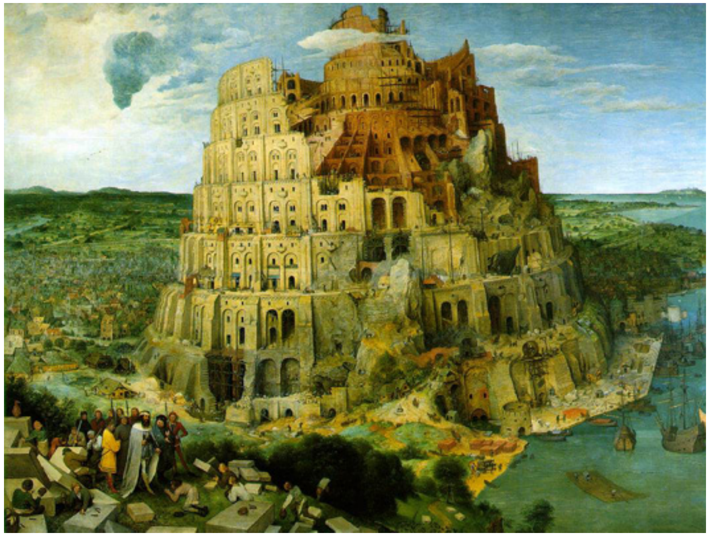
Karlsruhe Institute of Technology

Multilingual Automatic Speech Recognition for Code-switching Speech



University of the State of Baden-Württemberg
and National Research Center of the
Helmholtz Association

Tanja Schultz

Cognitive Systems Lab, Institute for Anthropomatics and Robotics, KIT

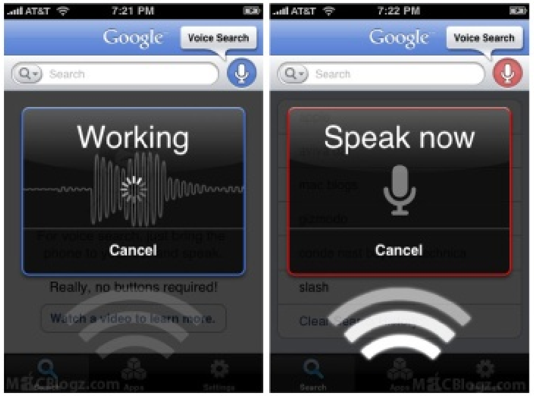
Cognitive Systems Lab

9th CSLP 2014

The 9th International Symposium on Chinese Spoken Language Processing
12-14 September 2014, Singapore

Multilingualism: An Engineer's View

- Multilingual Individuals and Communities
  - Multilingual speakers outnumber monolingual ones (wikipedia)
- Results in frequent *Code-Switching,* which happens …
  - Between and within utterances, between phrases, words
  - Occurrence depends on several factors (see below)
- What is the impact on voice-driven applications (Apple's Siri, Google voice search, …)?
- **Which language to use for the application?**
  - Push for only one?
  - Provide many?
  - All-in-one or several ones?
  - Identify spoken language.
  - Detect Code-Switching.



Code-Switch Conversational Speech

Definition: Code Switching (CS) is the phenomenon of changing languages within an utterance or discourse.

Code-Switch Points: CS may appear at sentence boundary, at phrase boundary or freely at any point in time

Code-Switching may depend on:

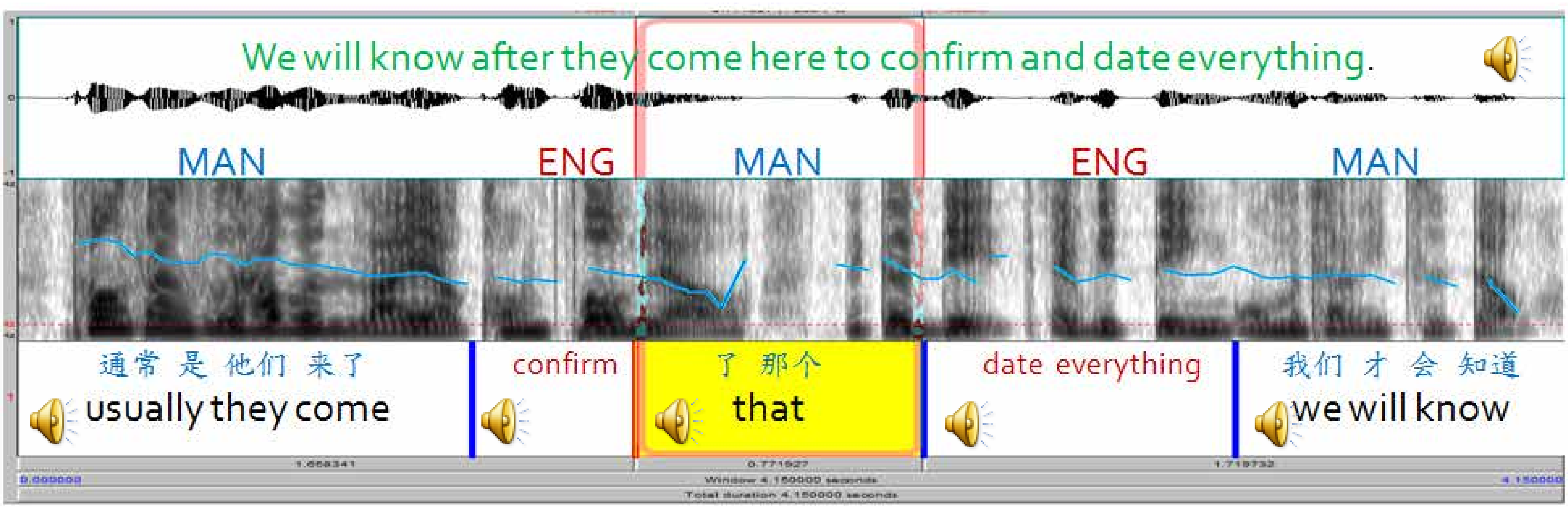
- Speakers' preferences,
- Languages involved,
- Topic / domain,
- Situation, context,
- Bystanders, audience,
- Mood, emotion, …

Code-Switch Conversational Speech

Definition: Code Switching (CS) is the phenomenon of changing languages within an utterance or discourse.

CS may appear at sentence boundary, at phrase boundary or freely at any point in time

Example (from the SEAME corpus*):



* Dau-Cheng Lyu, Tien Ping Tan, Eng-Siong Chng, Haizhou Li:
SEAME: A Mandarin-English Code-Switching Speech Corpus in South-East Asia. INTERSPEECH 2010: 1986-1989

Code-Switching – Related Work

- While Code-Switching (CS) frequently occurs … IMHO it received **too little attention in the spoken language community so far**
- Occurrences: CS positions follow the syntactical rules of the involved languages (Poplack 1978; Bokamba 1989; Muysken 2000),
- Speaker Dependencies: Some CS patterns are shared across speakers (Poplack, 1980), in general CS seems to be speaker dependent (Auer 1999; Vu et al., 2013),
- CS Detection:
  - Bi-phone probabilities: Chan/Ching/Lee/Meng 2004
  - Linguistic features and their combination: Solorio/Liu 2008
  - Textual features: Burgmer/Fung/Schultz 2009
- CS ASR:
  - Chan/Ching/Lee/Cao 2006: foreign words into POS classes to predict CS
  - Lyu/Lyu/Chiang/Hsu 2006: ASR for Chinese Dialects (Mandarin-Taiwanese)
  - C-F Yeh/L-S Lee et al. 2010, 2011, 2012-: Chinese-English lectures, Taiwan
  - Li/Fung 2012- : integrate equivalence constraint into LM for Mandarin-English
  - Davel/Barnard et al. 2010- : English/South African languages (Sepedi)
  - AM: Stemmer/Nöth 2001, White/Baker 2008, Imseng/Bourlard 2011
  - LM: Fügen/Schultz et al. 2003: Multilingual LMs and Grammars, integration

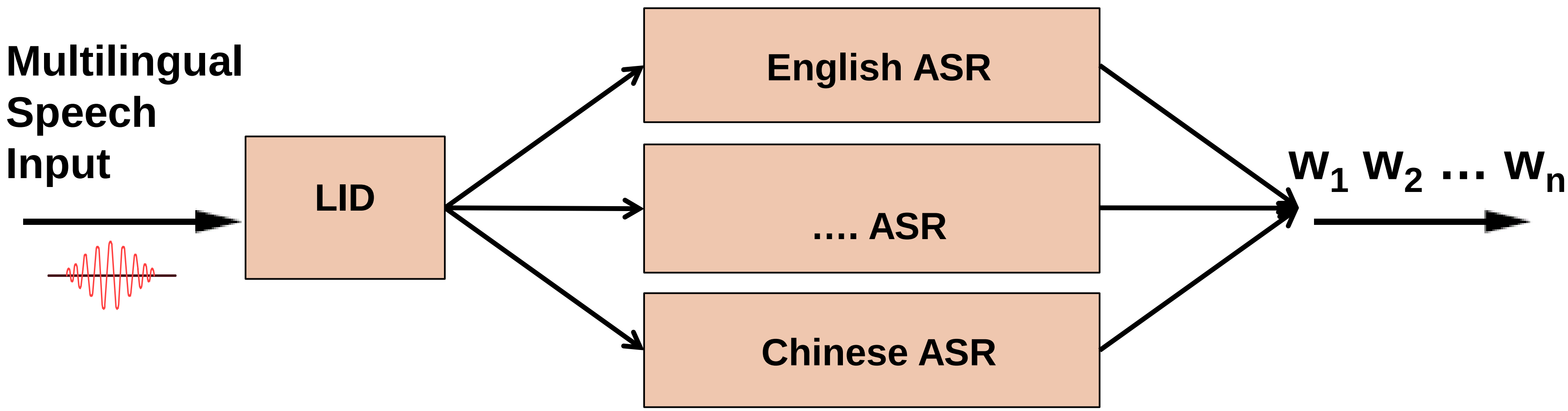
Automatic Speech Recognition for CS

Code-Switching combines **several challenges** in ASR:

1. CS is a spoken phenomenon
   - Recognition of conversational speech is tough by itself
   - No large / if any written text corpora (web)
2. CS is highly speaker dependent (on all levels)
   - No fixed rules when to code-switch
   - May also depend on the language combination – few data
3. Requires truly **multilingual** ASR components
   - Identification/prediction of code-switches
   - Large variety of language combinations
     (among Chinese dialects e.g. Mandarin/Taiwanese; with English e.g. Mandarin/English, Cantonese/English, Malay/English, Indian languages, South African languages, smaller bilingual pockets e.g. Swiss/German, Flemish/French, …)

CS ASR – Approach 1: Monolingual ASR

Language Identification (LID) followed by monolingual ASR

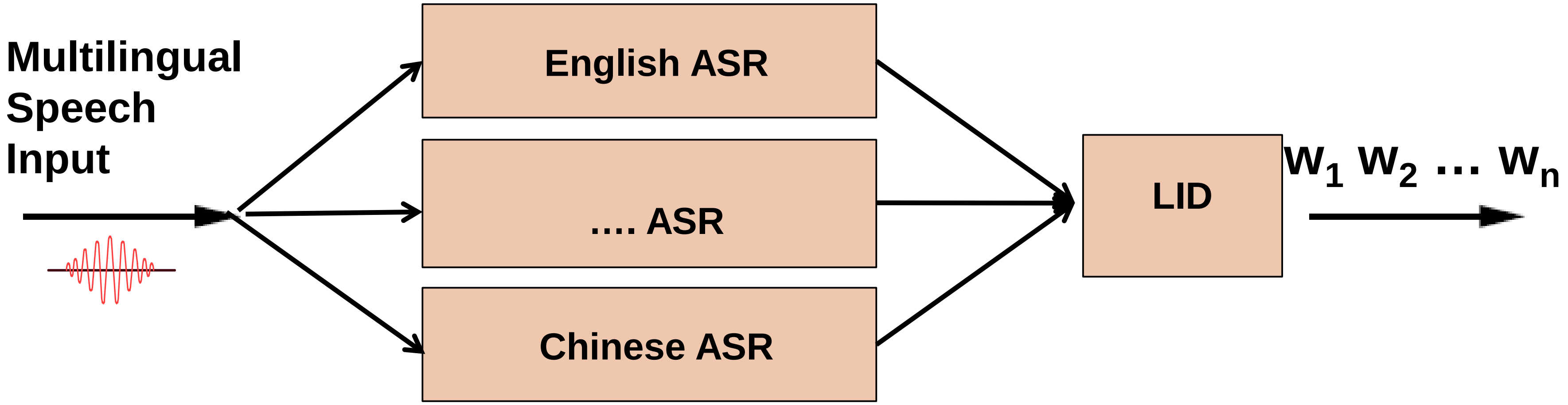


Benefits	Drawbacks
Straight-forward to implement	LID errors are not recoverable
Only monolingual ASR, data required	Semantic context is lost
	No true support of bilingualism

T. Schultz et al., *Multilingual Speech Recognition.* Chapter in: Verbmobil - Foundations of Speech-to-Speech Translation, Wolfgang Wahlster (Hrsg.), Springer Verlag, 2000.

