Karlsruher Institut für Technologie

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation

Cognitive Systems Lab

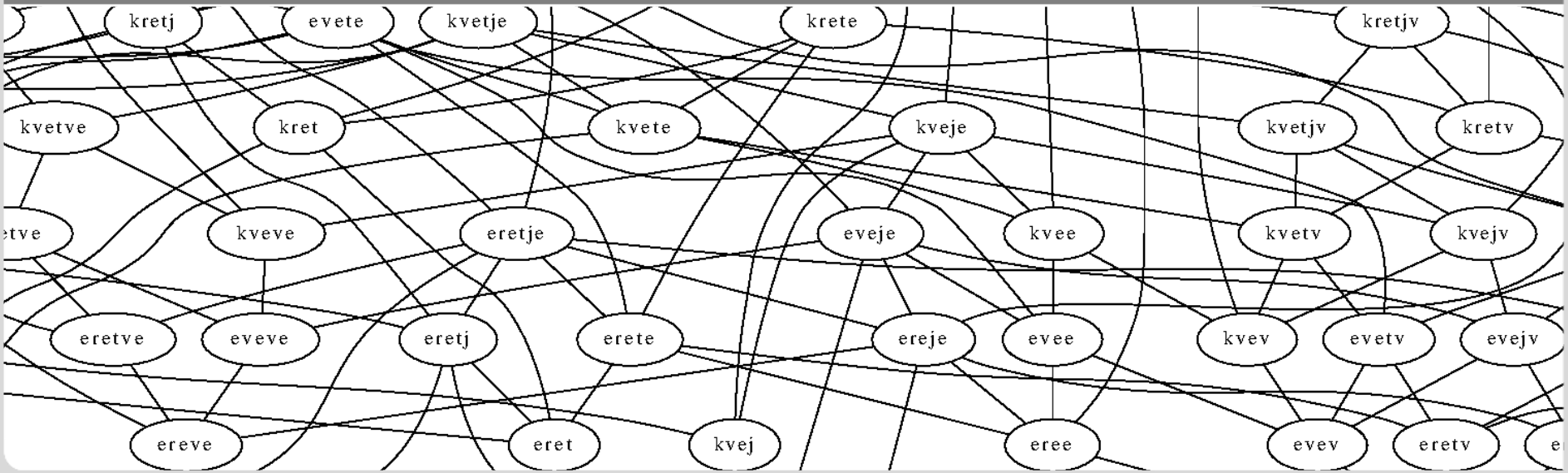
Towards ASR Without Pronunciation Dictionary, Transcribed Speech and Text Resources in the Target Language Using Cross-Lingual Word-to-Phoneme Alignment

Felix Stahlberg, Tim Schlippe, Stephan Vogel, Tanja Schultz

14 May 2014

Cognitive Systems Lab
Karlsruhe Institute of Technology

4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'14)



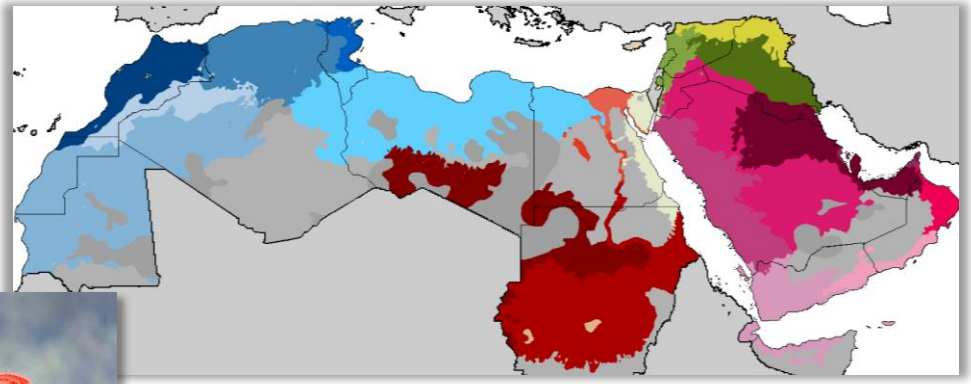
Goal

- Exploit the phonetic output of a human simultaneous translator
  - Translating between a resource-rich source language and an under-resourced target language
- Bootstrap an ASR system without any linguistic knowledge of the target language

Applications

http://en.wikipedia.org/wiki/File:Akha_laos_11_03d.jpg

Speech processing for non-written and under-resourced languages



Dialects

Scenario

Say “I am sick.” in your mother tongue.

