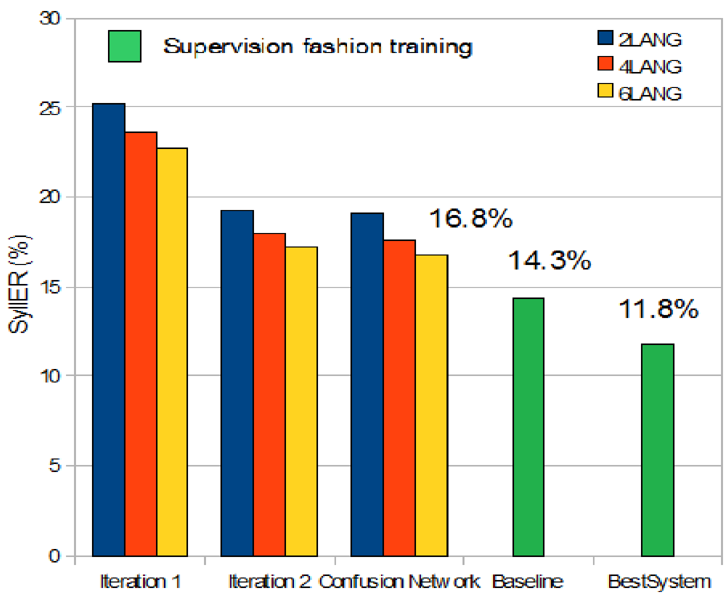
(23h, PPL 1880, 276k vocab, 3.7% OOV, 2000 quintphones)

Impact of Amount of Source Languages

Cognitive Systems Lab

- Target Language: Vietnamese
- Source: English, French, German, Spanish, Bulgarian, Polish
- Finding: More languages help to improve (more data, better quality)
  - Performance within range of VT baseline (16.8% vs. 14.3%)
  - But: Significant gap to language optimized system (11.8%) (Tone modeling, pitch feature, multi-syllables, large text corpus)





N.T. Vu, T. Schultz. Vietnamese Large Vocabulary Continuous Speech Recognition, ASRU 2009.

Karlsruhe Institute of Technology

Proposed Solutions

Lack of data resources for speech processing

- No Transcripts
  - MUT: Multilingual Unsupervised Transcription System
- No Pronunciation Dictionaries
  - G2P, Wiktionary, Keynounce

Lack of a writing system

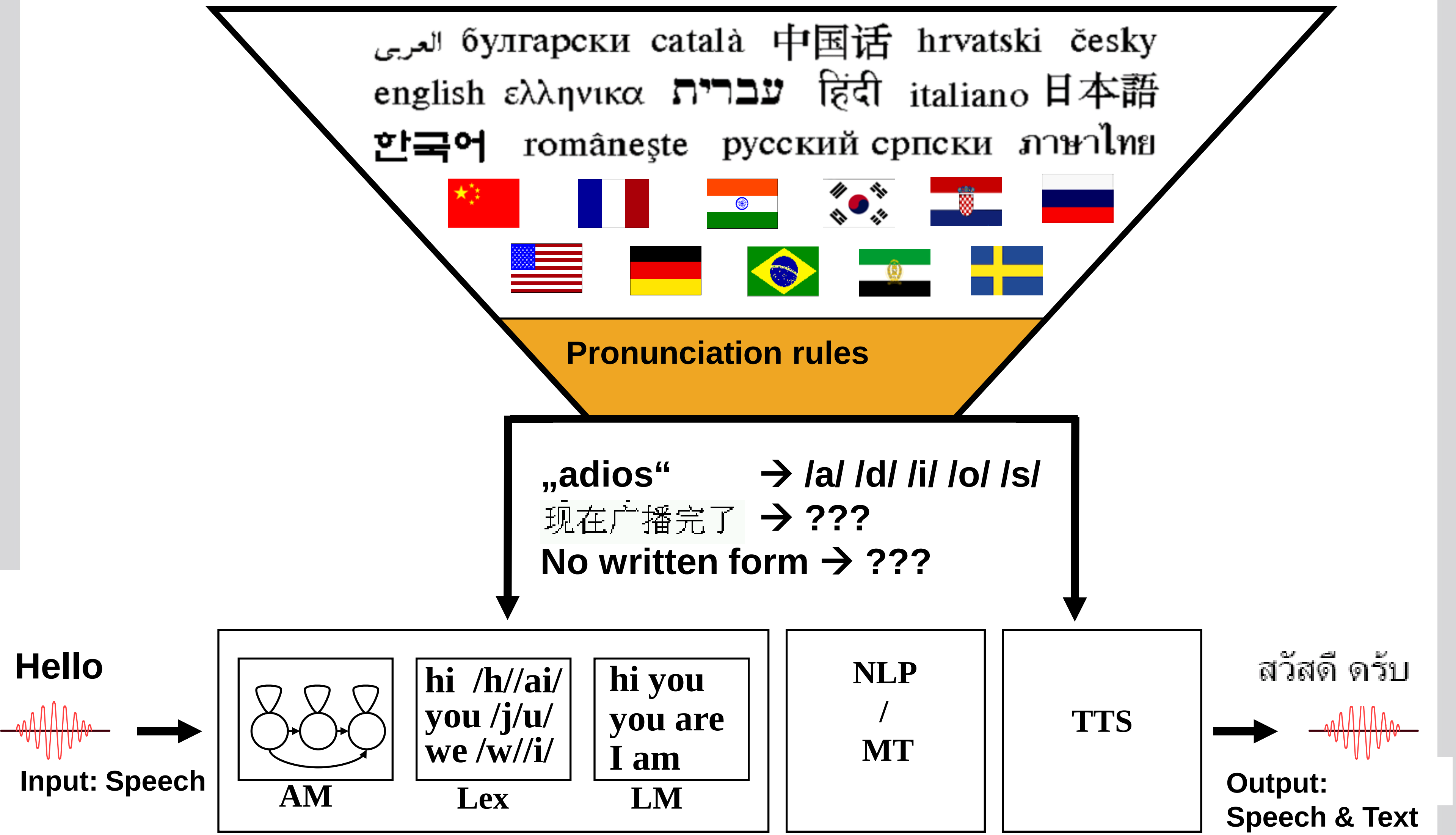
- Cross-lingual Word-2-Phoneme alignments

Lack of linguistic expertise

- Web-based Tools RLAT and SPICE

General Approach: Leverage off existing knowledge and data resources from many languages

Rapid Portability: Pronunciation Dictionary



Writing Systems of Languages

<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, (Simon Ager, http://www.omniglot.com)

<u>Writing systems:</u>

Logographic: based on semantic units, grapheme represents meaning

Phonographic: based on sound units, grapheme represents sound

- *Segmental:* grapheme roughly corresponds to phonemes (*Abjads* =consonantal segmental phonographic),
- *Syllabic:* grapheme represents entire syllable, (*Abugidas* = mix of segmental and syllabic systems)
- *Featural:* smaller than phone, articulatory features