CS ASR – Approach 1a: Monolingual ASR

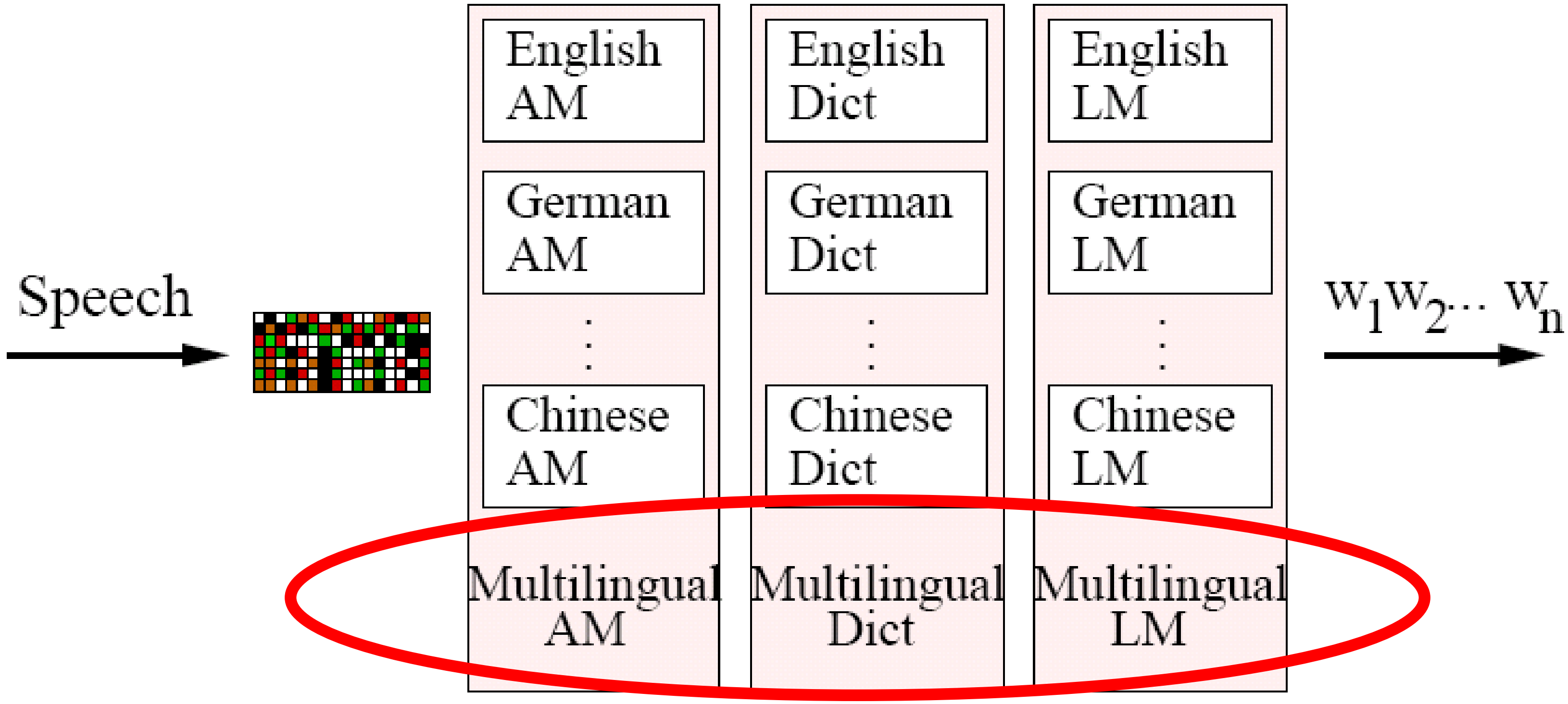
Language Identification (LID) followed by monolingual ASR



Benefits	Drawbacks
Straight-forward to implement	Computationally more expensive
Only monolingual ASR, data required	Semantic context is lost
LID error low if segment << cs	No true support of bilingualism

T. Schultz et al., *LVCSR-based Language Identification, ICASSP 1996, pp 781-784*

Approach 2: Integrated Multilingual ASR



Benefits	Drawbacks
Semantic Context is preserved	Challenging task since it requires techniques which are independent of language for AM, Dict, LM
Supports Bilingualism	
Maintenance, Scalability	

Ngoc Thang Vu , D.C Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.S. Chng, T. Schultz, H. Li, A First Speech Recognition System For Mandarin-English Code-Switch Conversational Speech, ICASSP 2012

Approach 2a: Integrated ML ASR + LID

Goal: Integrate explicit Language Detection score into ML ASR

- Two basic ideas: at word or at frame level

1. At Word-level

- Include Language Identification (LID) at word level
- Modify Language Model n-gram probabilities
- = Dynamically decrease probability of “wrong” language during the decoding process

2. At Frame-level

- Include LID at the acoustic model level
- Create two streams, one AM stream, one LID stream
- Calculate final acoustic score based on the two streams

J. Weiner, N.T. Vu, D. Telaar, F. Metze, T. Schultz, D.C. Lyu, E.S. Chng, H. Li. , Integration of Language Identification Into A Recognition System For Spoken Conversations Containing Code-Switches. SLTU 2012.

CS ASR – Options and Challenges

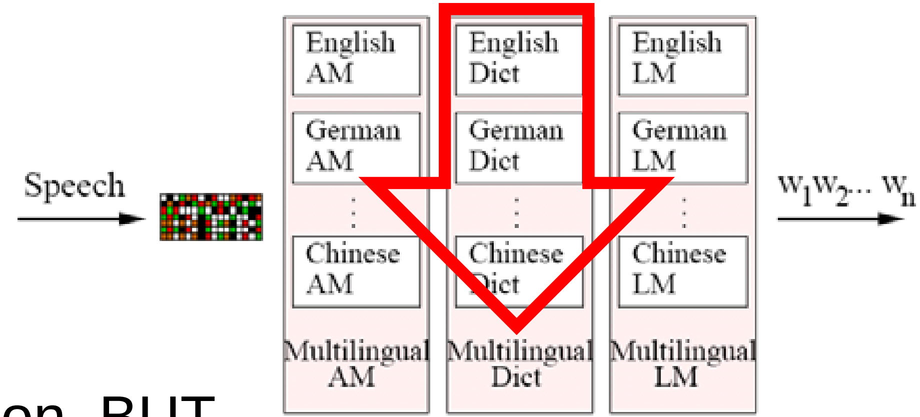
- ASR Components in multiple languages
  - Sound system and acoustic models for multiple languages
  - Pronunciation dictionaries for multiple languages
  - Borrow models/data from monolingual systems
- Share data/models across languages?
- What about language models for code-switching speech?
- Perform several of these tasks with no / little data !!!
- CS meets the definition of **“under-resourced”** languages (Krauwer 2003); A language with some (if not all) of the following aspects:
  - Lack of **electronic resources** for speech & language processing
  - Limited **presence on the web**
  - Lack of a unique **writing system** or stable orthography
  - Lack of **linguistic expertise**

L. Besacier, E. Barnard, A. Karpov, T. Schultz, Automatic Speech Recognition for Under-resourced Languages: A Survey, **Speech Communication, vol. 56, pp. 85-100, January 2014**

Multilingual ASR for CS: Dictionary

Case (1): Monolingual Dictionaries given in many languages

English	German
ONE /wʌn/	EINS /aɪ̯ns/
TWO /tu/	ZWEI /t͡svaɪ̯/
THREE /θɹi/	DREI /dʀaɪ̯/
:	:



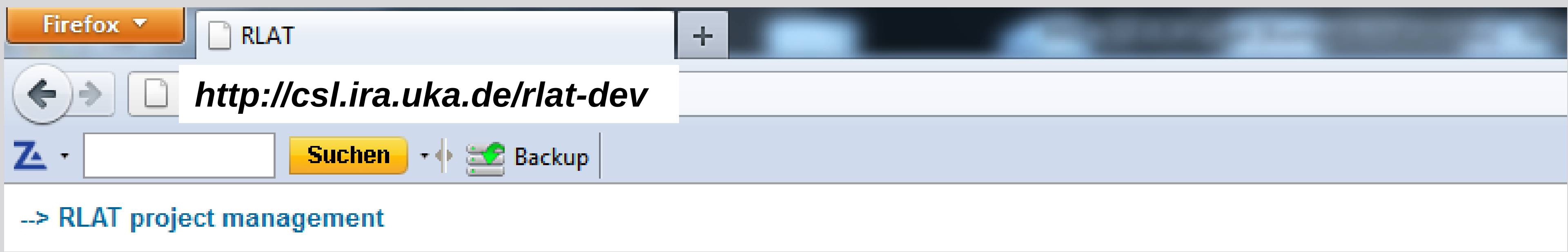
- Merge the dictionaries
  - Straightforward concatenation, BUT …
  - … Watch out for homographs (same surface form, different language, different meaning, different pronunciation), e.g.
    bald /bɔːld/ = hairless (English)
    bald /balt/ = soon (German)
  - Solution: Add a language tag “GE_bald”, “EN_bald”
- What about the phone set?
  - German and English /b/ share the same IPA symbol – same sound? (see later on Multilingual Phone Inventories)

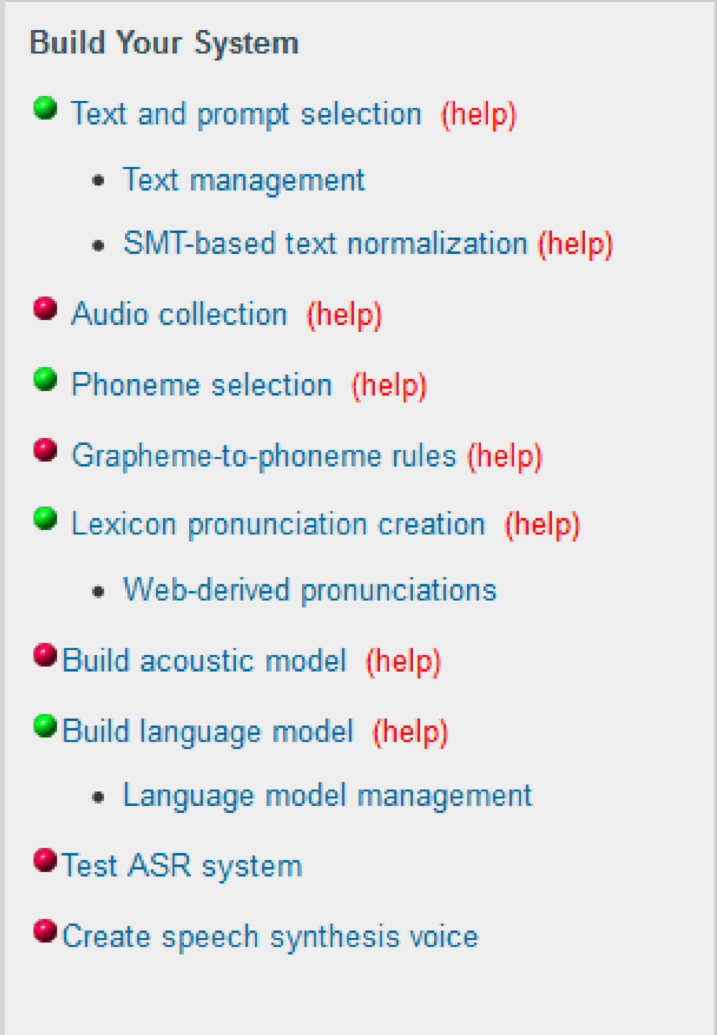
Multilingual ASR for CS: Dictionary

Cognitive Systems Lab

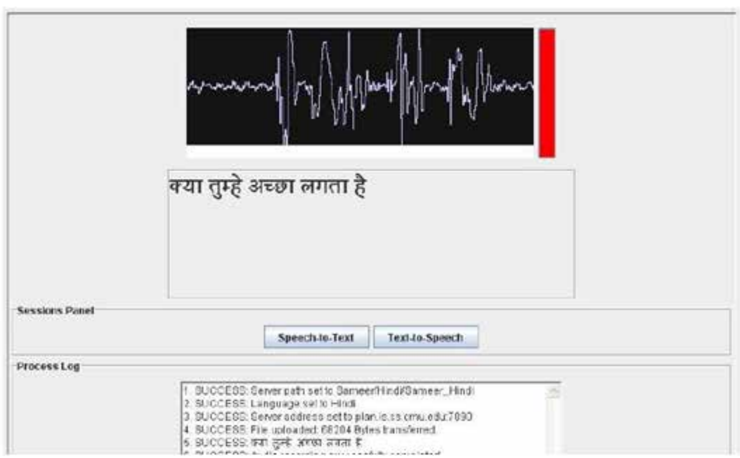
Case (2): No vocabulary / dictionary given – Create from scratch

- Use RLAT toolkit – many helpful tools to … :
  - Crawl time and topic relevant (monolingual) text corpora Snapshot functions, RSS-feeds, twitter, …
  - Automatically clean and normalize text corpora
  - Generate vocabulary lists (frequency based, tfidf)
  - Generate pronunciations using different strategies:
    - Manual time consuming), Rules (*Black et al., 1998*),
    - Heuristics and statistical models (*Besling, 1994*), (*Maskey, 2004*) (*Davel and Barnard, 2003*), (*Bisani and Ney, 2008*) Grapheme-2-Phoneme-based (e.g. Sequitur, Phonetisaurus)
    - Wiktionary and other web-resources,
    - Crowdsourcing marketplaces (e.g. Amazon Mechanical Turk)
  - Perform automated correction and filters to remove errors
  - Evaluation in terms of OOV, Phone Error Rate, WER …