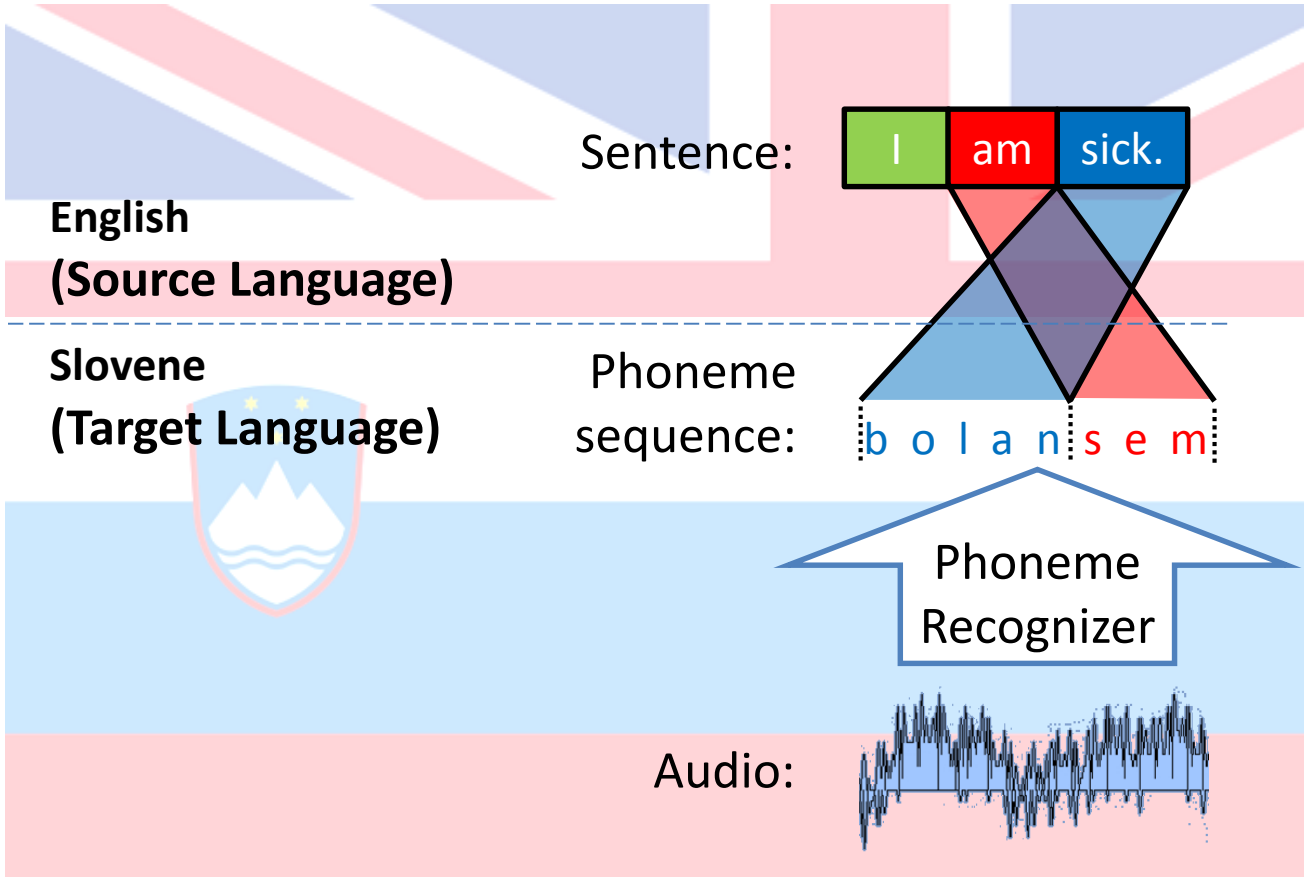
/b/ /o/ /l/ /a/ /n/ /s/ /e/ /m/

Say “I am healthy.” in your mother tongue.

