Karlsruhe Institute of Technology

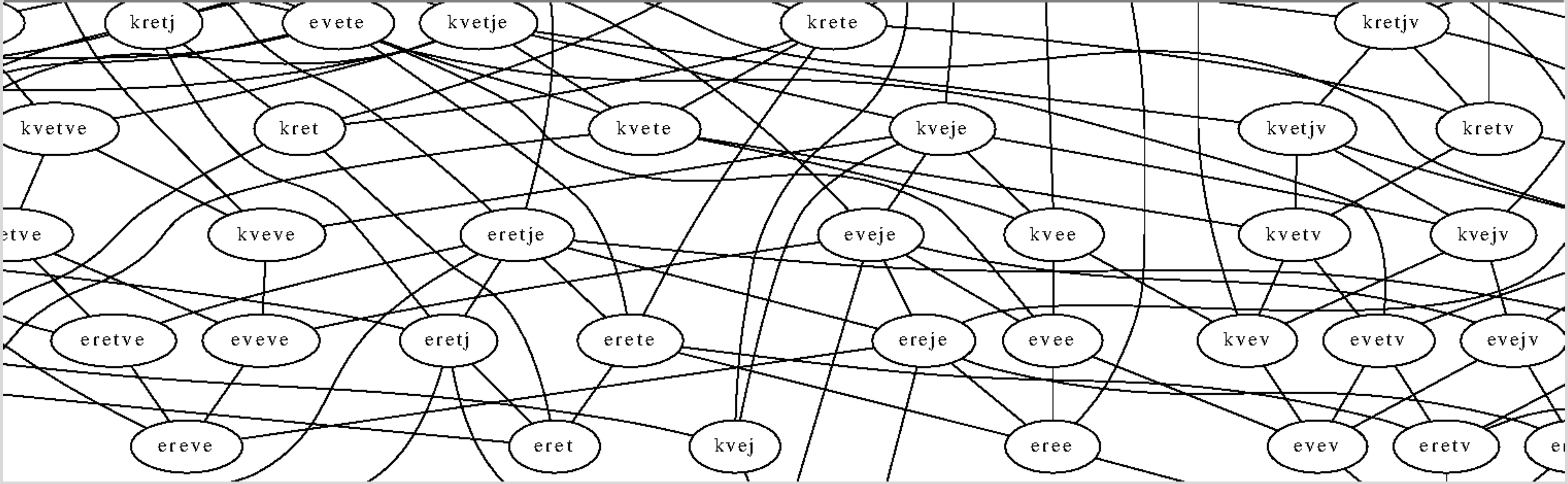
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

Cognitive Systems Lab

Pronunciation Extraction Through Cross-Lingual Word-to-Phoneme Alignment

Felix Stahlberg, **Tim Schlippe**, Stephan Vogel, Tanja Schultz

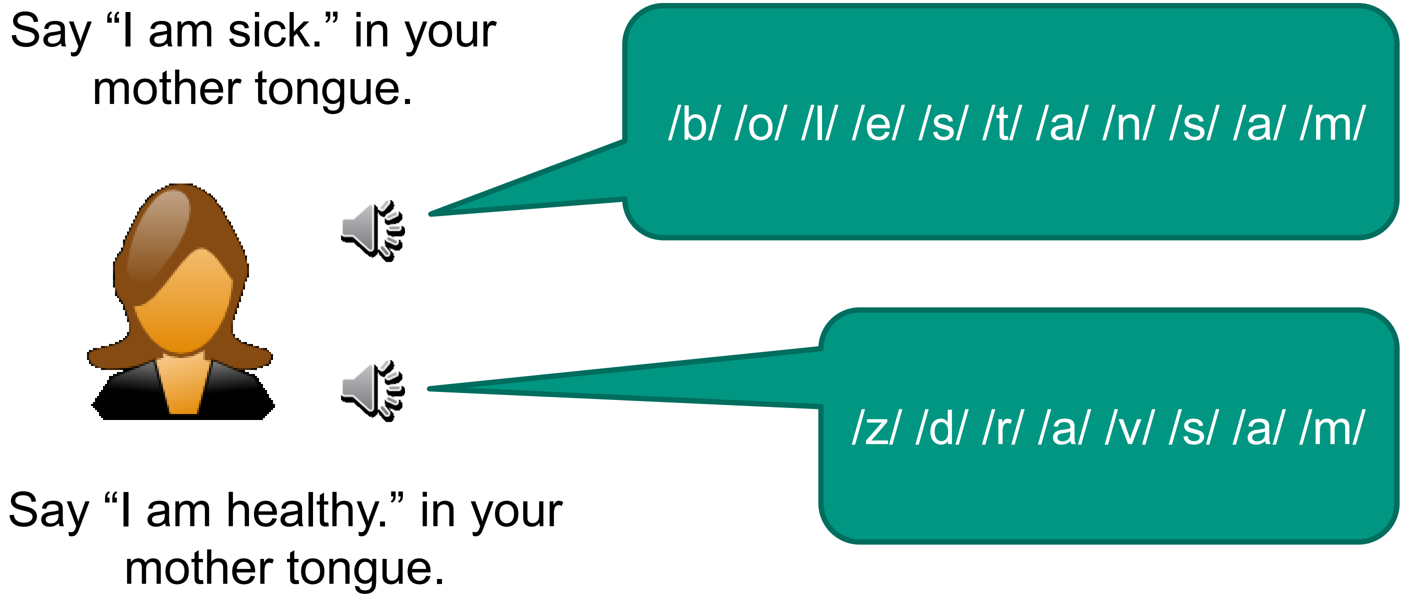
SLSP 2013 – 1st International Conference on Statistical Language and Speech Processing
Tarragona, Spain