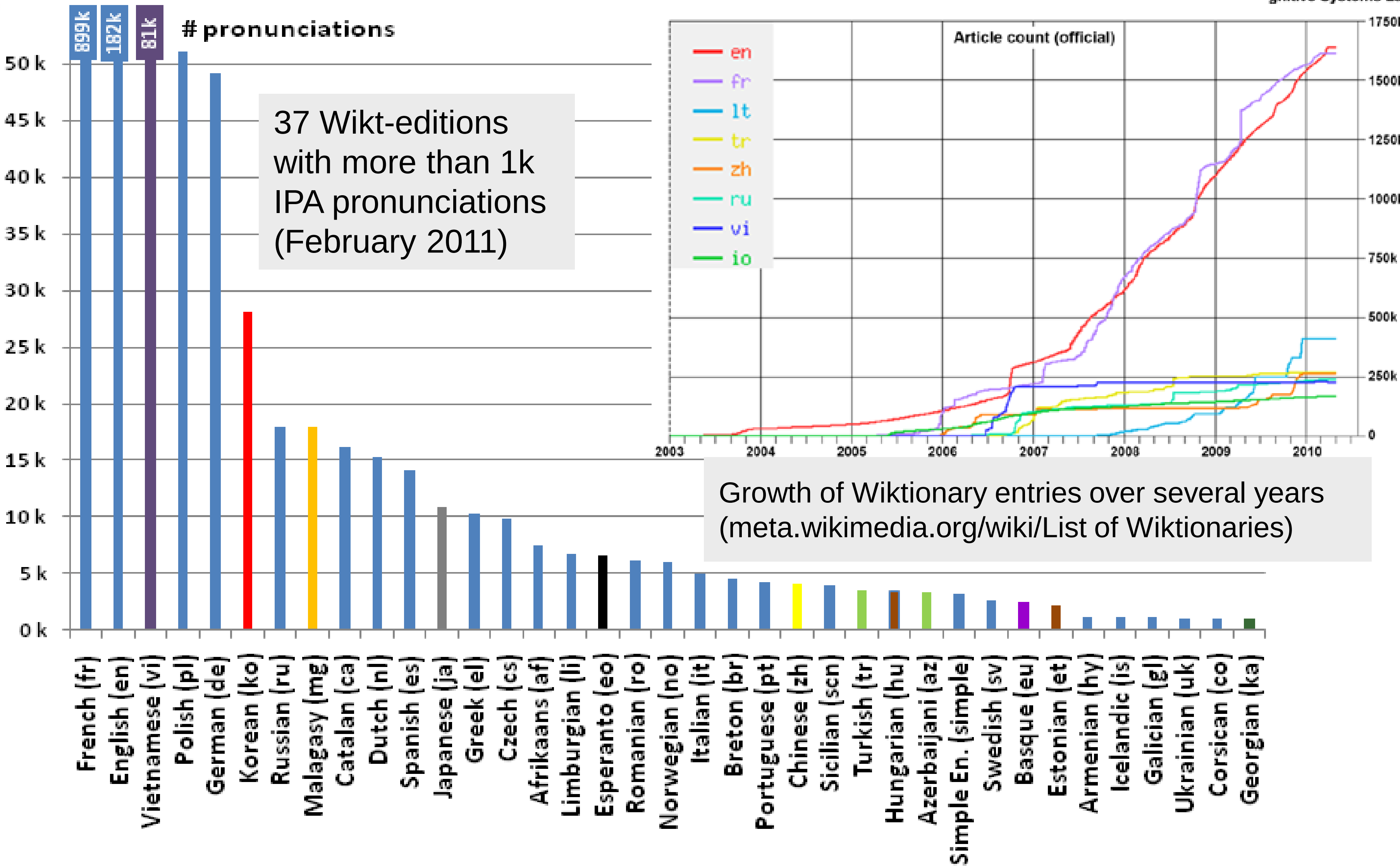


- Collect appropriate text and audio data
- Define phoneme set, prompt set
- Define and Refine pronunciation dictionary
- Produce:
  - Vocabulary / Word lists (ASR, TTS, SMT)
  - Pronunciation model (ASR, TTS)
  - Acoustic model (ASR, TTS)
  - Language model (ASR, SMT)
  - Synthetic voices (TTS)
- Maintain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



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

Pronunciations in many languages: Wiktionary

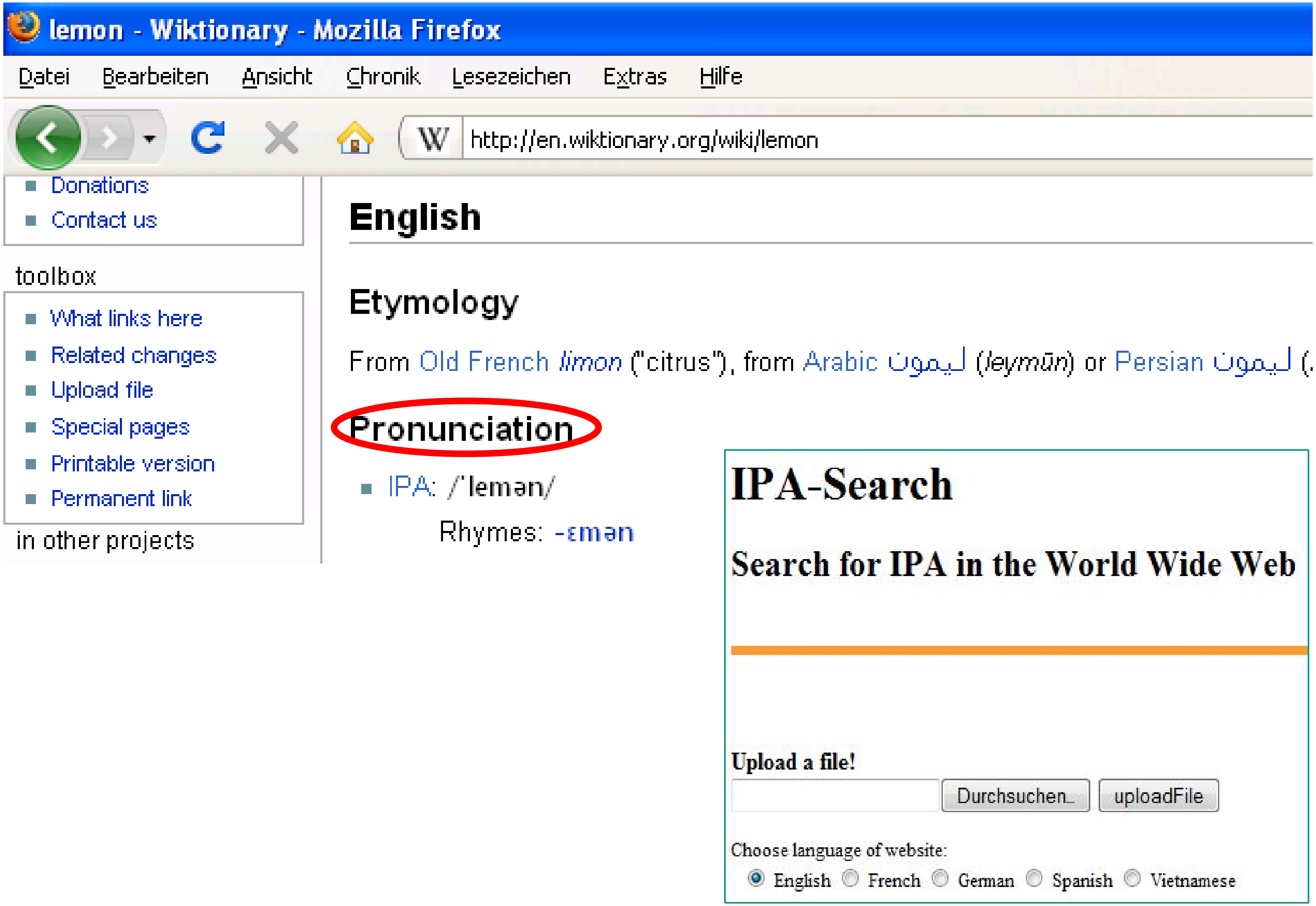


37 Wikt-editions with more than 1k IPA pronunciations (February 2011)

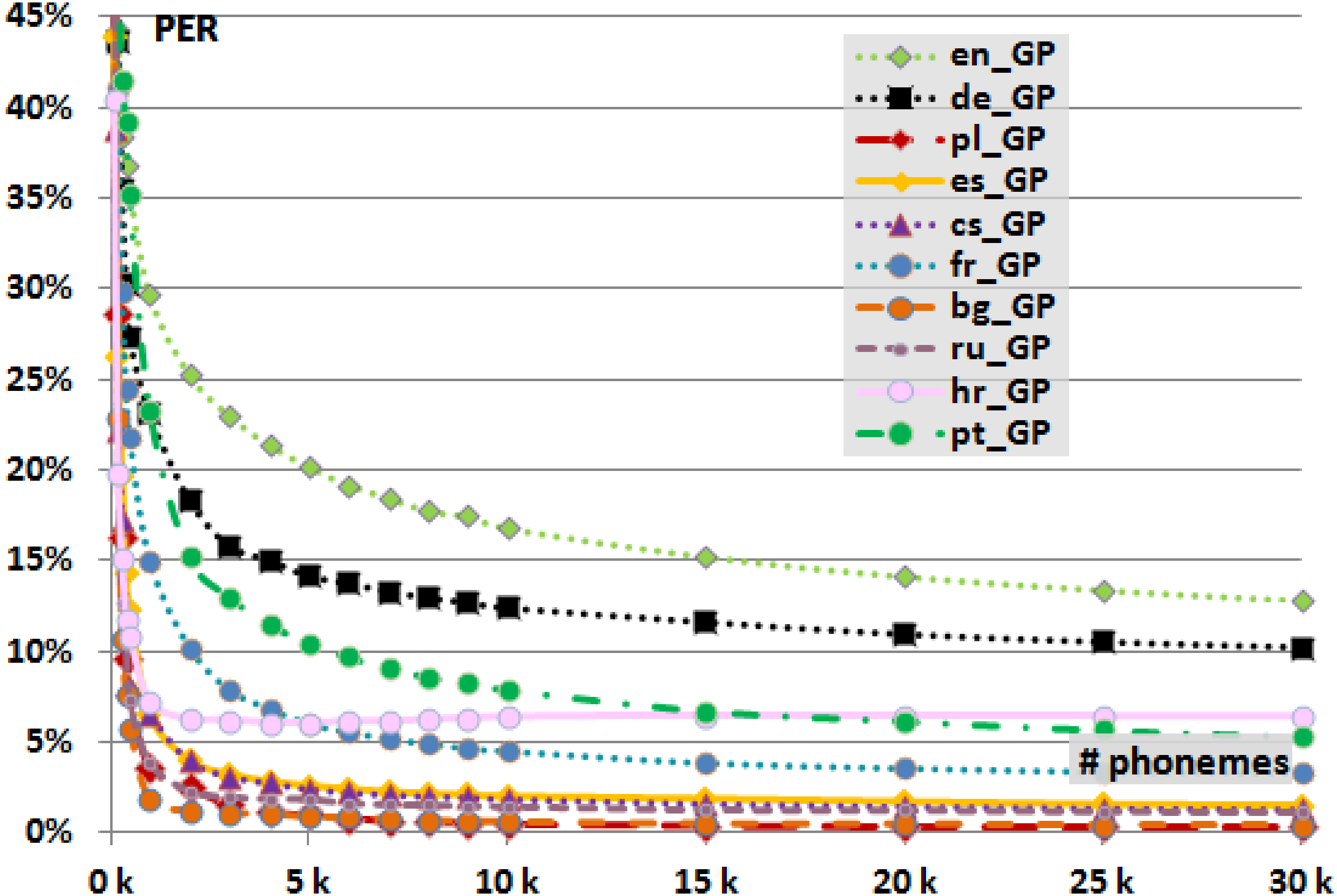
Growth of Wiktionary entries over several years (meta.wikimedia.org/wiki/List of Wiktionaries)

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



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



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

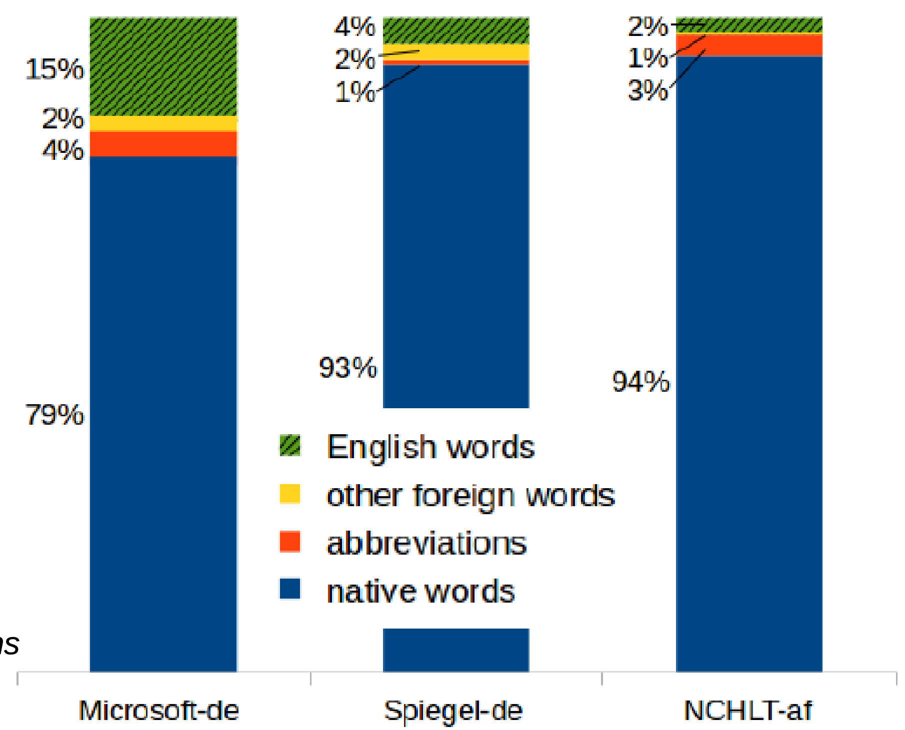
Multilingual ASR for CS: Dictionary

Additional challenges for Code-Switching Speech:

- Identify language of a word BEFORE pronunciation generation
- Phaenomenon like “Anglicisms” very prominent
- How to detect and handle Anglicisms and hybrids?