/z/ /d/ /r/ /a/ /v/ /s/ /e/ /m/

- ***/b/ /o/ /l/ /a/ /n/*** seems to be a word (meaning **sick**)
- ***/z/ /d/ /r/ /a/ /v/*** seems to be a word (meaning **healthy**)
- ***/s/ /e/ /m/*** seems to be a word (meaning **I am**)

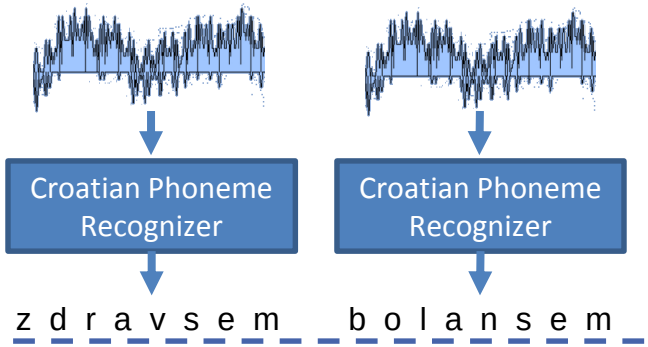
Word-to-Phoneme Alignments



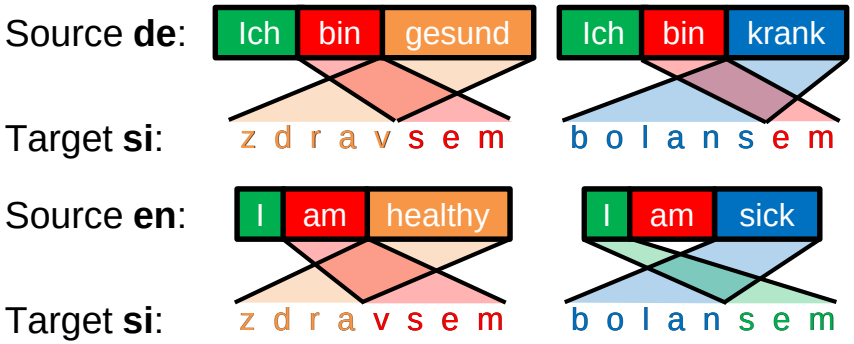
(Stahlberg et. al., 2012)
(Stüker and Besacier, 2009)
(Stüker and Waibel, 2008)
(Besacier et. al., 2006)