Outline

1. Motivation
2. Word Segmentation
3. Word Pronunciation Extraction
4. Experiments
    1. Corpus
    2. Evaluation Measures
    3. Which Translation Is Favorable?
    4. Combining Multiple Translations
    5. Analysis of the Results – Common errors
5. Conclusion and Future Work

Scenario



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

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

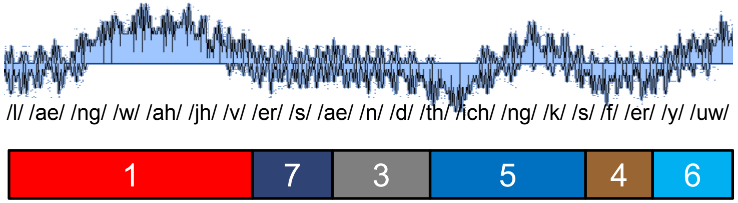
Say “I am healthy.” in your mother tongue.

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

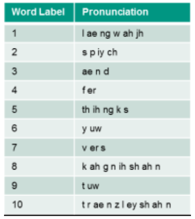
- ***/s/ /a/ /m/*** seems to be a word (meaning **I am**)
- ***/b/ /o/ /l/ /e/ /s/ /t/ /a/ /n/*** seems to be a word (meaning **sick**)
- ***/z/ /d/ /r/ /a/ /v/*** seems to be a word (meaning **healthy**)

Long Term Goal

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

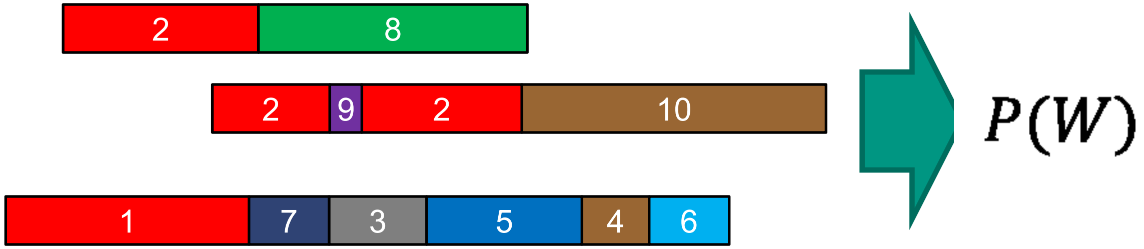

/l/ /ae/ /ng/ /w/ /ah/ /jh/ /v/ /er/ /s/ /ae/ /n/ /d/ /th/ /ich/ /ng/ /k/ /s/ /f/ /er/ /y/ /uw/
1
7
3
5
4
6