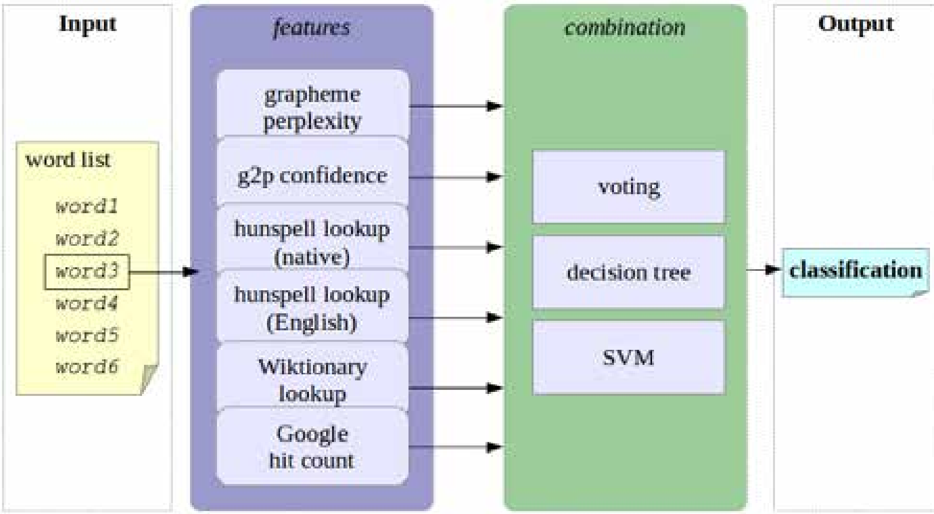
Anglicisms: words borrowed from English into the matrix language

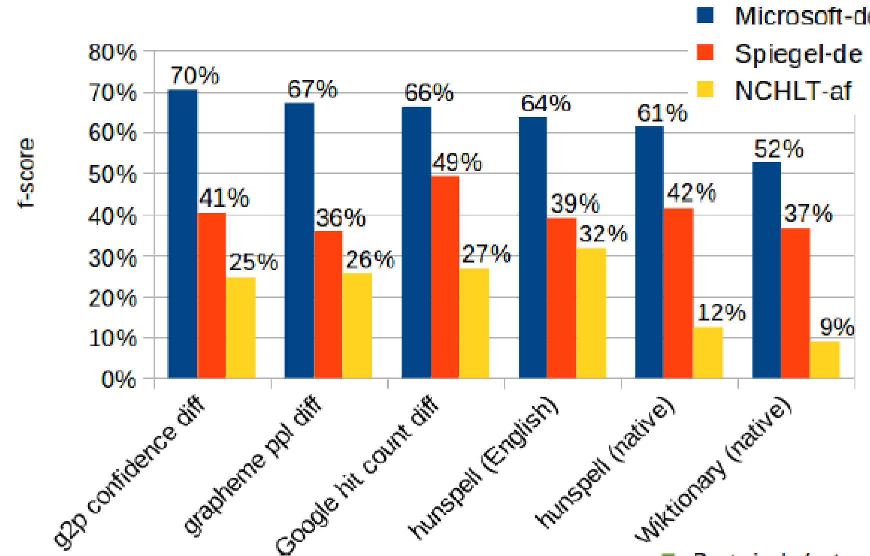
- Hybrid foreign word: Contain English plus a matrix-language part
- Example for German:
  - Compounds: “*Schad*software”
  - Inflected forms: “*ge*download*et*”
- Experiments on 2 German corpora and one Africaans NCHLT (thanks to M. Davel/E. Barnard, North-West University, SA)



S. Leidig, T. Schlippe, T. Schultz, *Automatic Detection of Anglicisms for the Pronunciation Dictionary Generation: A Case Study on a German IT Corpus*. SLTU, St. Petersburg, Russia, 2014.

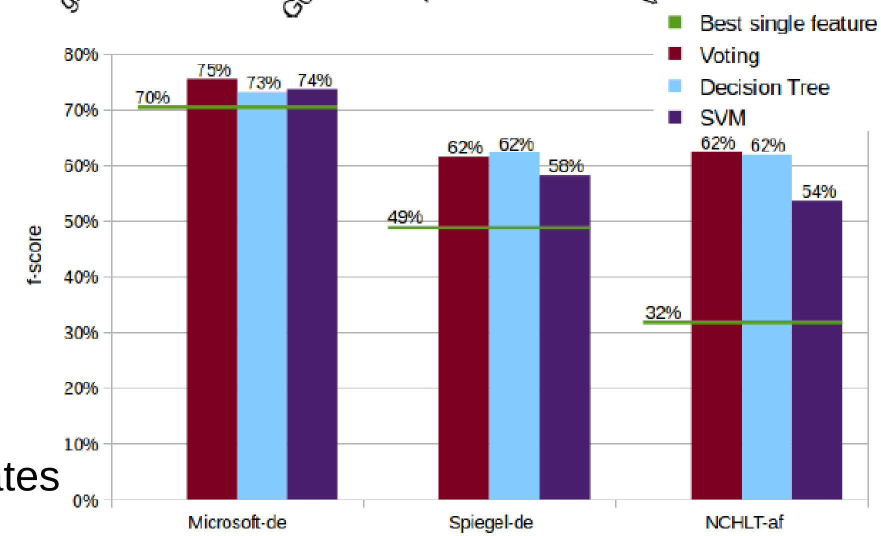
Automatic Anglicism Detection





	PER
Automatic Anglicism Detection	1.61%
German G2P Model	4.95%
Mixed Language 80:20	4.97%
Mixed Language 50:50	5.46%
English G2P Model	39.66%

Dictionary generation (German IT) with 5 different approaches; Quality in terms of Phoneme Error Rates



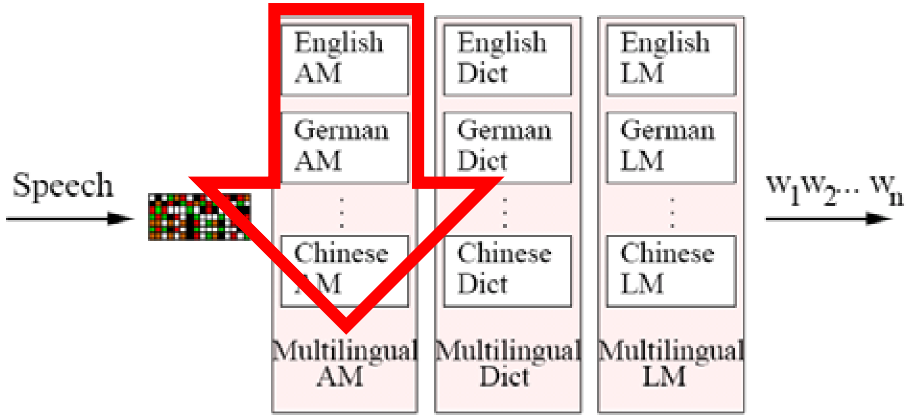
Leidig et al. *Automatic Detection of Anglicisms for the Pronunciation Dictionary Generation,* SLTU, St. Petersburg, Russia, 2014.

Multilingual ASR for CS: Acoustic Model

- Provide Acoustic Models for many languages

English
ONE /wʌn/
TWO /tu/
THREE /θɹi/
:

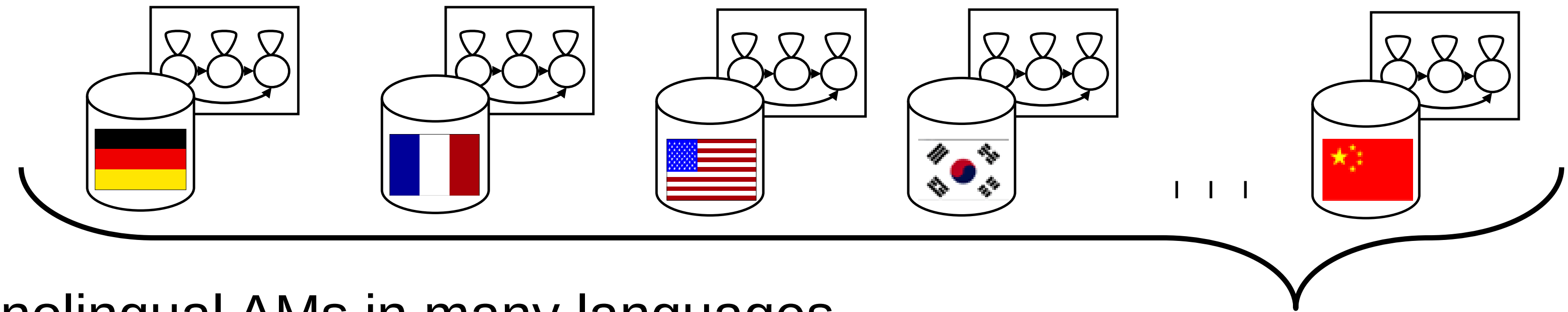
German
EINS /aɪ̯ns/
ZWEI /t͡svaɪ̯/
DREI /dʀaɪ̯/
:



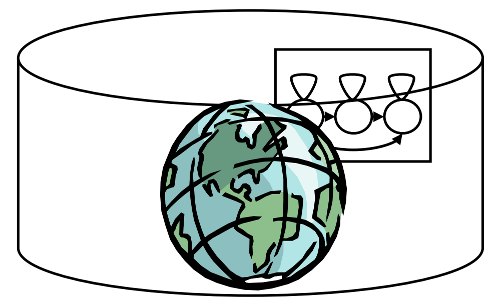
- Merge two monolingual acoustic models into ONE
  - Keep language dependent sets, i.e. $/n/_{GE}$, $/n/_{EN}$, $/t/_{GE}$, $/t/_{EN}$, …
- Same IPA symbol – same sound: $/n/_{GE}/ = /n/_{EN}$?
  - If so, shall we share data across languages to create truly multilingual acoustic models?
  - What's the best strategy to build multilingual acoustic models?

T. Schultz and A. Waibel: Language Independent and language adaptive acoustic modeling for speech recognition, Speech Communication, 35, pp. 31-35, 2001.

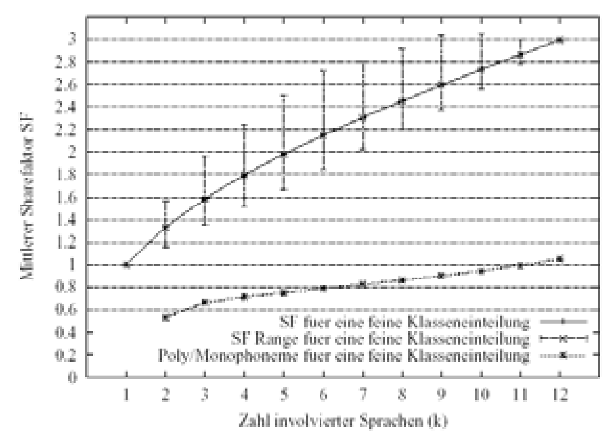
Multilingual Acoustic Modeling



Monolingual AMs in many languages
® Multilingual AM “language independent”

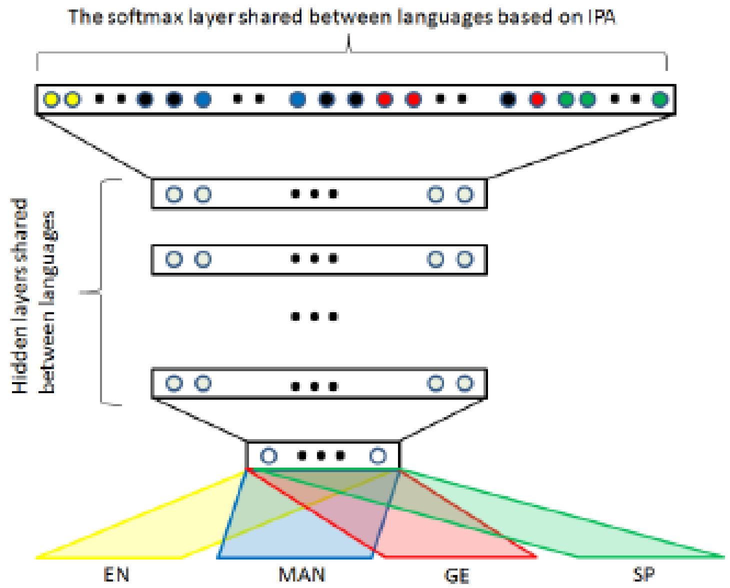


- IPA-based data sharing
  - Köhler 1997, Schultz 1997 (Context-independent)
  - On 12 languages: 485 ® 162 (sharing factor ~3)
  - Context-dependent AMs, PDTS (Schultz, 1999)
  - Articulatory features (Stüker et al. 2003)
- Mono outperformed ML on training language
- BUT: ML gives benefits on unseen 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 on 6 languages
  - Kulback-Leibler HMM-based (Imseng)
  - Hybrid decoding (pseudo-likelihoods
    instead of state emission probs in HMM)