Pronunciation Extraction

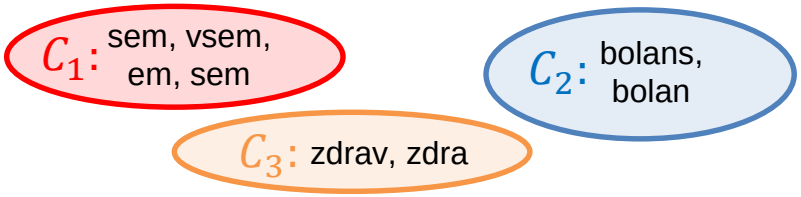
Step 1: Phoneme Recognition



Step 2: Alignment



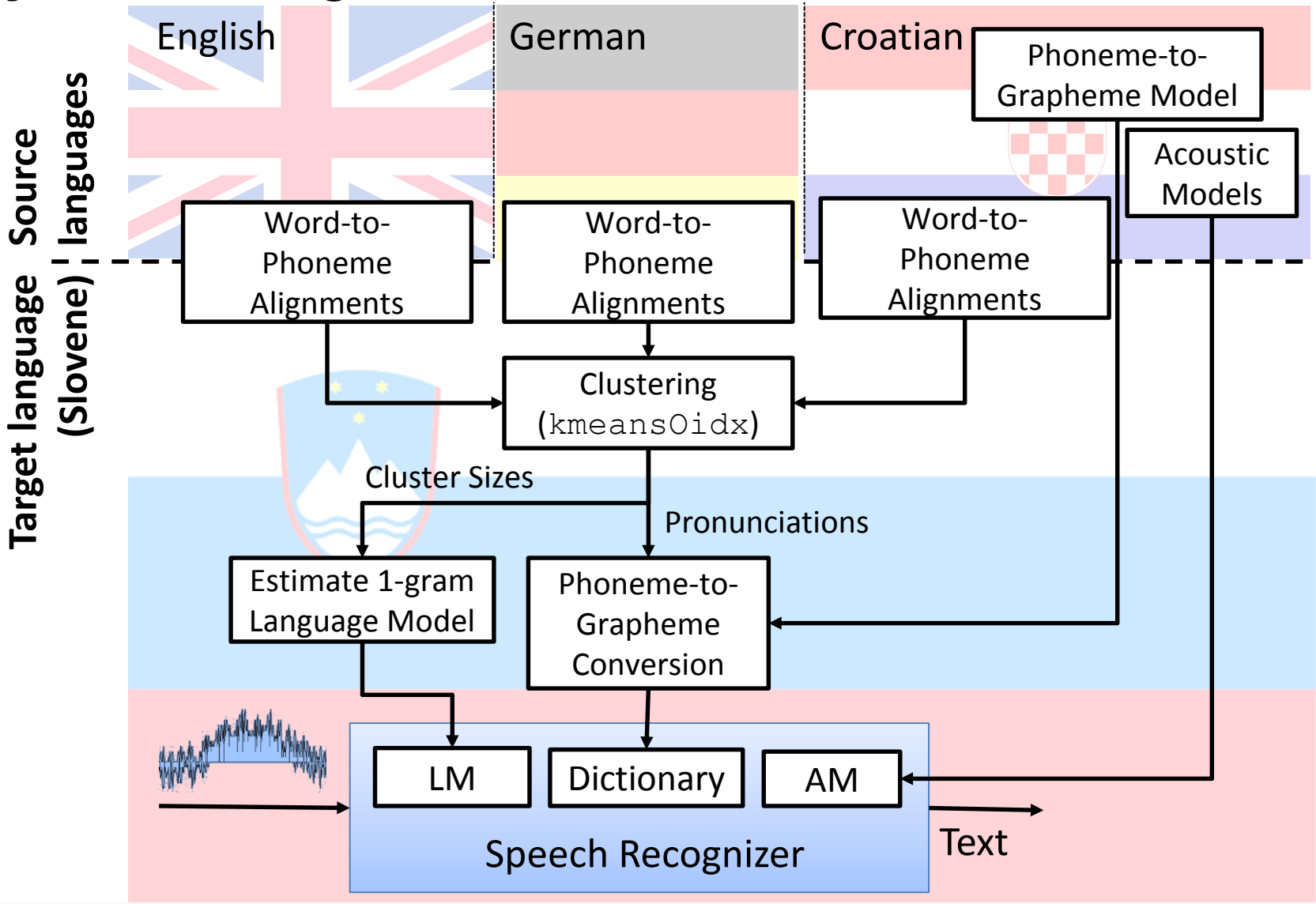
Step 3: Clustering



Step 4: Dictionary Generation

i	Written Form: $\mathrm{p2g}_{\mathrm{hr}}(\mu(C_i))$	Pronunciation: $\mu(C_i)$
1	sem	s e m
2	bolan	b o l a n
3	zdrav	z d r a v