  - 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



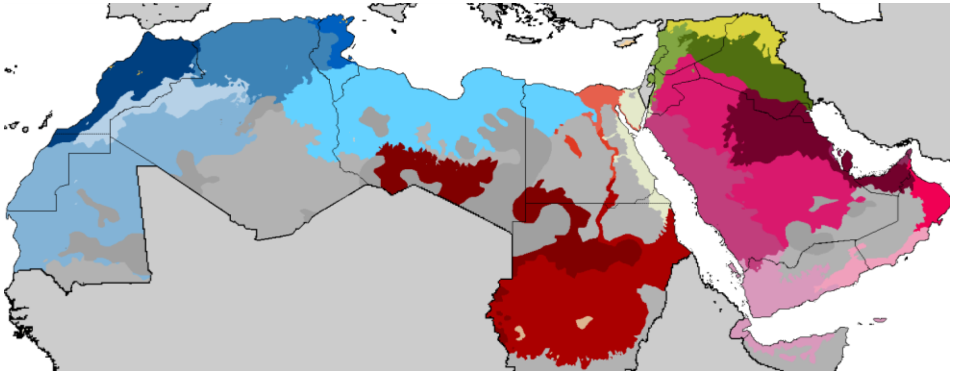
$P(W)$

Train ASR System (future work)

Applications

http://www.fotopedia.com/items/_avPIZmqM3w-6716j3F1J-U

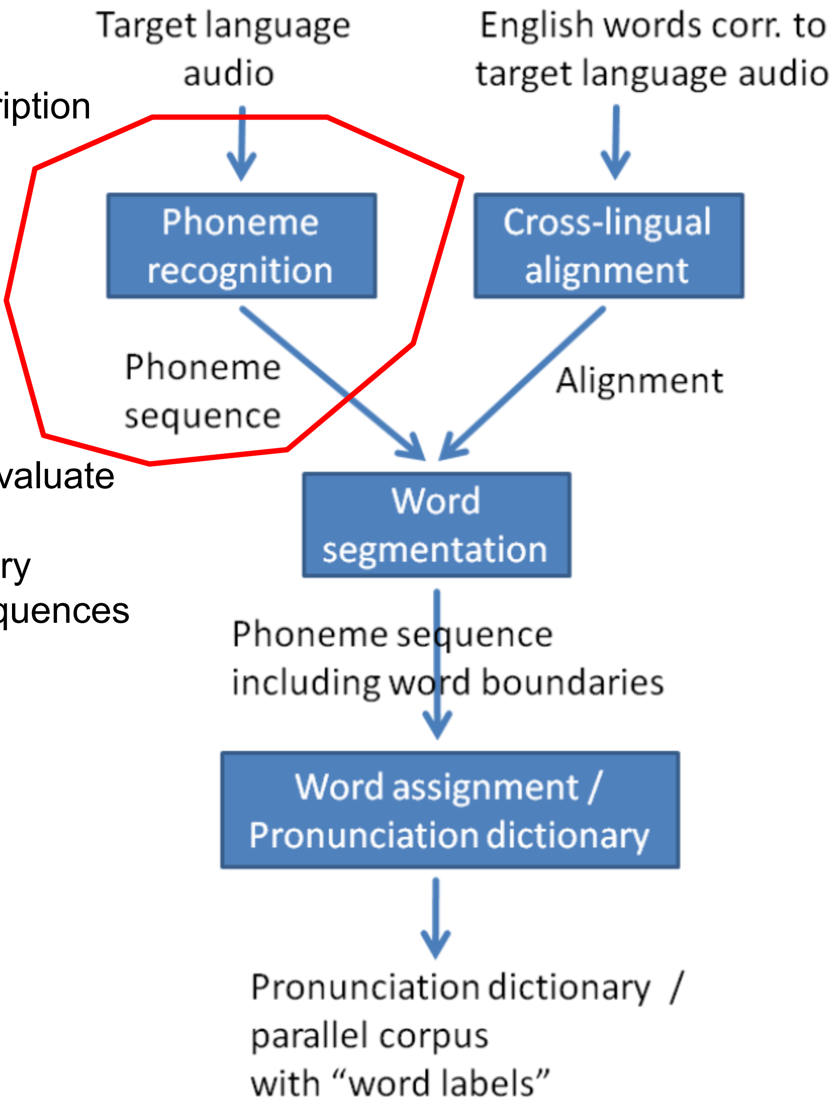
Speech processing for non-written and under-resourced languages