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

<u>Multilingual Database</u>

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

<u>Corpus</u>

- 21 Languages … counting
- ³ 2000 native speakers
- ³ 450 hrs Audio data
- Read Speech
- Filled pauses annotated

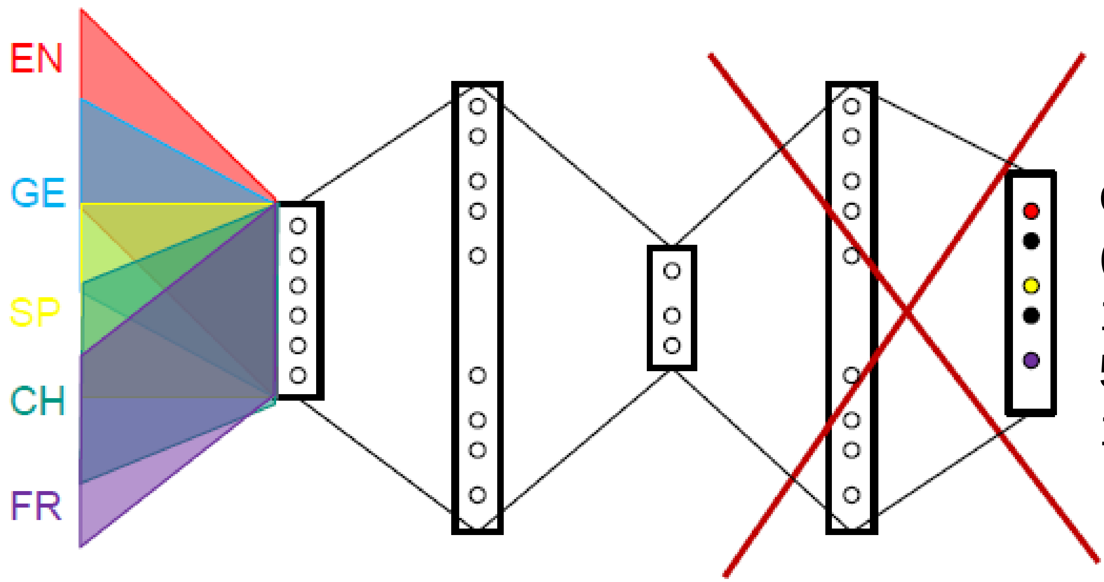
Available from ELRA, Appen

Tanja Schultz (2002): GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University, ICSLP Denver, CO
Tanja Schultz (2013): GlobalPhone: A Multilingual Speech and Text Database in 20 Languages, ICASSP, Vancouver 2013.

Multilingual Bottleneck Features

Idea:

- Classify acoustic units → language independent
- Using multilingual data resources



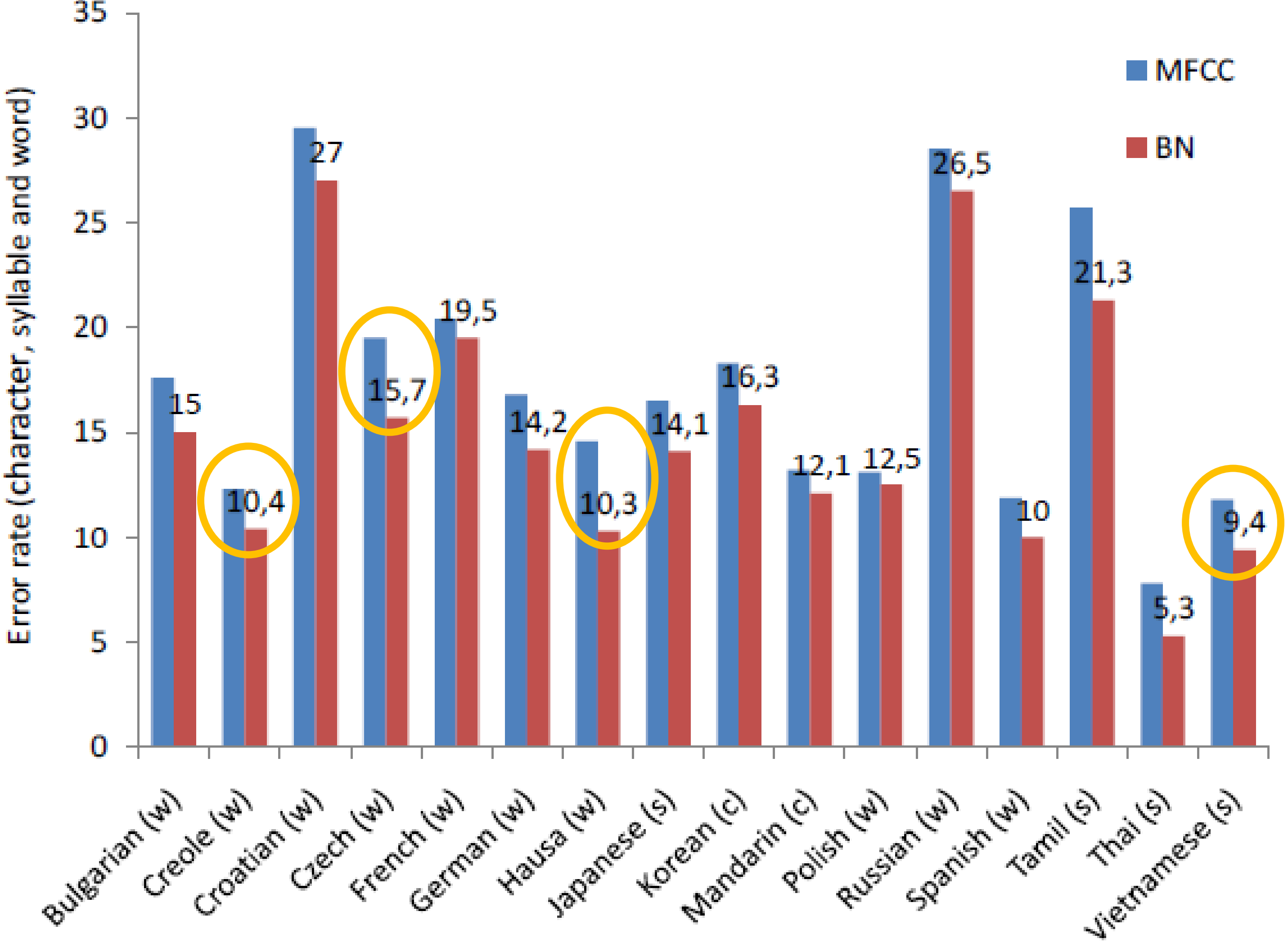
Global phone set CI
(based on IPA)
12 languages
5-layer MLP, Quicknet
143-1500-42-1500-**x**

Benefit:

- Robust due to large amount of data
- Combine knowledge between languages
- Allow training with less data

Vu, Metze, Schultz, Multilingual Bottle-Neck features and its application to under-resourced languages, SLTU 2012

Multilingual BNF on GlobalPhone-16



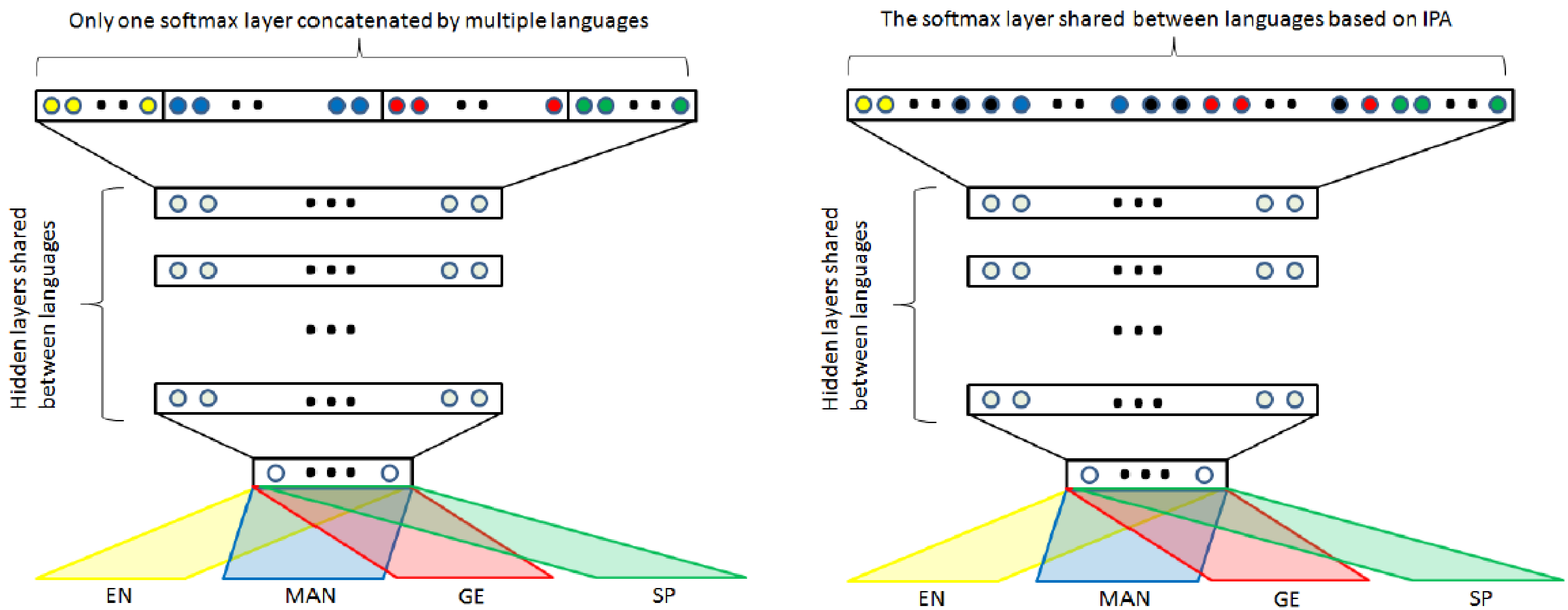
BN trained from on 12 languages: BL, EN, FR, GE, HR, JA, KO, MA, PL, RU, SP, TH
ASR performance improves with number of languages AND amount of data

Multilingual AM with DNNs

Cognitive Systems Lab

Sharing Phone sets:

No sharing (left) versus IPA-based sharing (right)



Ngoc Thang Vu, David Imseng, Daniel Povey, Petr Motlicek, Tanja Schultz, Hervé Bourlard,
Multilingual Deep Neural Network Based Acoustic Modeling For Rapid Language Adaptation, ICASSP 2014

Karlsruhe Institute of Technology

Multilingual DNN

- **Setup:** Used 6 GlobalPhone languages (BG, EN, GE, JA, MA, SP)
  - DNNs trained on second DNN implementation of KALDI (no RBMs but greedy layer-wise supervised training GL-ST)
  - 11 frames, 13 MFCCs, DNN 6000 tied-state triphones, 5 layers, 1500 units
- **Monolingual baselines**: greedy layer-wise supervised DNNs, fine-tuned
- **Crosslingual transfer**:
  - Transfer all hidden layers of the DNN-MUL to target language,
  - Replace softmax layer with new output for target language,
  - Random initialization of weights and biases of last hidden to output layer
- **Results:** Table 7.2: *Word error rates (WER) on BG, EN, GE, JA, MAN, and SP test data using greedy layer-wised supervised training DNN and DNNs which were pre-trained using multilingual DNNs*

Systems	BG	EN	GE	JA	MAN	SP
DNN (GL-ST)	17.4	9.9	6.2	16.8	12.3	14.9
DNN-MUL-SEP	16.8	9.5	**5.8**	16.2	**11.8**	**14.3**
DNN-MUL-IPA	**16.7**	**9.2**	**5.8**	**16.1**	**11.8**	**14.3**

→ *Crosslingual transfer better than monolingual (GL-ST)*

→ *IPA phone sharing gives same or lower WER than SEP*

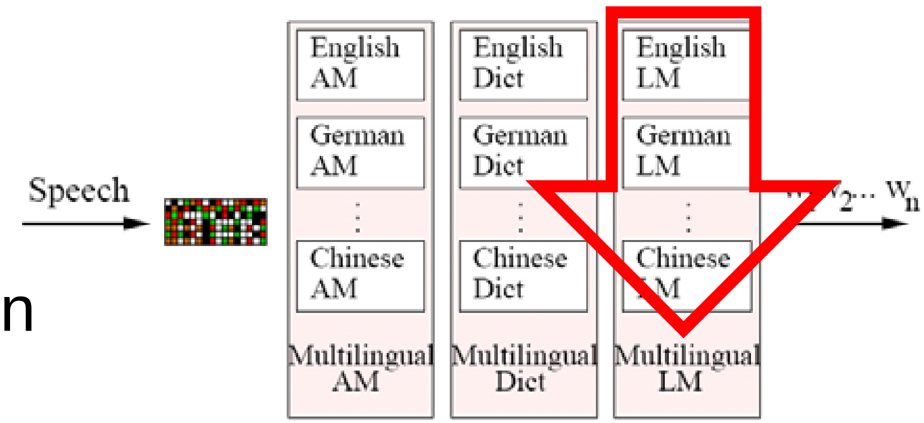
Challenges CS ASR: Language Models

- Language Models for CS – stochastic model: N-grams

  3-gram: $P(w_n|w_{n-2}w_{n-1})$

  2-gram: $P(w_n|w_{n-1})$

  1-gram: $P(w_n)$



- CS is a spoken phenomenon
  - no /little written text
  - Manual transcription costly
    multilingual experts, conversational data, tough task
  - Statistical modeling: requires HUGE amounts of data
  - Grammars: not feasible for conversational speech