Wikipedia: August 2007

Segmental: Latin, Cyrillic, Latin&Cyrillic, Greek, Georgian or Armenian

Abjads: Arabic, Arabic&Latin, Hebrew&Arabic

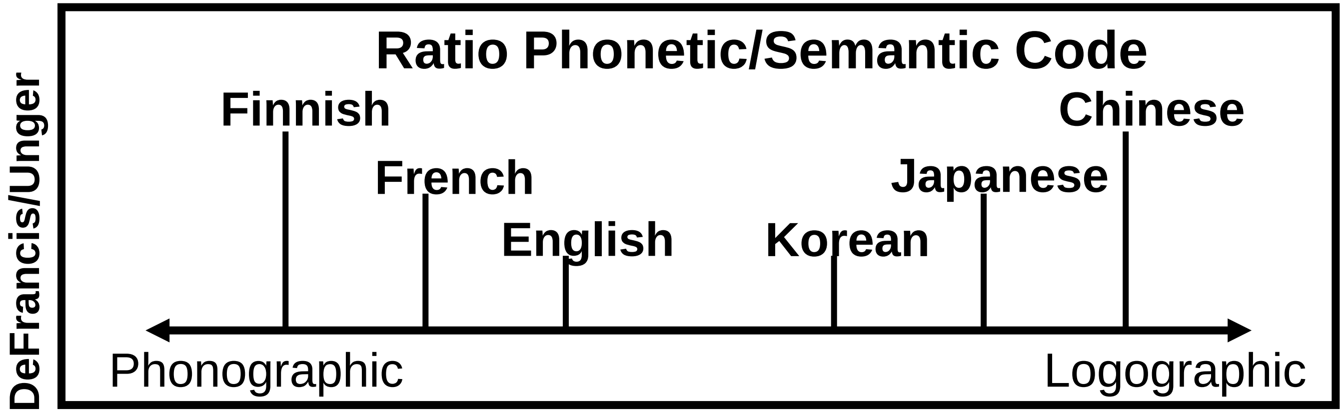
Abugidas: North Indic, South Indic, Ethiopic, Thaana, Canadian Syllabic ,

Logographic+syllabic: Pure logographic, Mixed logographic&syllabaries, Featural syllabary+lmtd logographic Featural-alphabetic syllabary

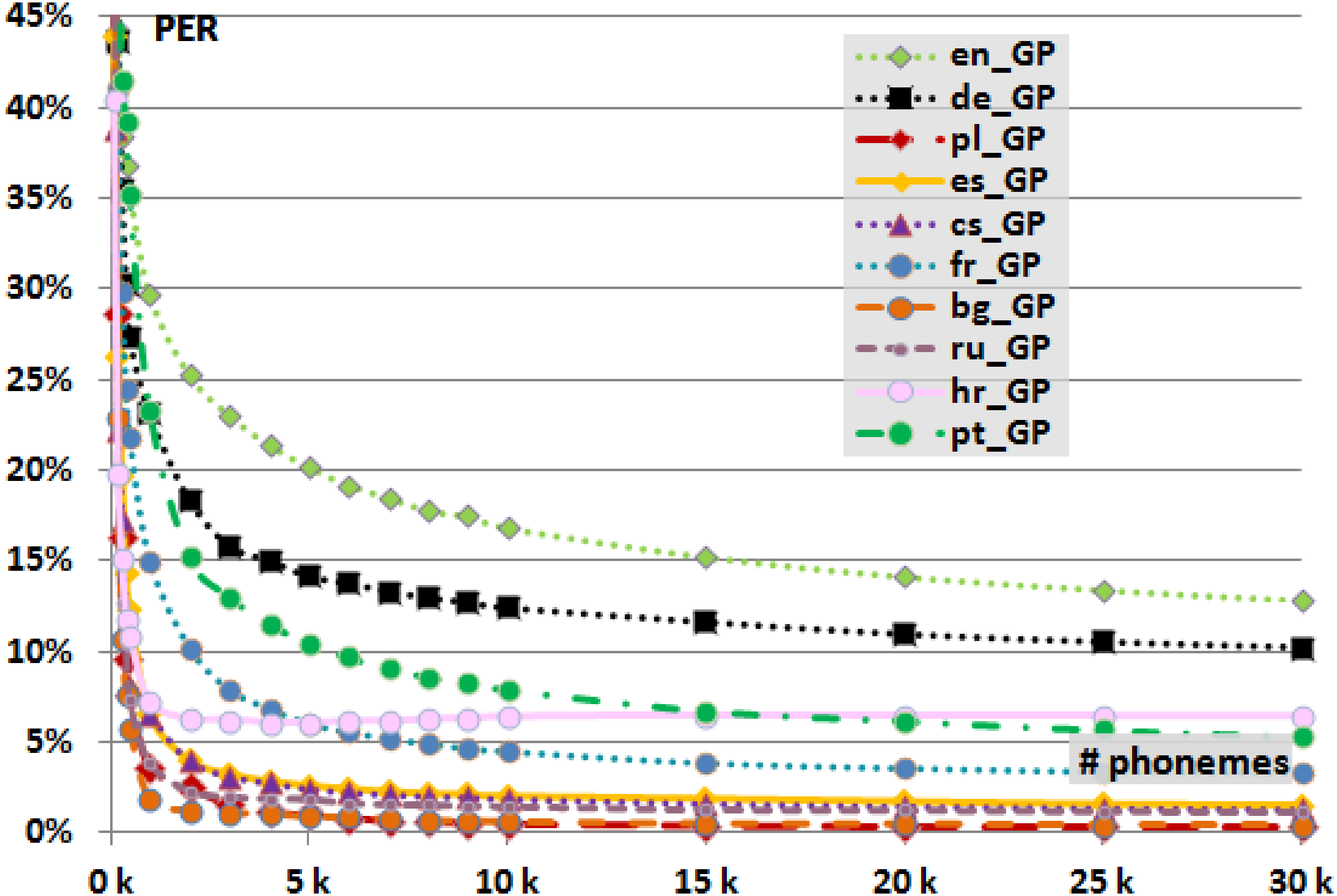
Impact: Grapheme-to-Phoneme Relation

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

- Logographic: NO relationship at all
  - Chinese (>10.000 *hanzi*), Japanese (7000 *kanji*), Korean (some)
- 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 / Phonographic: featural
  - e.g. Thai, Devanagari: C-V flips / Korean (~5600 *gulja*)