Dialects

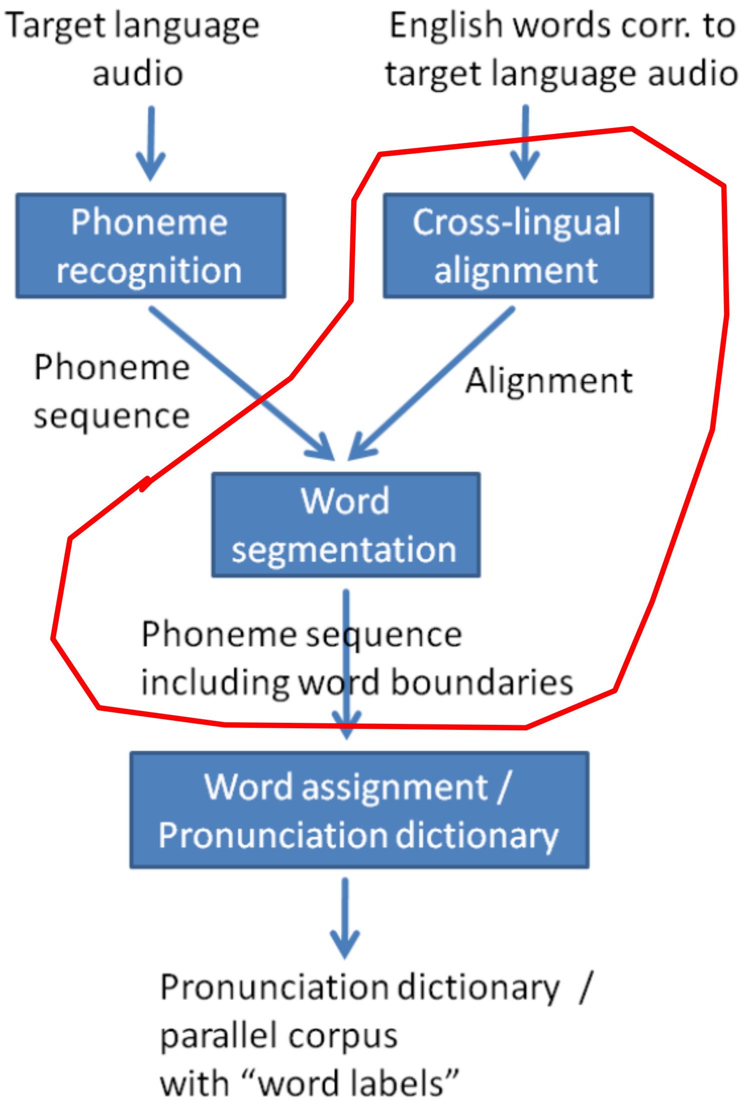
Roadmap

- Need phonetic transcription of what is said
- Usually phoneme recognizer
- In this work: Perfect phonetic transcriptions
- Focus to define and evaluate steps for extracting a pronunciation dictionary from the phoneme sequences


Target language audio
English words corr. to target language audio
Phoneme recognition
Cross-lingual alignment
Phoneme sequence
Alignment
Word segmentation
Phoneme sequence including word boundaries
Word assignment / Pronunciation dictionary
Pronunciation dictionary / parallel corpus with “word labels”

Roadmap

- How can we find word boundaries and segment phoneme sequences into word units?
- Inproved segmentation with cross-lingual information
- Alignment between word units in written translation and phoneme sequences of target language