- Simply stringing together monolingual text resources
  - BUT: CS within utterances and at phrase boundaries
    requires transcripts of valid code-switching speech

SEAME Data

<u>SEAME: **S**outh-**E**ast **A**sia **M**andarin-**E**nglish</u>

- Speaking style: conversations and interviews
- High quality audio: 16KHz sampling, 16-bit resolution
- Orthographically transcribed, UTF-8 code
- CS 49.6hrs, Mandarin-only 7.6hrs, English-only 4.2hrs, others 2.5hrs

SEAME ALL	NTU		USM	ALL
	conversation	interview	interview	
number of speakers	61 (F 34, M 27)	67 (F 36, M 31)	29 (F 14, M 15)	157 (F 84, M 73)
number of utterances	13,112	19,586	19,447	52,145
number of hours	11.54	22.65	28.75	62.94

D. Lyu, T. Tan, E. Chng and H. Li, "An Analysis of a Mandarin-English Code-switching Speech Corpus: SEAME", Interspeech, Japan, 2010

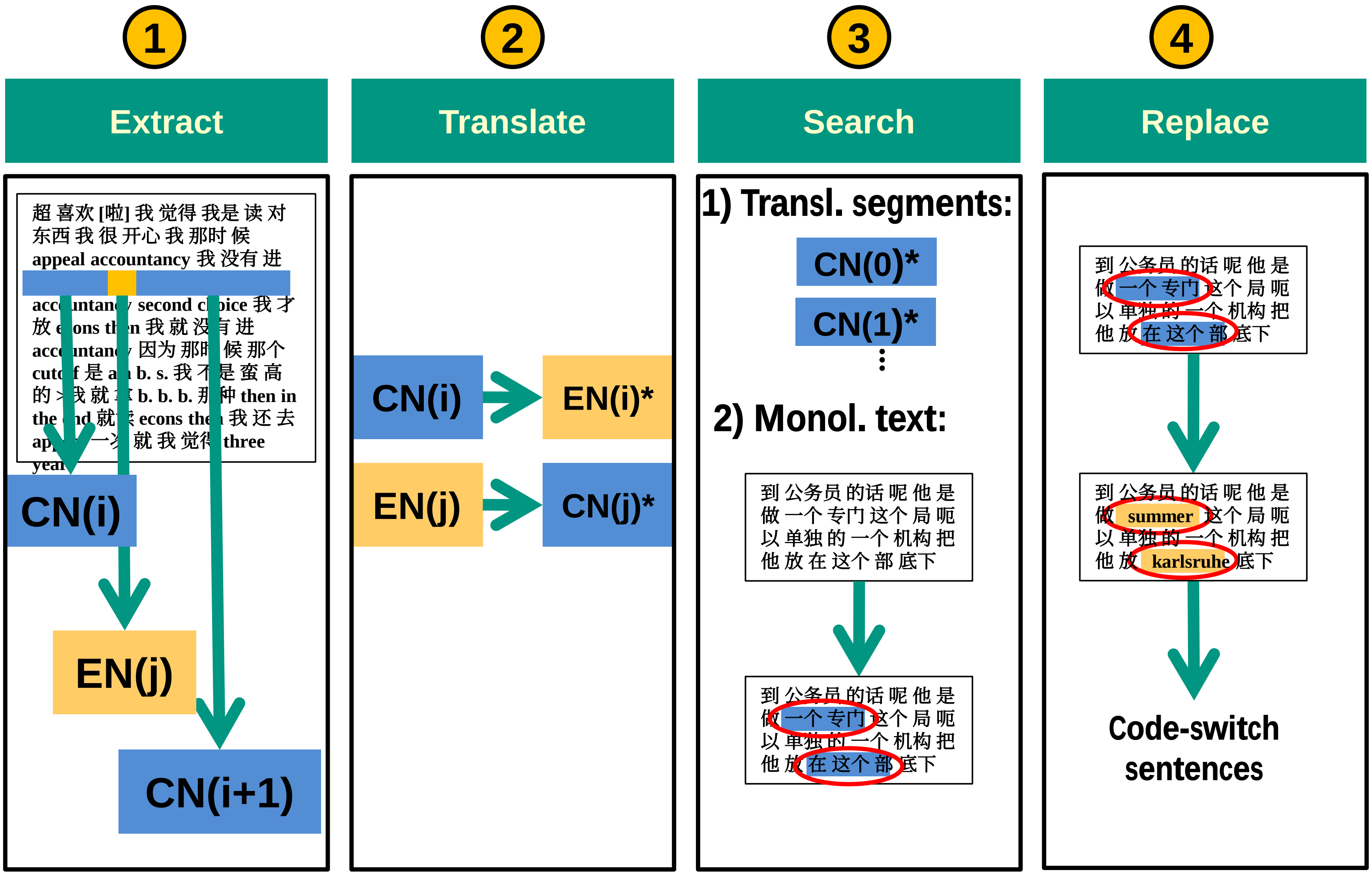
29 T. Schultz, http://csl.anthropomatik.kit.edu

N-gram Language Models for CS-ASR

- We implemented several ideas to create statistical LMs
  - Evaluated on the SEAME corpus
  - Evaluation criteria: PPL and Mixed Error Rate MER = (CER + WER)/2

Approach for Language Modeling	Dev (1)	Dev (2)	Eval
Oracle Experiment: full coverage of CS trigrams	28.5%		
Idea 1: 3-gram on 50hrs CS transcripts only	50.5%		
Idea 2: Interpolate 3-gram with monolingual data (Giga)	50.1%		
Idea 3: MT to create artificial CS text (large PPL reduc.)	49.8%	36.9%	
Idea 4: Class-based LM (Auto/POS)	No improvements		
Idea 5: Recurrent NN LM + Output factorization + Feature integration		35.6% 34.7%	29.3% 29.2%
Idea 6: Factored LM (POS+LID)		35.2%	29.7%
Idea 7: Combined RNNLM + FLM (POS + LID)		34.4%	29.2%
Idea 8: Speaker Clustering + Adaptation of combined LM		34.0%	28.8%

Idea 3: Generate CS Text from monolingual source



Repeat the analogous approach for monolingual English text

Idea 4: Part-of-Speech (POS) for LM

- Bilingual LM that predicts language changes

  => **Analysis**: Do words or features predict CS-points?

word	frequency	CS-rate
那个(that)	5261	53.43 %
我的(my)	1236	52.35 %
那些(those)	1329	49.44 %
一个(a)	2524	49.05 %
他的(his)	1024	47.75 %
then	6183	56.25 %
think	1103	37.62 %
but	2211	36.23 %
so	2218	35.80 %
okay	1044	34.87 %

Mandarin trigger words

English trigger words

Trigger POS for Code-Switching

Tag	meaning	count CS	count all	cs rate
DT	determiner	4560	11276	40.44 %
DEG	associative 的	1622	4395	36.91 %
VC	是	1598	6183	25.85 %
DEC	的 in a relative-clause	1375	5763	23.86 %
M	measure word	610	2612	23.35 %
NN	noun	24073	49060	49.07 %
NNS	noun (plural)	1883	4613	40.82 %
RB	adverb	6716	21096	31.84 %
JJ	adjective	2875	10856	26.48 %
CC	coordinating conjunction	1058	4400	24.05 %

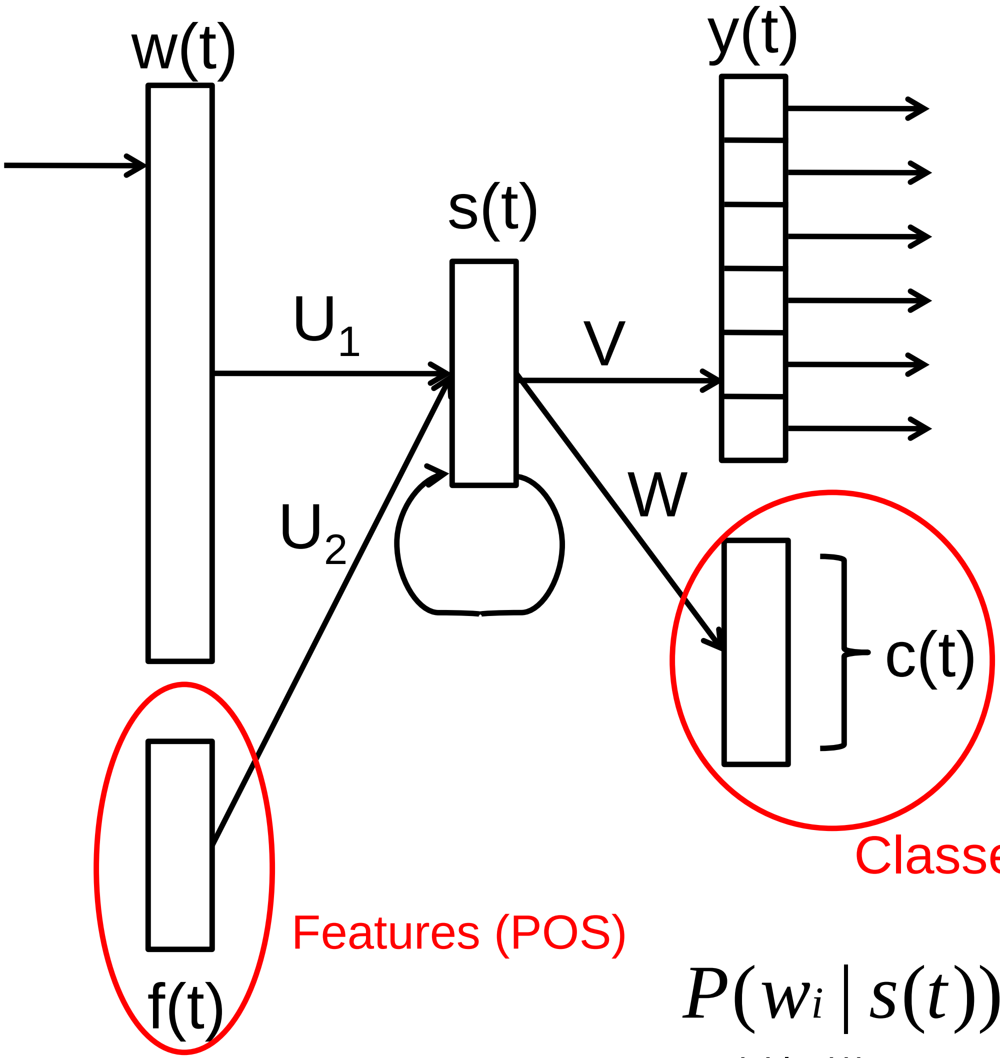
Mandarin trigger POS

English trigger POS

- Using POS in a standard 3-gram LM gave no gains

=> Is POS-Information useful for something else?

Idea 5: RNNLM for Code-Switching



Features (POS)

POS get propagated into hidden layer and back-propagated into its history

Based on:
Tomas Mikolov, M. Karafiat, L. Burget, J. Cernocky, S. Khudanpur (2010) „Recurrent neural network based language model", Interspeech 2010.

t = time
w(t) = current word
s(t) = hidden layer
y(t) = next word
c(t) = class of next word
U, V, W = weights

Classes = language ID of the words

$$P(w_i \mid s(t)) = P(c_i \mid s(t)) \times P(w_i \mid c_i, s(t))$$

P(ci | s(t)) computes the next language ci using information of previous words and previous features.