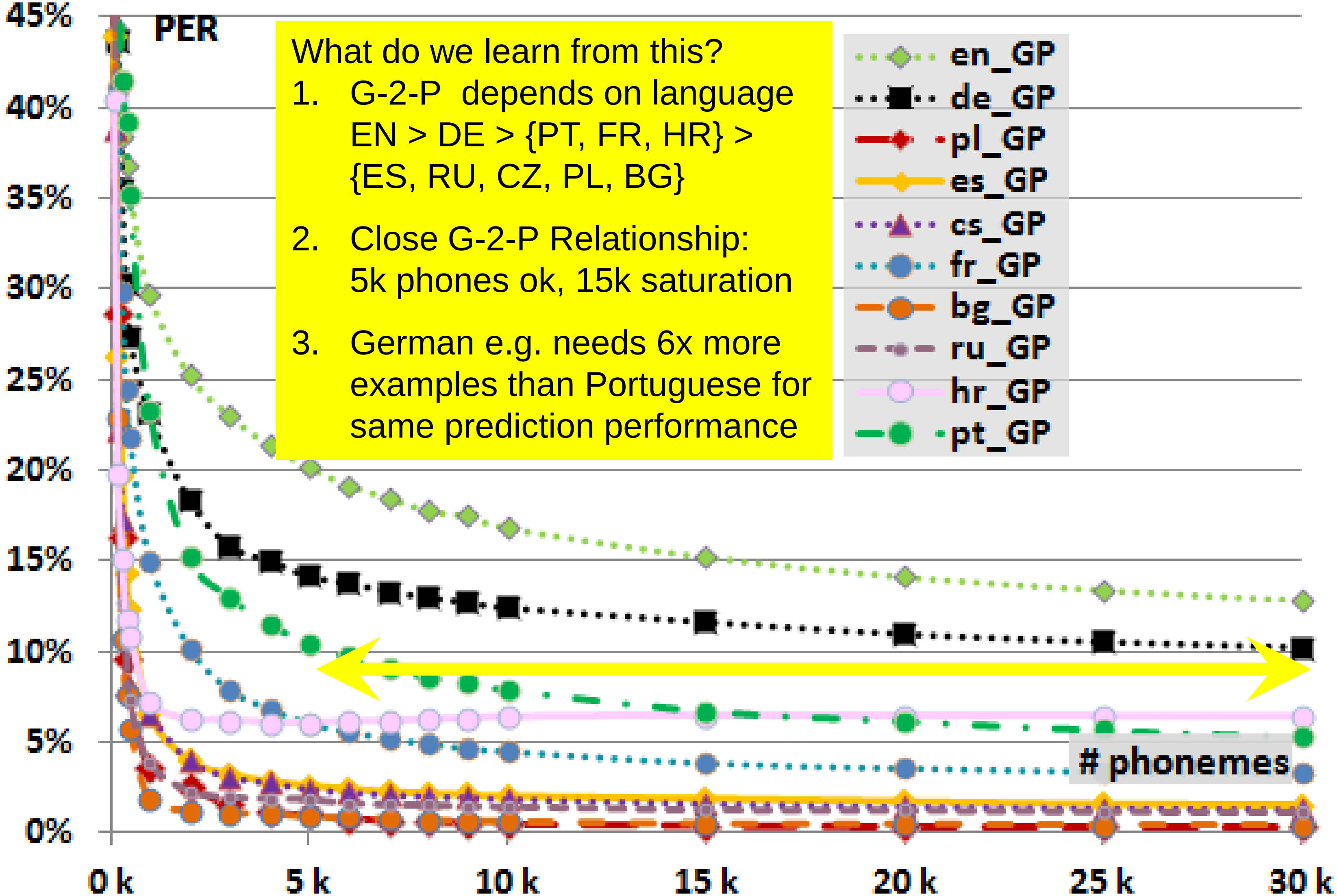


G-2-P: Accuracy over Data (10 languages)



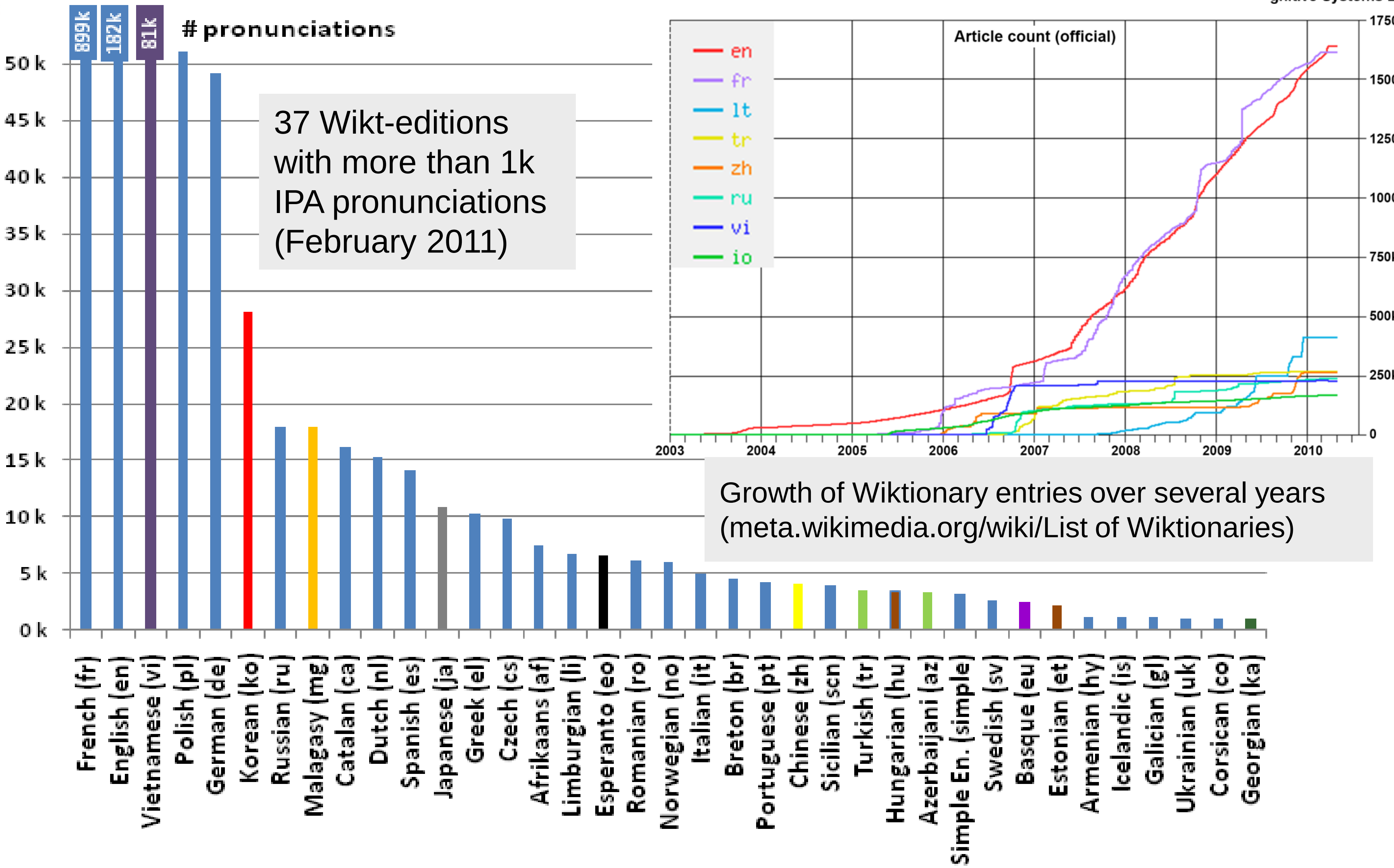
GlobalPhone Dictionaries, G-2-P generation with Sequitur (Bisani & Ney, 2008)

G-2-P: Accuracy over Data (10 languages)



GlobalPhone Dictionaries, G-2-P generation with Sequitur (Bisani & Ney, 2008)

Wiktionary



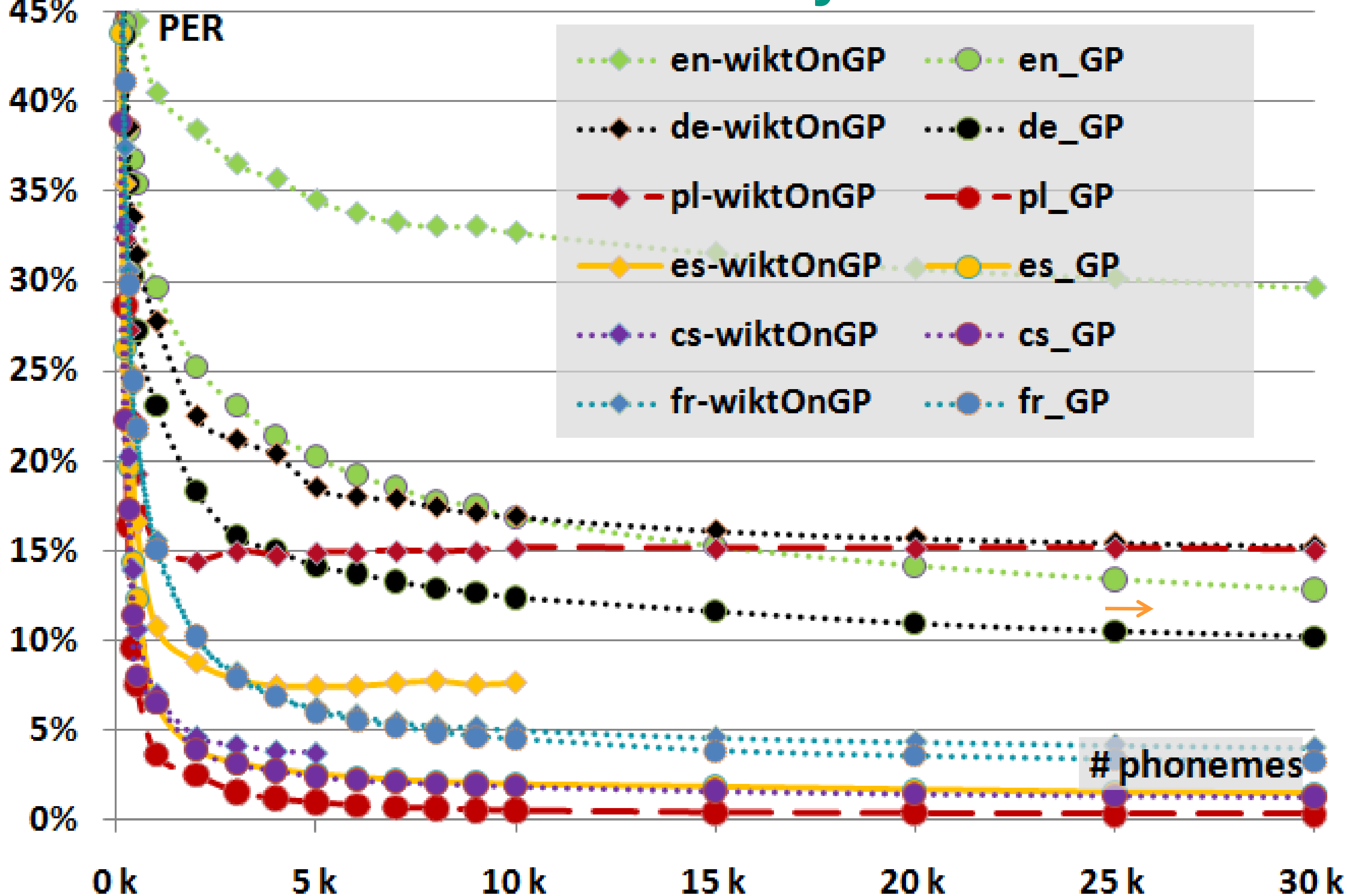
37 Wikt-editions with more than 1k IPA pronunciations (February 2011)