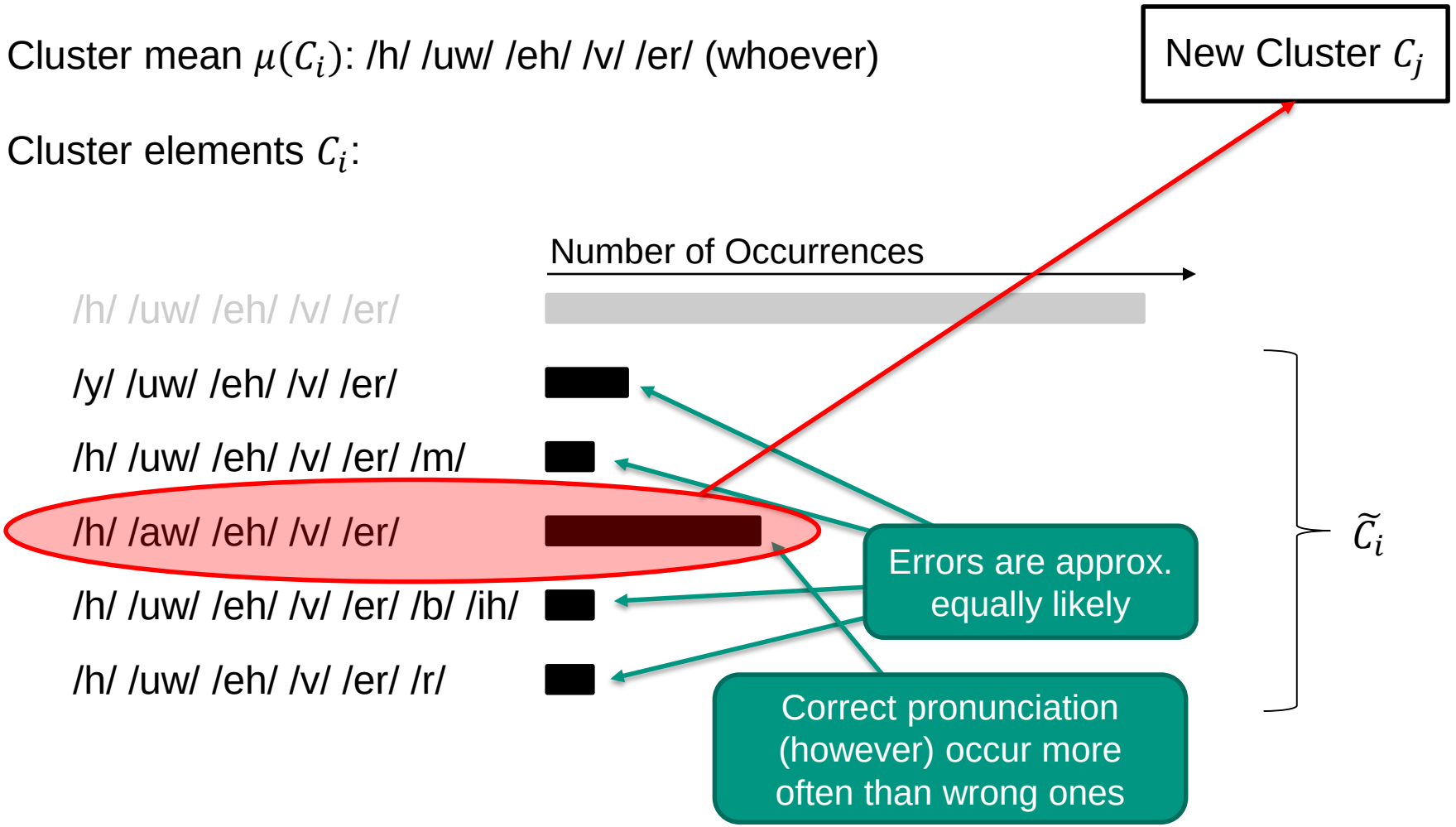
System Design



Issues with *k*-means Clustering

Cluster mean $\mu(C_i)$: /h/ /uw/ /eh/ /v/ /er/ (whoever)

Cluster elements C_i:

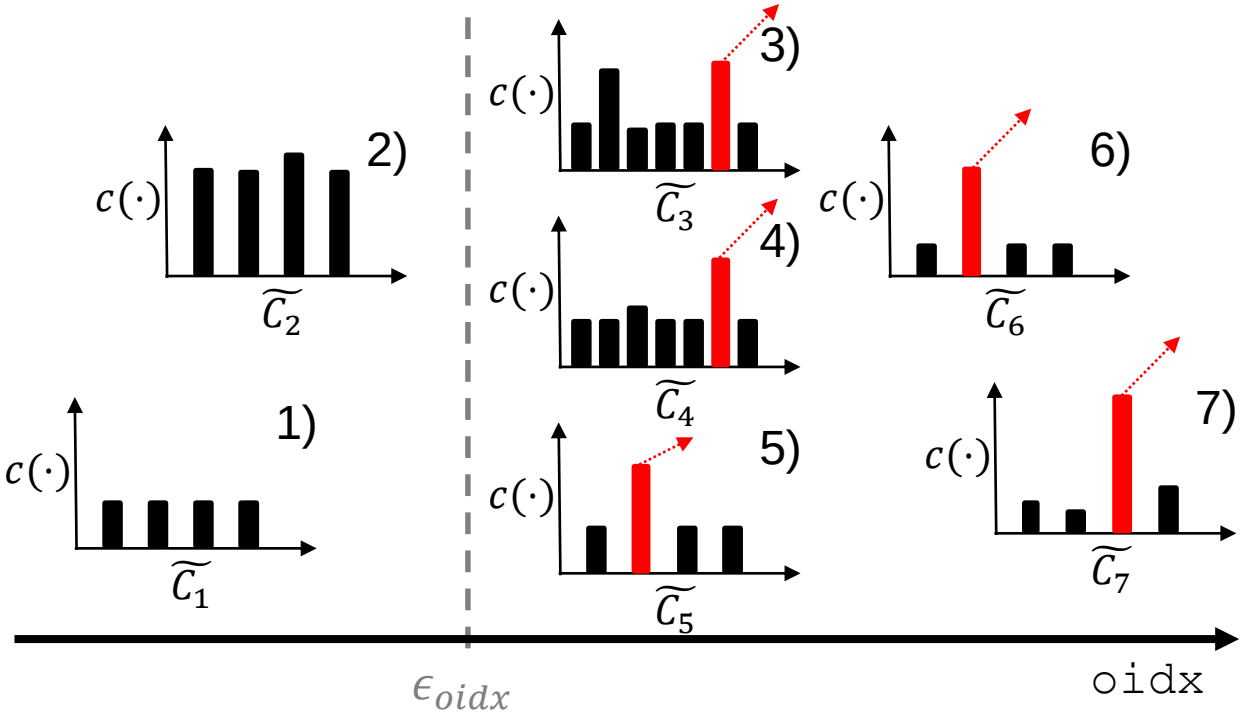


kmeansOidx Clustering Algorithm

$$\widetilde{C_i} := \{p \in C_i | p \neq \mu(C_i)\}$$

$$\texttt{oidx} : C_i \mapsto \begin{cases} 1 & \text{if } \widetilde{C_i} = \emptyset \\ \frac{\max_{p \in \widetilde{C_i}} c(p)}{\text{Median}(\{c(p) | p \in \widetilde{C_i}\})} & \text{otherwise} \end{cases}$$

C_i:	Set of elements in cluster i.
$\mu(C_i)$:	Mean of cluster i.
$c(p)$:	Number of occurrences of element p.



Basic Medical Expression Database (BMED)

	Vocabulary Size	Avg. Word Frequency	Avg. Sentence Length	Speakers	Audio
Croatian (hr)	280 words	3.19	4.47 words	8	96 min.
English (en)	163 words	6.90	5.62 words	-	-
German (de)	184 words	6.11	5.62 words	-	-
Slovene (si)	279 words	3.24	4.50 words	5	50 min.

Error-Free Phonetic Transcriptions (0% Phoneme Error Rate)

Method from (Stahlberg et. al., 2013):

Source Language	Character Error Rate (0-gram)
de	66.0%
en	62.4%
hr	51.3%

New Method `kmeansOidx`:

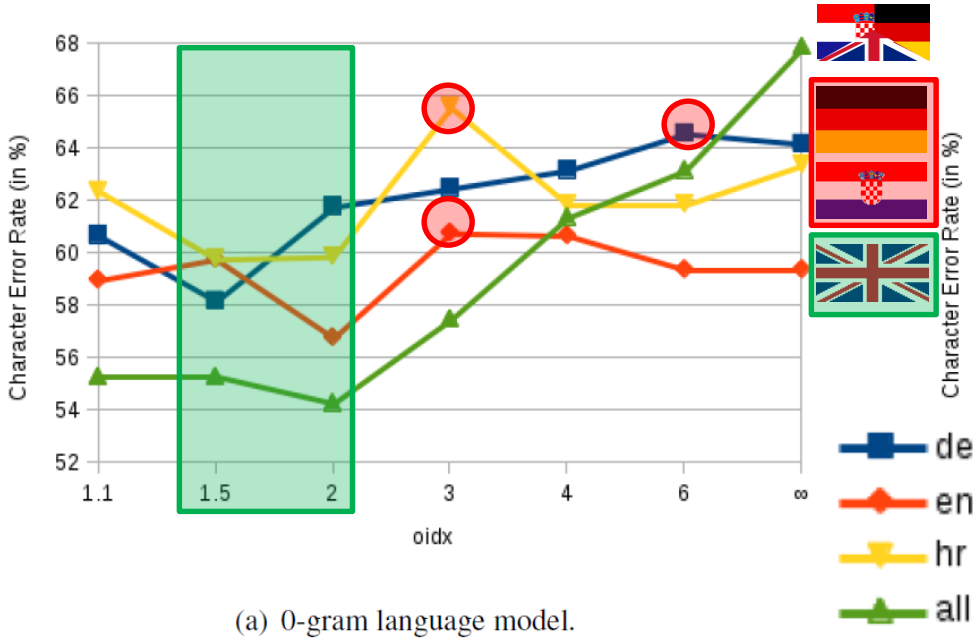
Src. Lang.	ϵ_{oidx}	CER (0-gram)	CER (1-gram)
de	∞	52.1 %	50.7%
en		51.4%	47.9%
hr		49.9%	48.4%
all		47.3%	46.1%
	3	46.6%	46.2%
	2	45.9%	45.1%
	1.5	45.8%	44.4%
	1	45.2%	44.2%

- Outperforms (Stahlberg et. al., 2013)
- 1-gram language model helps
- Source language combination helps
- Setting ϵ_{oidx} helps

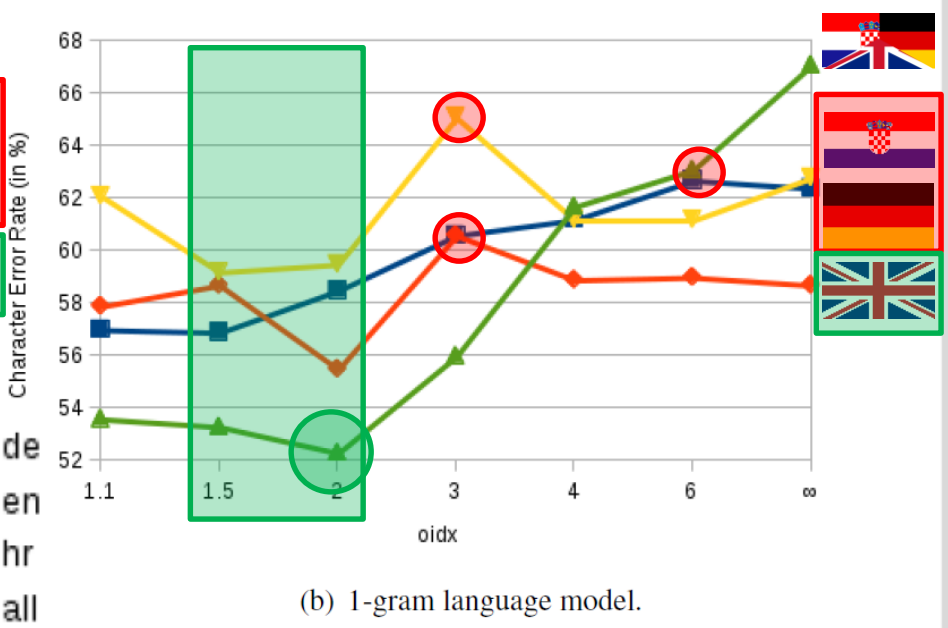
Slovene ASR with Croatian AM (Gold Standard):

Language Model	WER	CER
0-gram	36.2%	**15.7%**
1-gram	32.0%	**13.6%**

Recognized Phonetic Transcriptions (55.2% Phoneme Error Rate)



(a) 0-gram language model.



(b) 1-gram language model.