Results on RNNLM

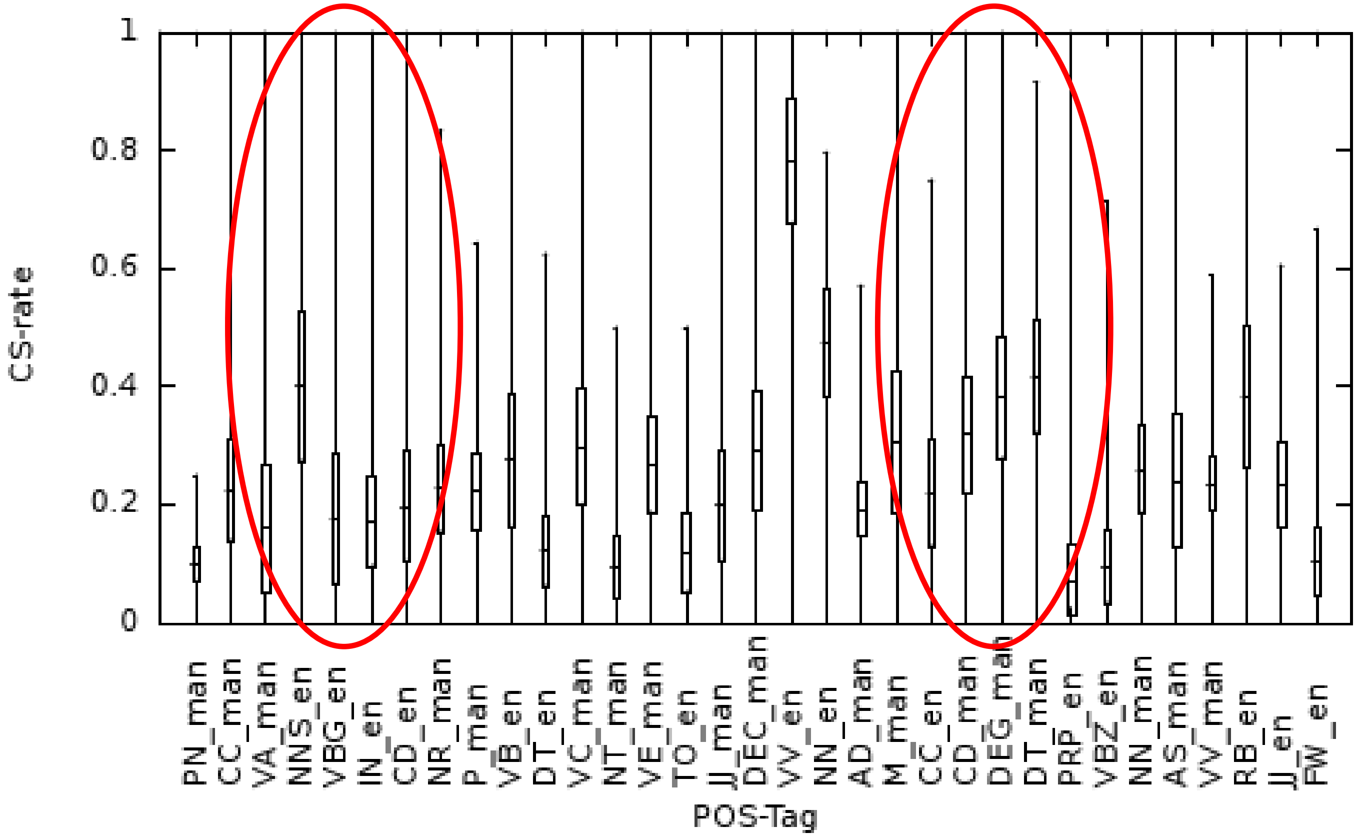
Model	PPL dev	PPL eval	MER dev	MER eval
3-gram	-	-	35.5 %	30.0 %
RNNLM	246.60	287.88	35.6 %	29.3 %
RNNLM + OF	239.64	269.71	34.9 %	29.4 %
RNNLM + FI	233.50	268.05	34.8 %	29.3 %
RNNLM + FI + OF	**219.85**	**239.21**	**34.7 %**	**29.2 %**

OF= Output factorization into language classes
FI = feature (POS) integration into the input layer

H. Adel, N.T. Vu, F. Kraus, T. Schlippe, H. Li, T. Schultz: Recurrent Neural Network Language Modeling for Code Switching Conversational, International Conference on Acoustics, Speech, and Signal Processing, 2013

Speaker Dependent Analysis

- Analysis: Is CS speaker dependent? ('CS attitude')



=> high spreads between min and max CS rates per speaker

=> high standard deviations of CS rates among speakers

Clustering Speakers

- **Idea**: cluster speakers according to their CS attitudes
- Define vector for each speaker:

$$\boxed{spk = [\ f_{CS}(POS_1) / f(POS_1), \ldots, f_{CS}(POS_n) / f(POS_n)\ ]}$$

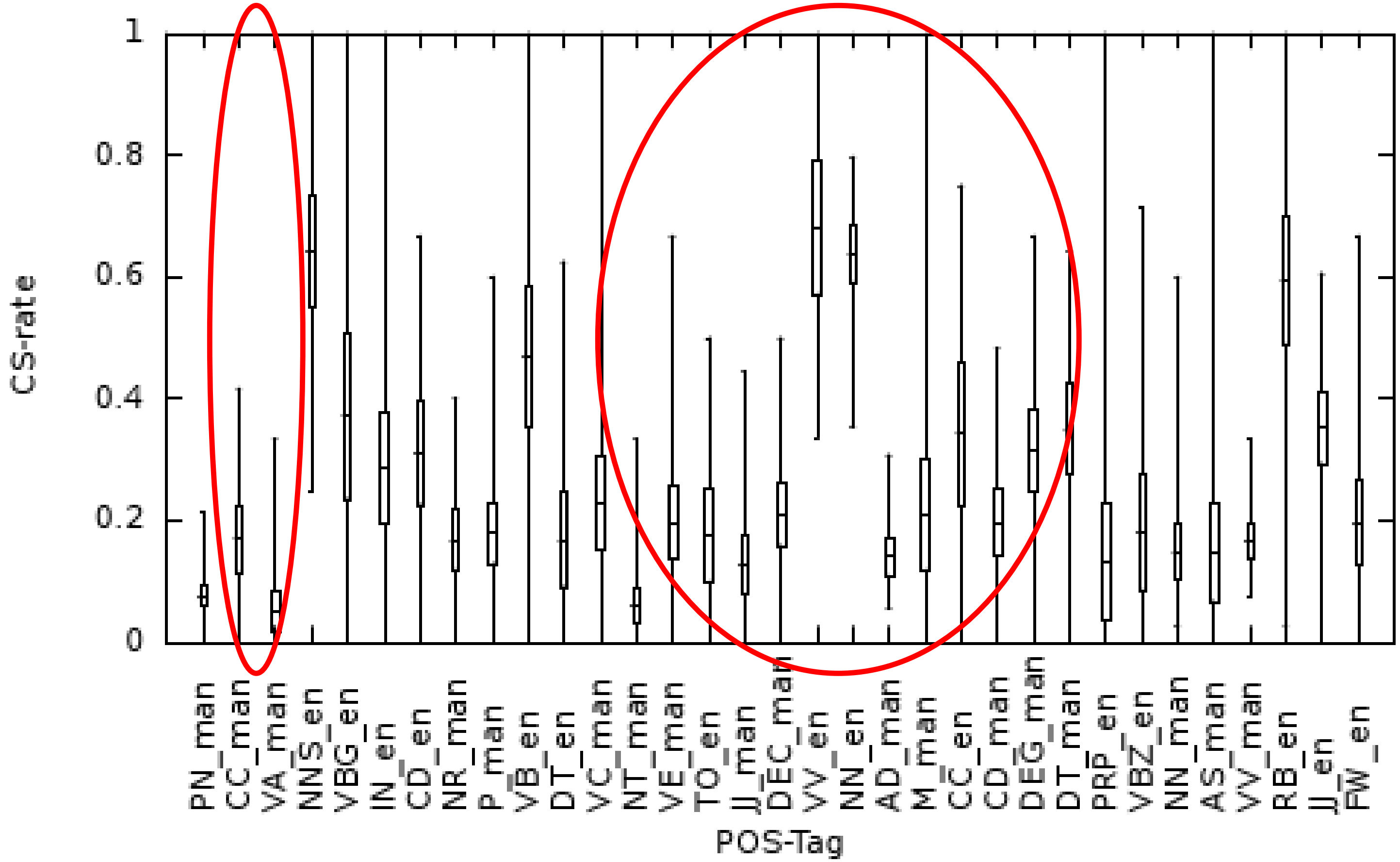
(f: frequency of POS tag, f_{CS}: frequency in front of CS-point)

- Cluster vectors into K classes using **k-means** and cosine similarity as distance measure
- **Cosine similarity**:

$$\boxed{Sim(spk1, spk2) = (spk1 \cdot spk2) / (||spk1|| \cdot ||spk2||)}$$

Example: Cluster of 3 Classes

- Example: Cluster 1:
- Clustering decreases the spread of CS rates:



Speaker Dependent Models

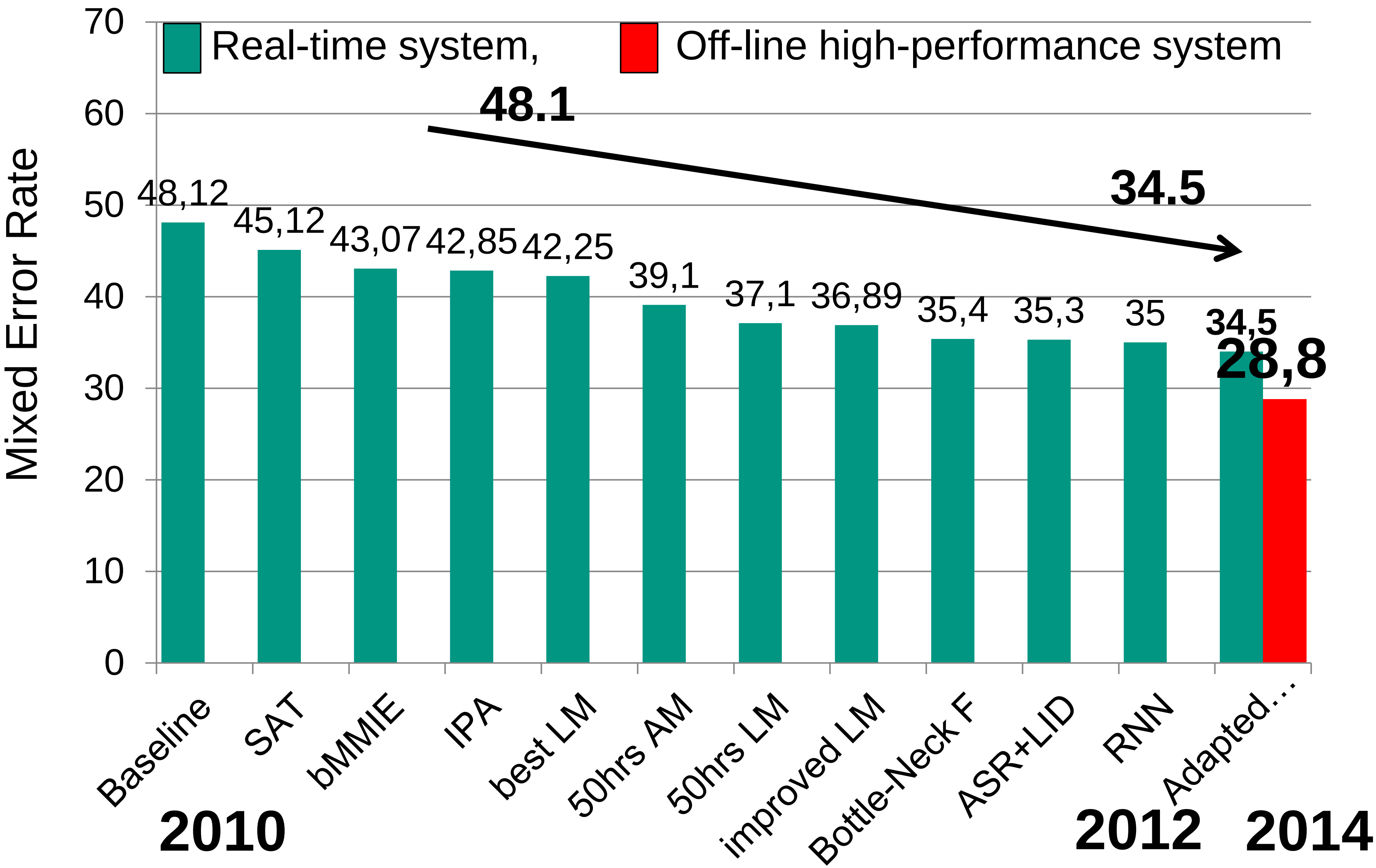
- Adapt speaker independent RNNLM to each class (one-iteration-retraining)
- Adapt the speaker independent 3-gram to each class (interpolation with a class-specific 3-gram)
- Perform speaker wise evaluation

Speaker	N-gram	Adapted N-gram	RNNLM	Adapted RNNLM
Speaker 1	317.84	**302.94**	200.66	**197.74**
Speaker 2	265.77	**253.73**	181.60	**175.85**
Speaker 3	327.09	**302.56**	187.04	**170.92**
Speaker 4	232.83	**213.33**	174.13	**160.58**