Growth of Wiktionary entries over several years (meta.wikimedia.org/wiki/List of Wiktionaries)

T. Schlippe, S. Ochs, T. Schultz: Web-based tools and methods for rapid pronunciation dictionary creation, Speech Communication, vol 56, pp. 101–118, January 2014.

Web-Interface for Pronunciation Retrieval

G2P Models from Wiktionary vs. GP



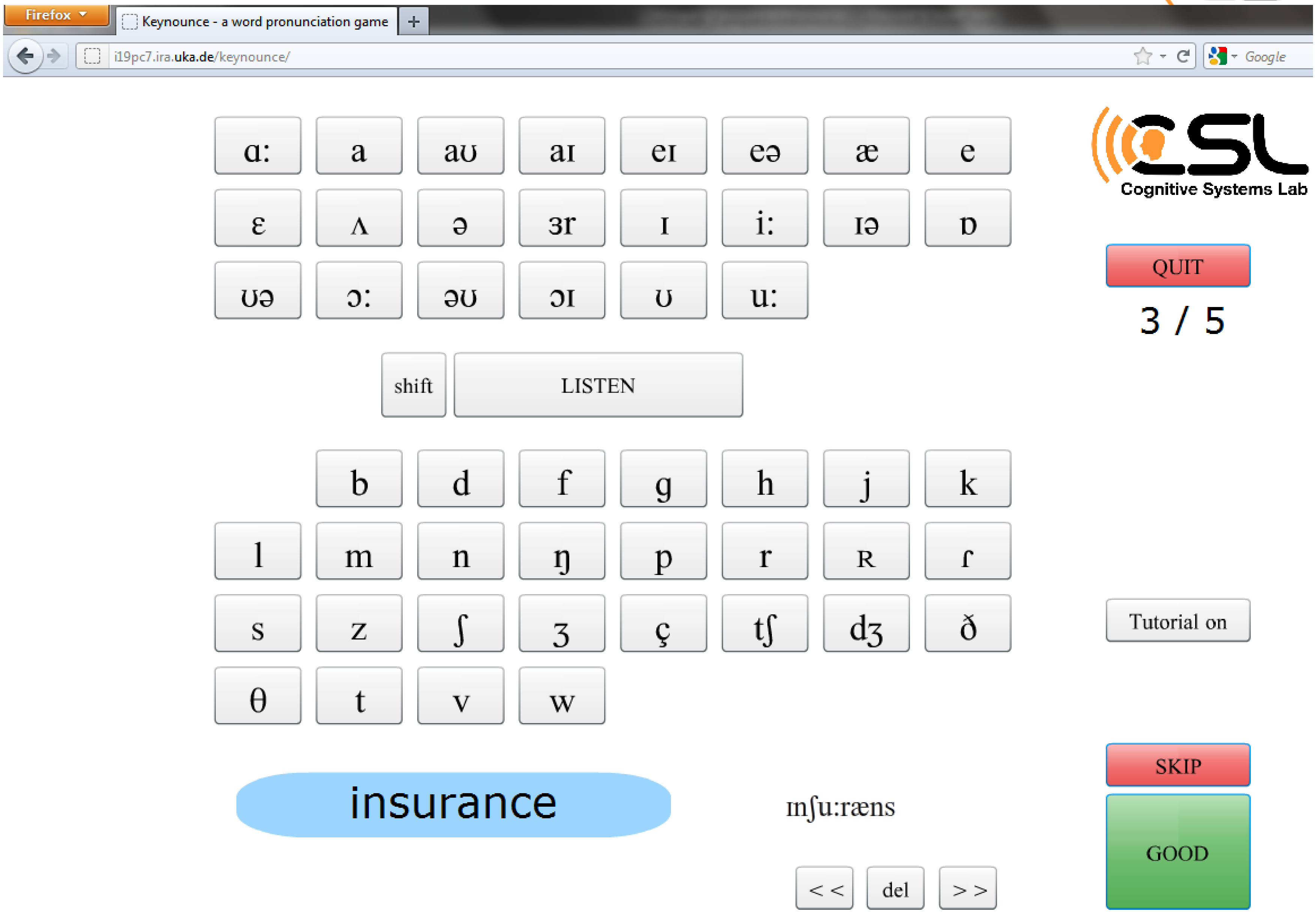
Crosslingual Dictionary Generation

- 0-data?: Apply G-2-P models of (related) languages
- Target: Ukrainian, Source: Russian, Bulgarian, German, English
  1. Crosslingual **G2G**: Map Ukrainian grapheme $\rightarrow$ Source grapheme
  2. Monolingual **G2P**: Apply Source Grapheme $\rightarrow$ Source Phone model
  3. Crosslingual **P2P**: Map resulting Source Phones $\rightarrow$ Ukrainian Phones
  4. Post-processing to fix shortcomings (**Post**-rules)

	# G2G	# P2P	PER [%]	WER [%]	# Post	PER [%]	WER [%]
RU	43	56	12.4	22.8	57	1.7	**21.63**
BG	40	79	10.3	23.7	65	2.8	22.1
DE	68	66	32.7	27.1	39	28.6	26.4
EN	68	63	46.8	34.9	21	36.6	34.0
Ukrainian Grapheme-based ASR							23.8
Ukrainian ASR with Hand-crafted dictionary (882 rules)							22.4
+ data-driven Semi-Palatalized Phone Modeling							**21.65**