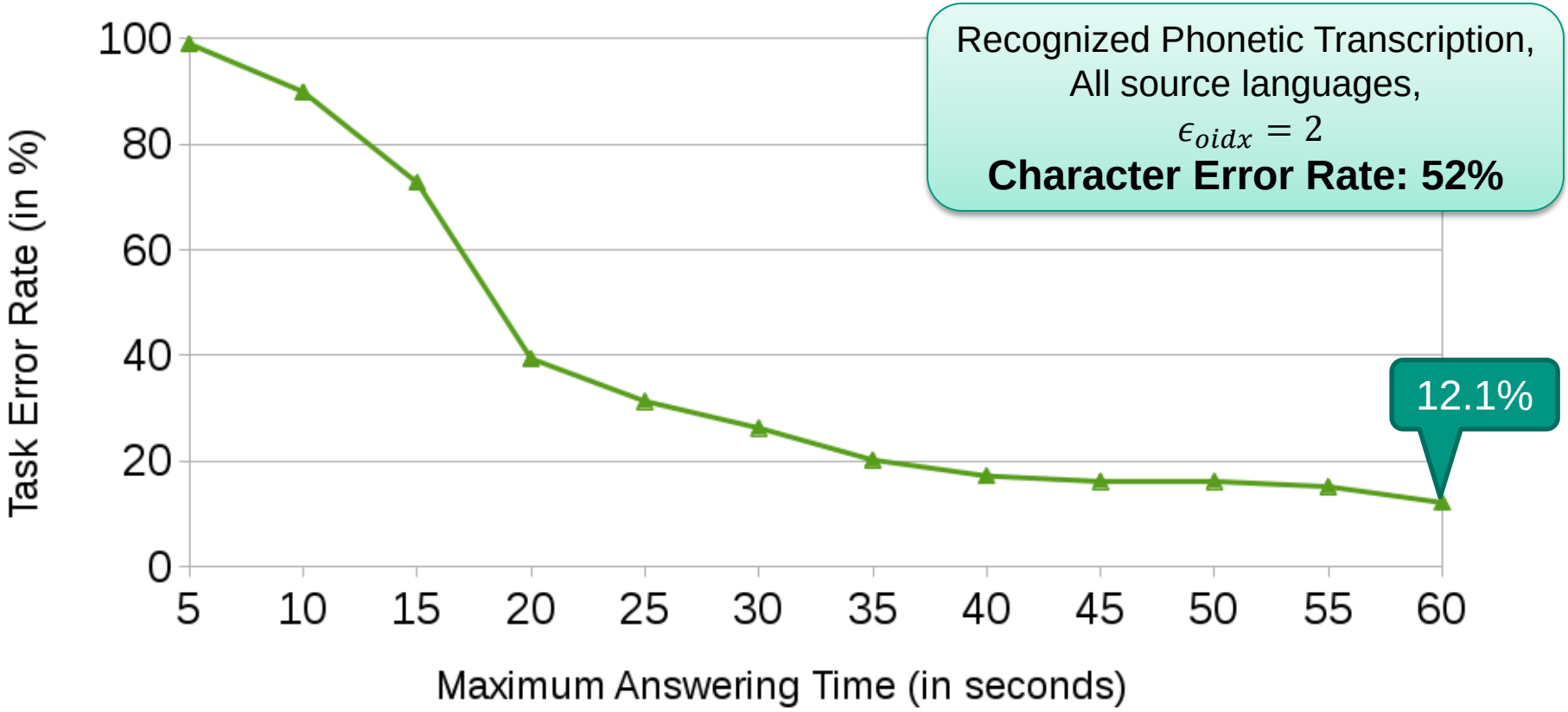
- English best single source language
- Fluctuations with single source languages.
- All minima between $\epsilon_{oidx} = 1.5$ and 2.
- Best system: 52% Character Error Rate

Slovene ASR with Croatian AM (Gold Standard):

Language Model	WER	CER
0-gram	36.2%	**15.7%**
1-gram	32.0%	**13.6%**

Human Evaluation

- Task: Select correct sentence from the 200 BMED sentences, given the recognizer transcript



Summary

- Speech recognition for non-written and under-resourced languages or dialects
  - Only using spoken translations from other languages and resources from resource-rich languages
  - No given pronunciation dictionary, transcribed speech and text resources in the target language
- Target Language: Slovene
  Source Languages: Croatian, English, German
  - 200 sentences, limited domain
- Best system: 52% Character Error Rate
  - Task Error Rate for selecting 1 of 200 sentences: 12.1%