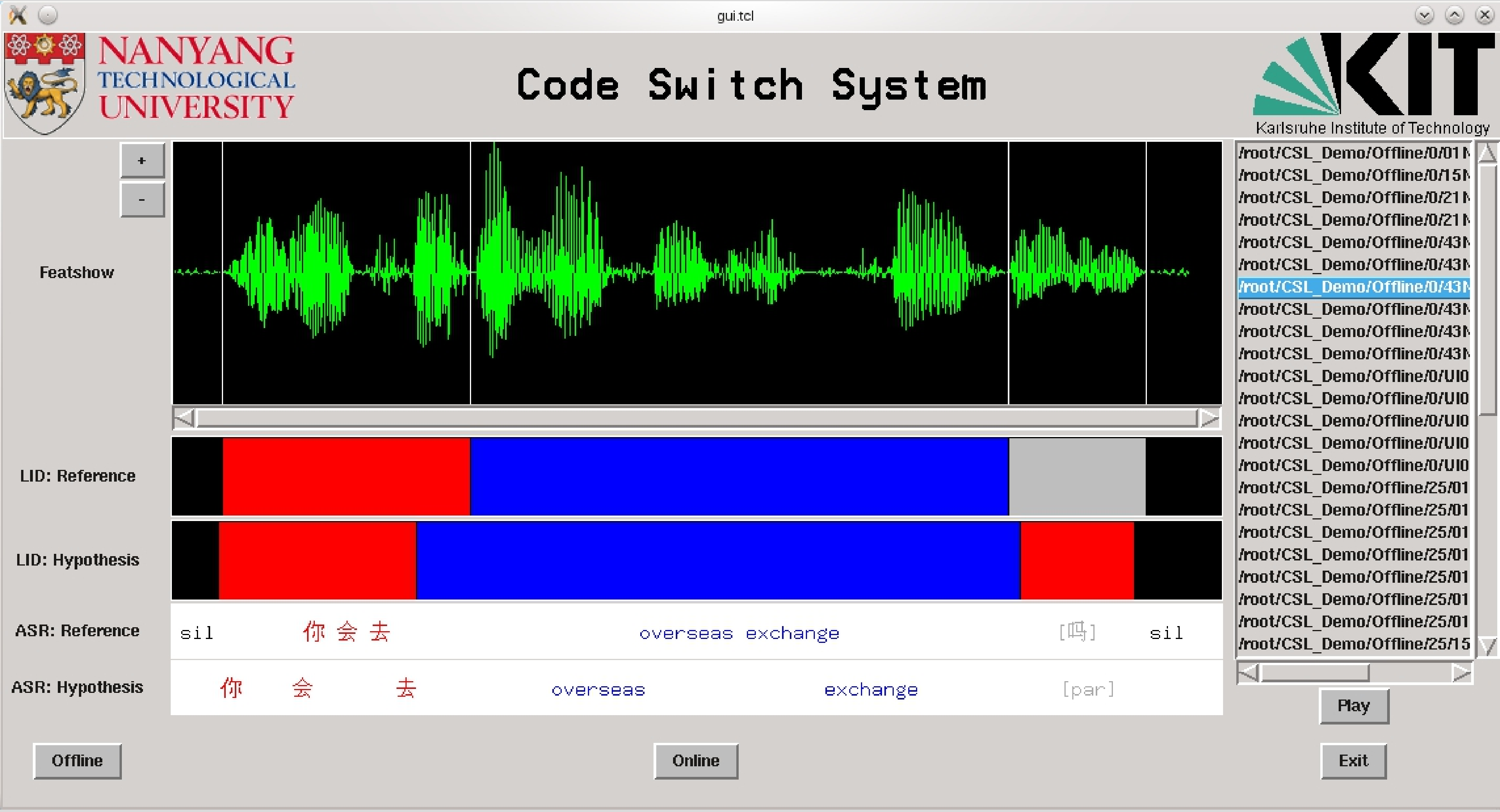
- Use adapted 3-gram to **decode**, then adapted RNNLM to rescore 100-best
- Baseline: ASR system without LM adaptation

Model	Dev set	Eval set
Baseline	34.74 %	29.23 %
Adapted 3-gram + RNNLM	34.47 %	28.89 %
Adapted 3-gram + RNNLM + FM	**34.0 %**	**28.80 %**

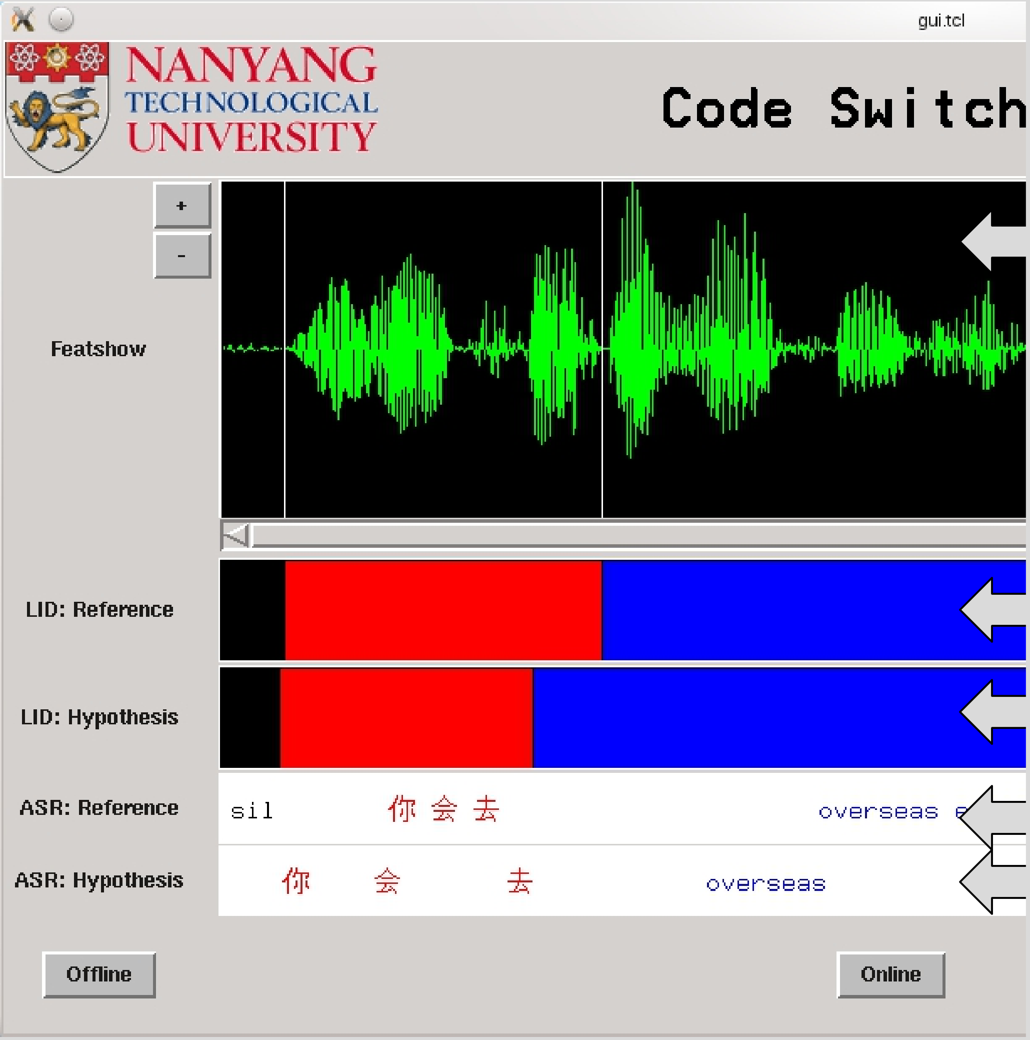
Progress Code-Switch System (MER)



Code-Switch Prototype System 2011



Code-Switch Prototype System 2011



First-time Joint Integration of Language Identification (LID) from NTU and Automatic Speech Recognition (ASR) from KIT

Display for recorded Audio File (y-axis = amplitude, x-axis = time)

Language ID Ground Truth

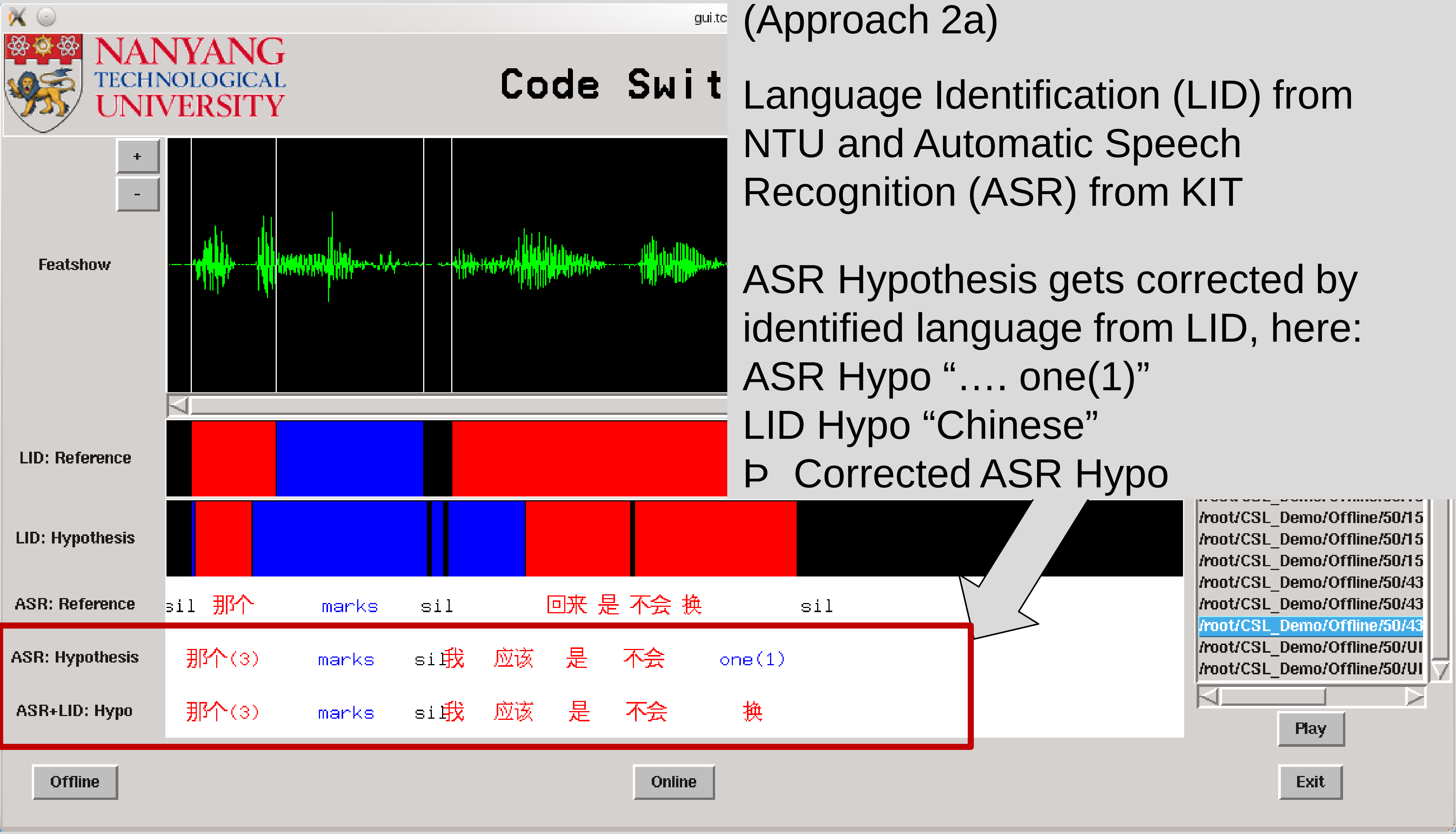
Automatic Language Identification

Manual Transcription of Audio File

ASR Output

- red script: Mandarin spoken language
- blue script: English spoken language
- gray script: silence, particle

Code-Switch Prototype System 2012



Tight Integration of LID and ML-ASR (Approach 2a)

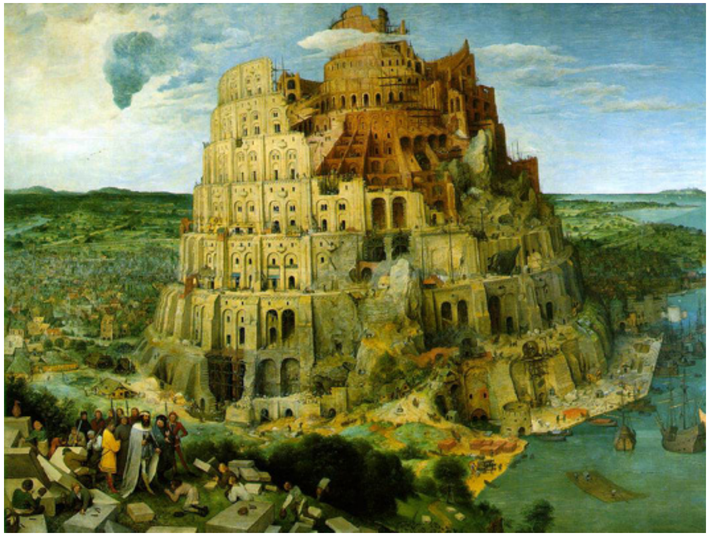
Language Identification (LID) from NTU and Automatic Speech Recognition (ASR) from KIT

ASR Hypothesis gets corrected by identified language from LID, here:
ASR Hypo "…. one(1)"
LID Hypo "Chinese"
Þ Corrected ASR Hypo

Summary and Remaining Challenges

- Experiments and Results
  - Acoustic level: sharing data gives good improvements
  - Dictionary level: straightforward, investigate phone sharing
  - Language level: very challenging
    - Words and POS can be used to predict Code-Switching points
    - Integration of POS and LID information into RNNLM significantly improves the LM perplexity
- Issues: Few data resources
  - Speech: Monolingual data from source languages?
  - Text: How to get more text?
  - CS-points: Speaker dependent
- Integrated End-to-End system
  - Oracle experiments indicate lots of room for improvement
  - For offline usage do multi-pass, CNC, larger models
  - Challenging problem, lack of benchmarks, lack of databases

THANK YOU



University of the State of Baden-Württemberg and National Research Center of the Helmholtz Association

Thanks to collaborators: Eng-Siong Chng, Pascal Fung, David Imseng, Katrin Kirchhoff, Haizhou Li, Dau-Cheng Lyu, and Dan Povey

Thanks to CSL-students: Heike Adel, Fabian Blaicher, Christoph Burgmer, Franziska Kraus, Sebastian Leidig, Sebastian Ochs, Tim Schlippe, Dominic Telaar, Ngoc Thang Vu, and Jochen Weiner

Cognitive Systems Lab