T. Schlippe, M. Volovyk, K. Yurchenko, T. Schultz. Rapid Bootstrapping of a Ukrainian LVCSR System, ICASSP 2013

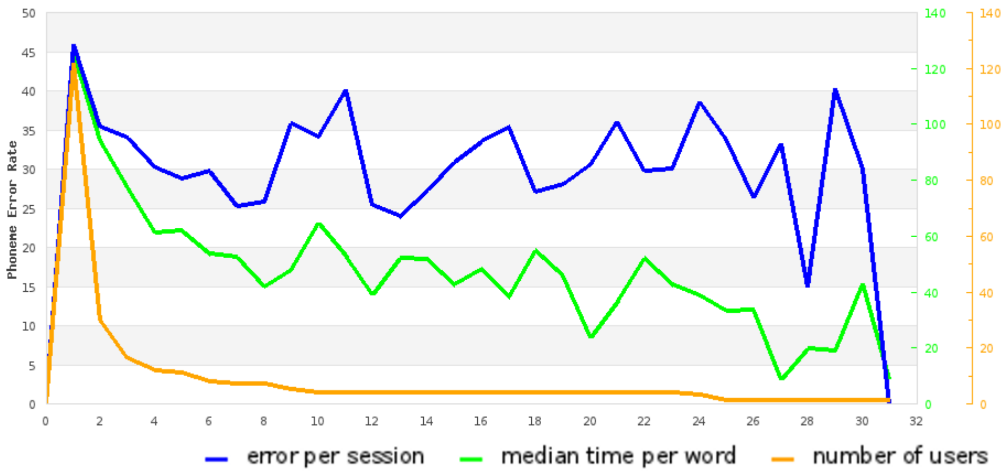
Keynounce – Pronunciation Generation via Crowdsourcing



Issues with Crowdsourcing

Cognitive Systems Lab

- Keynounce using mTurk (12 days):
  - Average time spent: 53 seconds
  - 387 approved / 531 rejected assignments
  - 1902 pronunciations, 55% rejected (1062)
  - Excessive SPAM accounts/bots to test HITs for easy money
  - Fast but sloppy, Incentives to provide “good” answers?
- Use Friends/Volunteers, Improved Interface:
  - Welcome page, Tutorial
  - Quality Feedback
  - Show current ranking
  - Get familiar with task
    1st word: 6 minutes
    2nd word: 2 minutes
    last words: 1:30min
  - Slower but higher quality



Daniel Lemke. Keynounce - A Game for Pronunciation Generation through Crowdsourcing, Student Paper, CSL KIT, 2013

Karlsruhe Institute of Technology

Proposed Solutions

Lack of data resources for speech processing

- No Transcripts
  - MUT: Multilingual Unsupervised Transcription System
- No Pronunciation Dictionaries
  - G2P, Wiktionary, Keynounce

Lack of a writing system

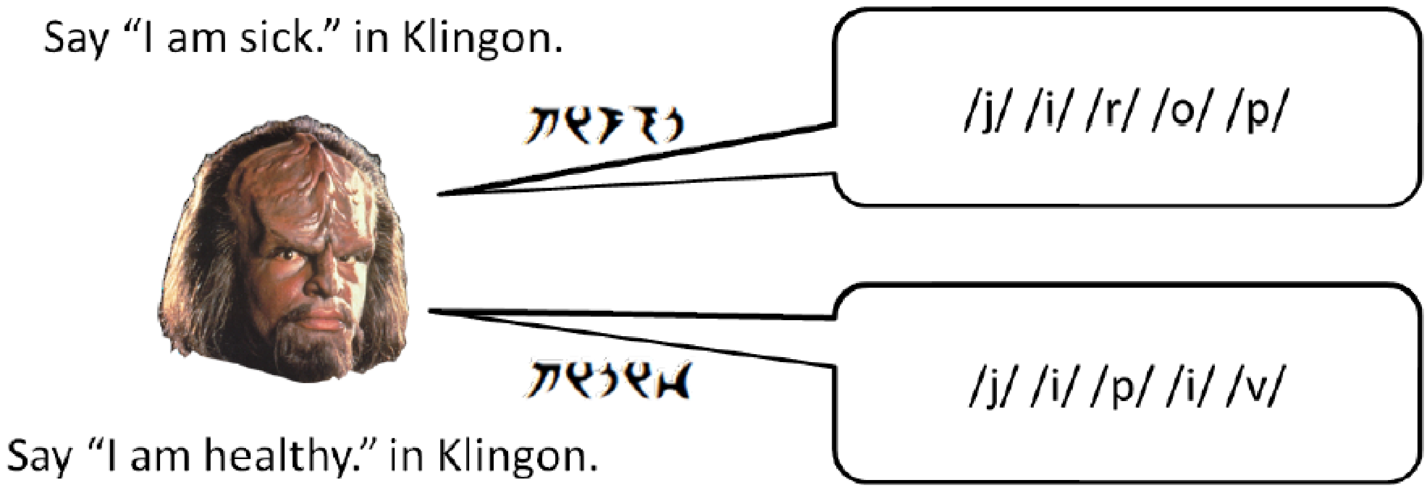
- Cross-lingual Word-2-Phoneme alignments

Lack of linguistic expertise

- Web-based Tools RLAT and SPICE

General Approach: Leverage off existing knowledge and data resources from many languages

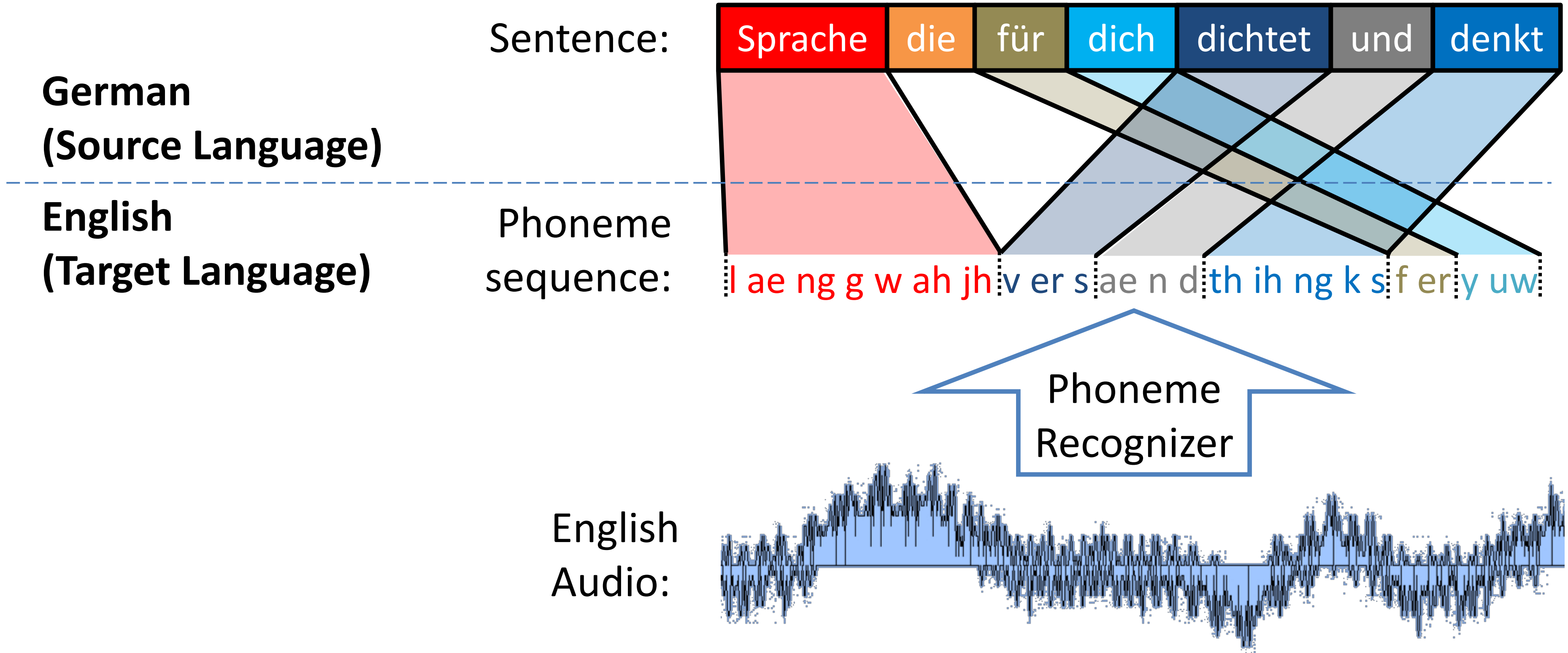
Languages without Written Form



- **/j/ /i/** seems to be a word (meaning **I am**)
- **/r/ /o/ /p/** seems to be a word (meaning **sick**)
- **/p/ /i/ /v/** seems to be a word (meaning **healthy**)

- Goal: ASR for spoken (only) languages // no linguistic knowledge available
- Approach: Exploit the phonetic output of a human simultaneous translator
- Cross-Lingual Word-to-Phoneme Alignment
  - Discover words, vocabulary, and pronunciations

Word-to-Phoneme Alignments



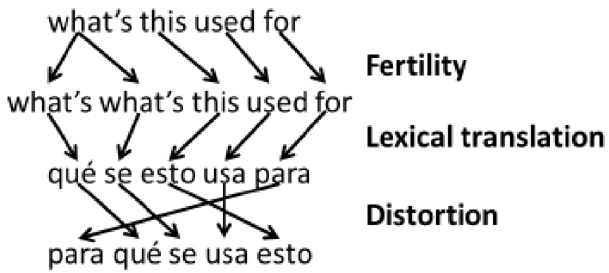
(Besacier et. al., 2006) – monolingual unsupervised segmentation of phone sequences into words
(Stüker and Waibel, 2008) – cross-lingual word-to-word alignment using Giza++
(Stüker and Besacier, 2009) – combine monolingual and Giza++ approach
(Stahlberg et. al., 2012) – use cross-lingual word-to-phoneme alignment approach

Step 1: Model 3P (Giza++ → Pisa)

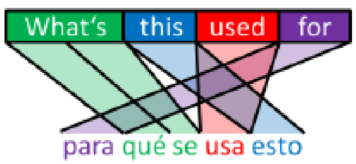
IBM Model 3

Problem: Generative story does not fit word-to-phoneme alignment

Generative Story



Alignment

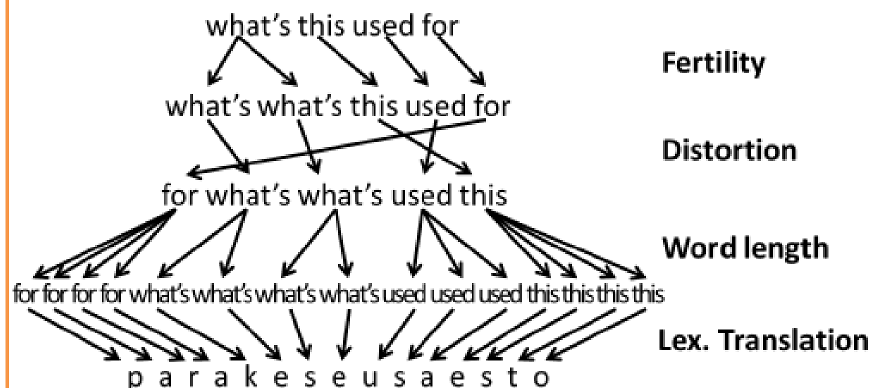


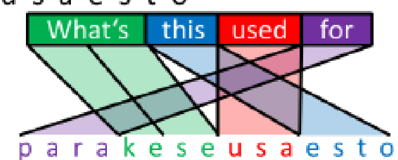
English word position (i)	4	1	1	3	2

Model 3P

Extend IBM Model 3
- apply word length probability, phoneme position in target word
- insert WB where phone neighbors align to different source words

http://code.google.com/p/pisa/ (Stahlberg et. al., 2012)

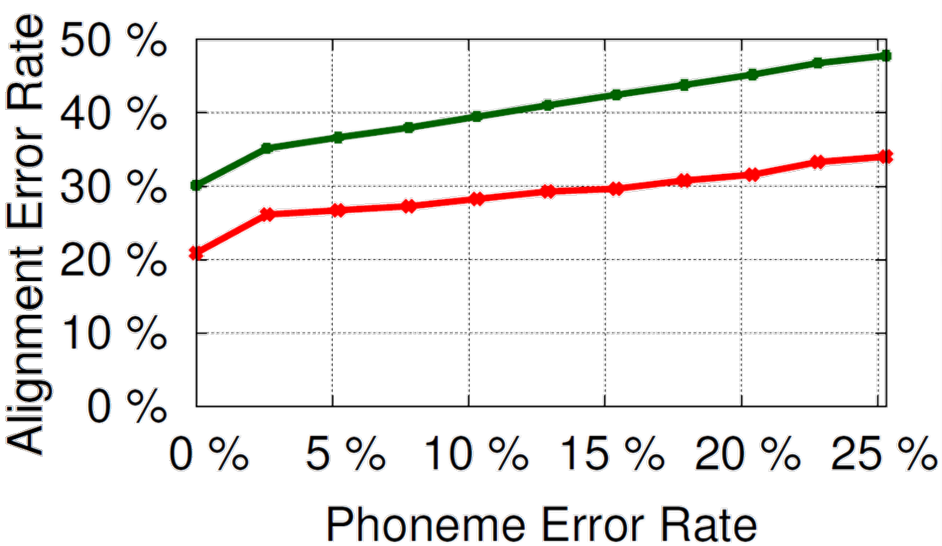




English word position (i)	4	4	4	4	1	1	1	1	3	3	3	2	2	2	2
Target word position (π_{ik})	1	–	–	–	2	–	3	–	4	–	–	5	–	–	–
Target word length (ψ_{ik})	4	–	–	–	2	–	2	–	3	–	–	4	–	–	–
Phoneme position in target word (j)	1	2	3	4	1	2	1	2	1	2	3	1	2	3	4

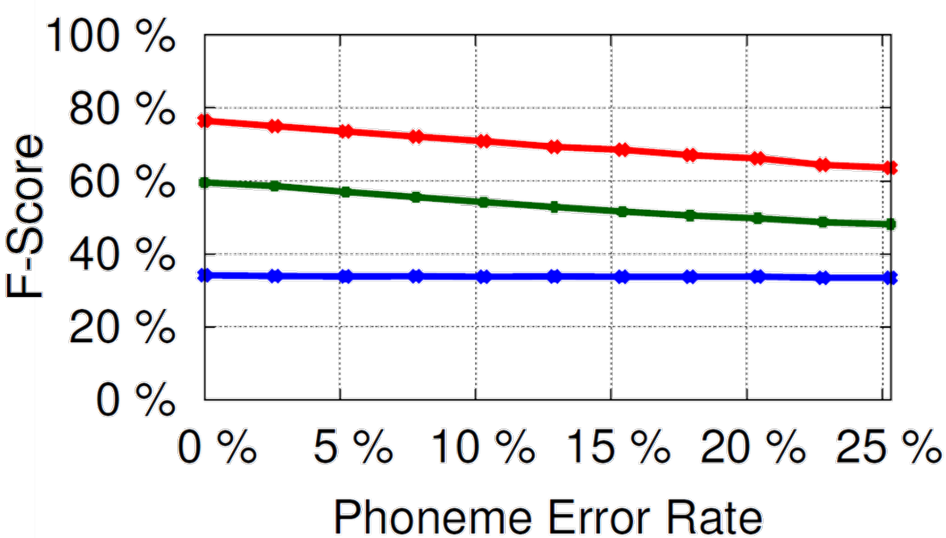
Results on BTEC English-Spanish

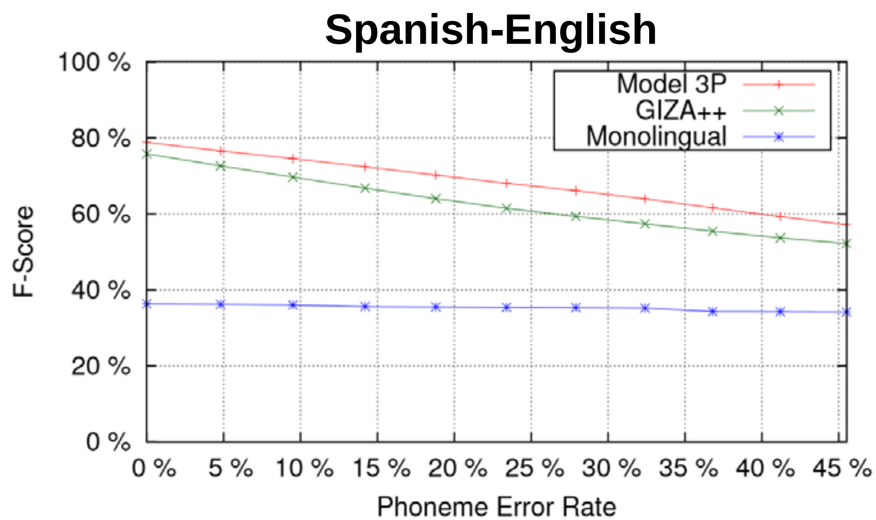
(Stahlberg et. al., 2012)



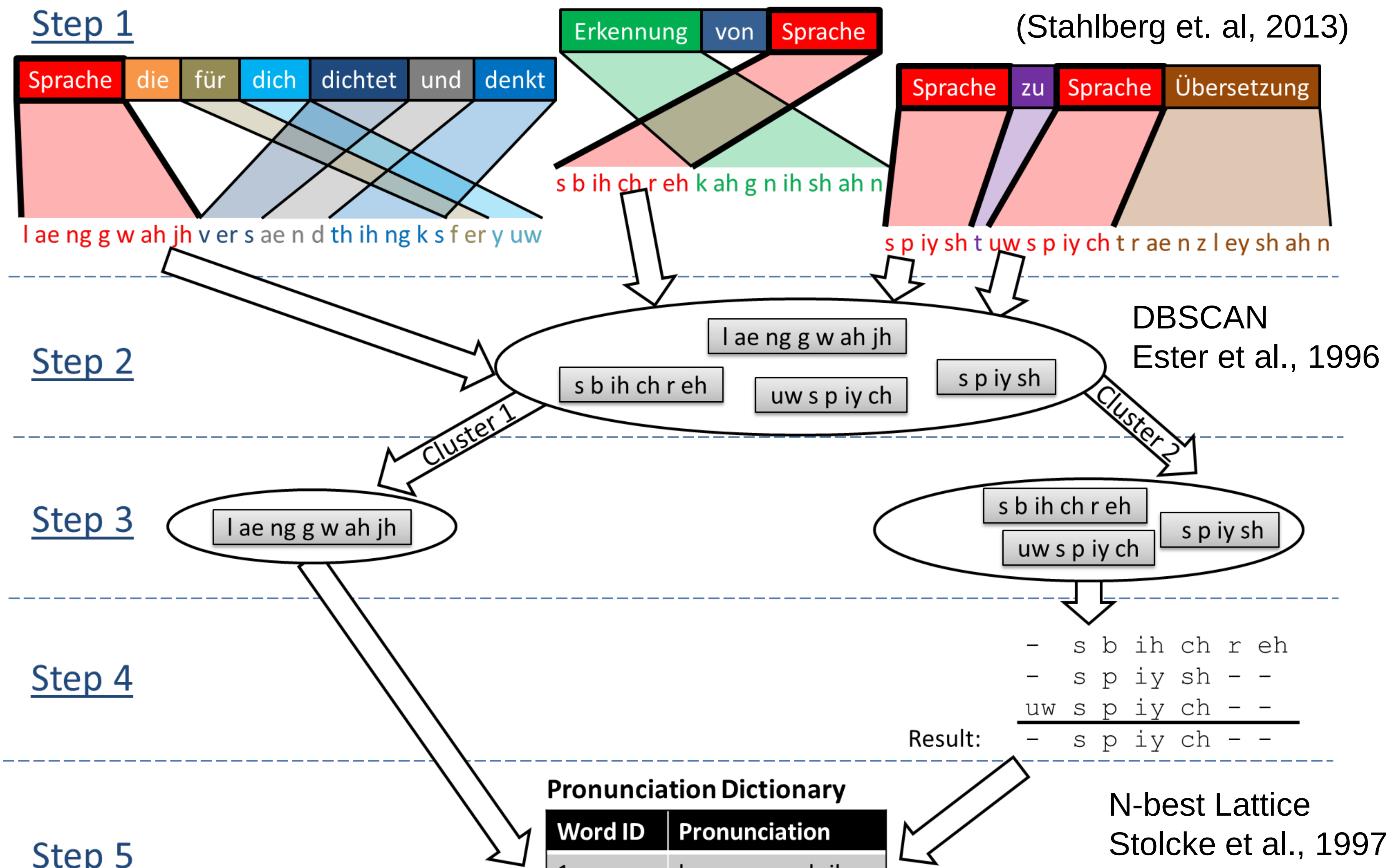
Adaptor Grammars (Monolingual)
GIZA++ word-to-phoneme alignments
Model 3P

BTEC 123k sentence pairs
Spanish PER=25.3%; English 45.5%
Ref: GIZA++ on word level





WordIDs and Pronunciations



Pronunciation Dictionary

Word ID	Pronunciation
1	l ae ng g w ah jh
2	s p iy ch

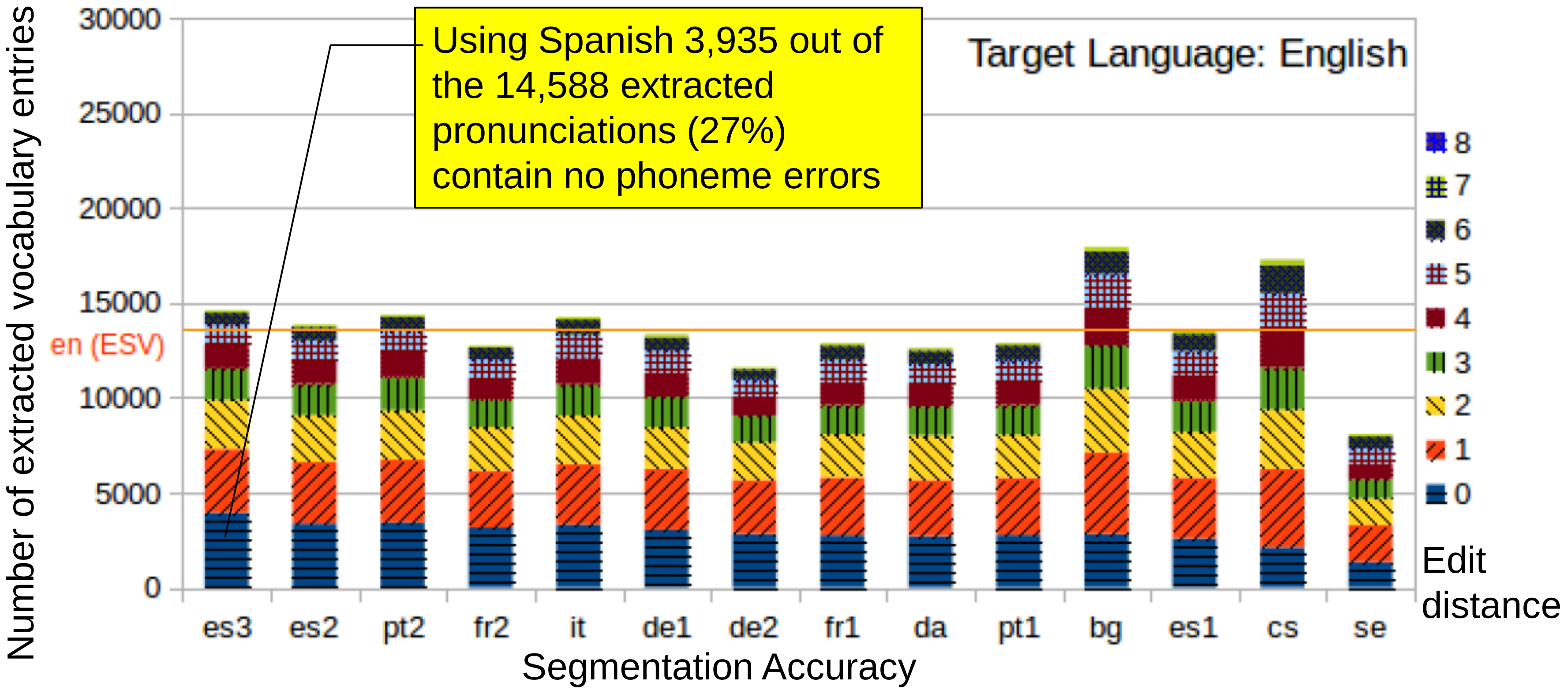
Pronunciation Extraction on Bible Data

Cognitive Systems Lab

ID	Language	Full Bible Version Name	# running words	Vocab. Size
bg	Bulgarian	Bulgarian Bible	643k	38k
cs	Czech	Bible 21	547k	48k
da	Danish	Dette er Biblen på dansk	653k	24k
de1	German	Schlachter 2000	729k	26k
de2	German	Luther Bibel	698k	21k
en	*English*	*English Standard Version*	*758k*	*14k*
es1	Spanish	Nueva Versión Internacional	704k	28k
es2	Spanish	Reina-Valera 1960	706k	26k
es3	*Spanish*	*La Biblia de las Américas*	*723k*	*26k*
fr1	French	Segond 21	756k	26k
fr2	French	Louis Segond	735k	23k
it	Italian	Nuova Riveduta 2006	714k	28k
pt1	Portuguese	Nova Versão Internacional	683k	25k
pt2	*Portuguese*	*João Ferreira de Almeida Atualizada*	*702k*	*26k*
se	Swedish	Levande Bibeln	595k	21k

Extracted from http://www.biblegateway.com (accessed on Nov 2013); verse aligned, 30k

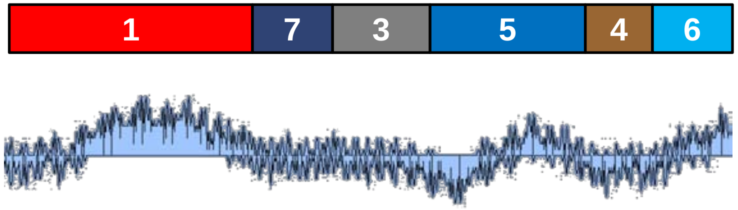
Pronunciation Extraction



- Distribution of the abs. phoneme errors in the extracted pronunciations
  - OOV: Assign each wordID to the written word with most similar pronunciation
  - PER: Calculate phoneme error based on this assignment
- ESV: English Standard Version (Crossway, 2001); Zipf distrib; 30% freq=1
- Target phoneme sequence: canonical pronunciation PER = 0%, no WB

Next Step: Putting the Pieces together

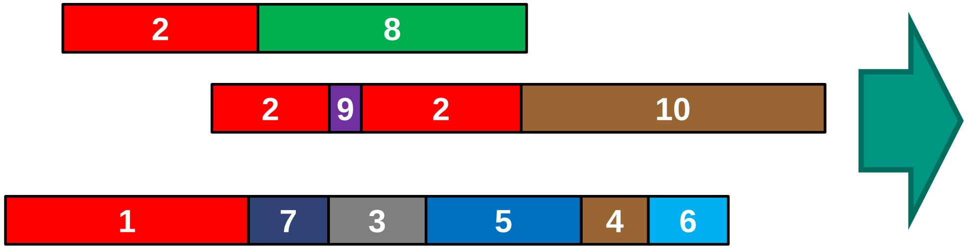
- We get
  - Transcribed audio data (in terms of IDs)



  - Pronunciation dictionary

Word Label	Pronunciation
1	l ae ng w ah jh
2	s p iy ch
3	ae n d
4	f er
5	th ih ng k s
6	y uw
7	v er s
8	k ah g n ih sh ah n
9	t uw
10	t r ae n z l ey sh ah n

  - Language model



$P(W)$

Train ASR System (future work)

Proposed Solutions

Lack of data resources for speech processing

- No Transcripts
  - MUT: Multilingual Unsupervised Transcription System
- No Pronunciation Dictionaries
  - G2P, Wiktionary, Keynounce

Lack of a writing system

- Cross-alignment word alignments

Lack of linguistic expertise

- Web-based Tools SPICE and RLAT

General Approach: Leverage off existing knowledge and data resources from many languages

Rapid Language Adaptation Tools

Speech Processing: Interactive Creation & Evaluation toolkit

- National Science Foundation, 2004-2008 (Schultz & Black)
- Bridge the gap between technology experts → language experts
  - Components for ASR, MT, TTS
- Develop web-based intelligent systems
  - Interactive Learning with user in the loop
  - Rapid Adaptation from universal models

Rapid Language Adaptation Toolkit (KIT)

- Massive Crawling (text, rss-feeds, twitter), text post processing
- Automatic Pronunciation Generation (wiktionary, crowd-sourcing)
- Two alternative Interfaces for data collection: Web-based and Telephone
- RLAT *webpage* *http://csl.ira.uka.de/rlat-dev*

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

Firefox | RLAT

i19pc5.ira.uka.de/rlat-dev/project.php

Suchen | Backup

--> RLAT project management

Build Your System

- Text and prompt selection (help)
  - Text management
  - SMT-based text normalization (help)
- Audio collection (help)
- Phoneme selection (help)
- Grapheme-to-phoneme rules (help)
- Lexicon pronunciation creation (help)
  - Web-derived pronunciations
- Build acoustic model (help)
- Build language model (help)
  - Language model management
- Test ASR system
- Create speech synthesis voice

- Collect appropriate text and audio data
- Define phoneme set, prompt set
- Define and Refine pronunciation dictionary
- Produce:
  - Vocabulary / Word lists (ASR, TTS, SMT)
  - Pronunciation model (ASR, TTS)
  - Acoustic model (ASR, TTS)
  - Language model (ASR, SMT)
  - Synthetic voices (TTS)
- Maintain user and projects, data, models

Recent Progress on RLAT

- Hands-on courses at CMU and KIT since 2007: Students build ASR and TTS in their language (Bulgarian, German, Hausa, Hindi, Konkani, Suaheli, Tamil, Telugu, Turkish, Ukrainian, Vietnamese, …)
- Collaboration / Crowd Sourcing
  - OK: Multiple people working on the same language / similar projects
  - Leverage archived expertise, Multiple views within and across projects
- Error-blaming
  - OK: Automatic Generation of Recommendations to improve systems
  - End-to-end system Evaluation versus Component Evaluation
- Address Language Peculiarities
  - OK: Enable users to customize to languages (e.g. normalization)
- Continuous Server Support
  - Improve Interface based on user feedback and lessons learned
  - Latest Version @ http://csl.ira.uka.de/rlat-dev

Conclusions

- Techniques to perform on low resources
  - Share data/models across system components
  - Reuse language independent aspects of data/models
- Lower the overall costs for system development
  - Automate data collection process, Leverage off Crowd Sourcing
  - Reduce the data needs without sacrificing (too much) performance
- Field Work and Community Outreach
  - Get tools to the people, i.e. flexible, portable, simple
  - Engage and actively involve native speakers
  - Identify language specific aspects
- Bridge the gap between technology and language experts
  - Technology experts do not speak all languages in question
  - Native users are not in control of the technology

Thank You

Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

Thanks to Ngoc Thang Vu, Tim Schlippe, Felix Stahlberg, Stephan Vogel, Sebastian Ochs, Alan Black, Franziska Kraus, Mykola Volovyk, Kateryna Yurchenko, Sameer Badaskar, Daniel Lemke, and Heike Adel

Cognitive Systems Lab