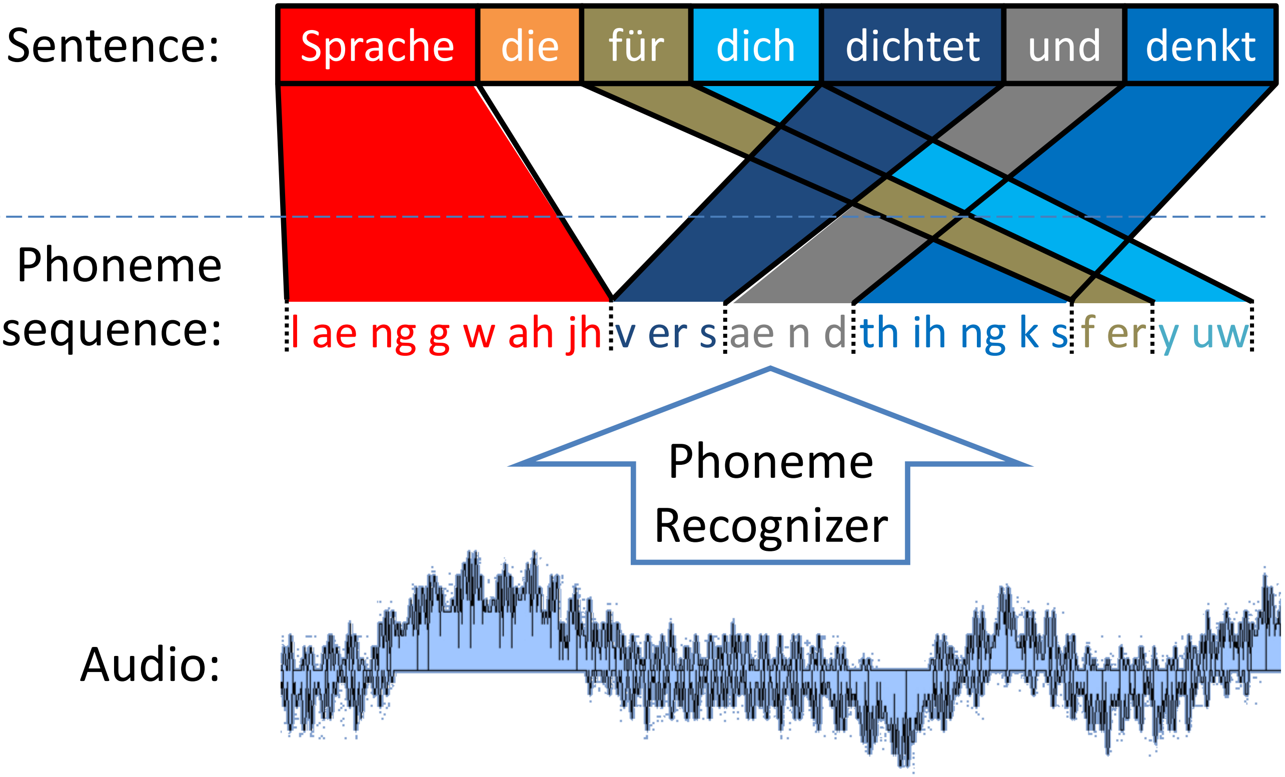


Target language audio
English words corr. to target language audio
Phoneme recognition
Cross-lingual alignment
Phoneme sequence
Alignment
Word segmentation
Phoneme sequence including word boundaries
Word assignment / Pronunciation dictionary
Pronunciation dictionary / parallel corpus with “word labels”

Word-Segmentation – Word-to-Phoneme Alignments

German (Source Language)

English (Target Language)

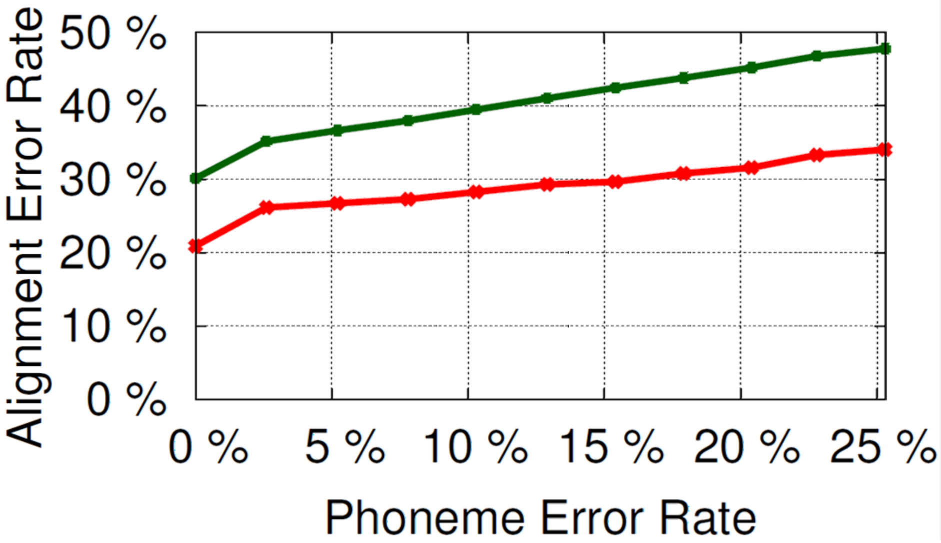


(Besacier et. al., 2006)
(Stüker and Waibel, 2008)
(Stüker and Besacier, 2009)
(Stahlberg et. al., 2012)

Word-Segmentation – Results

(Stahlberg et. al., 2012)

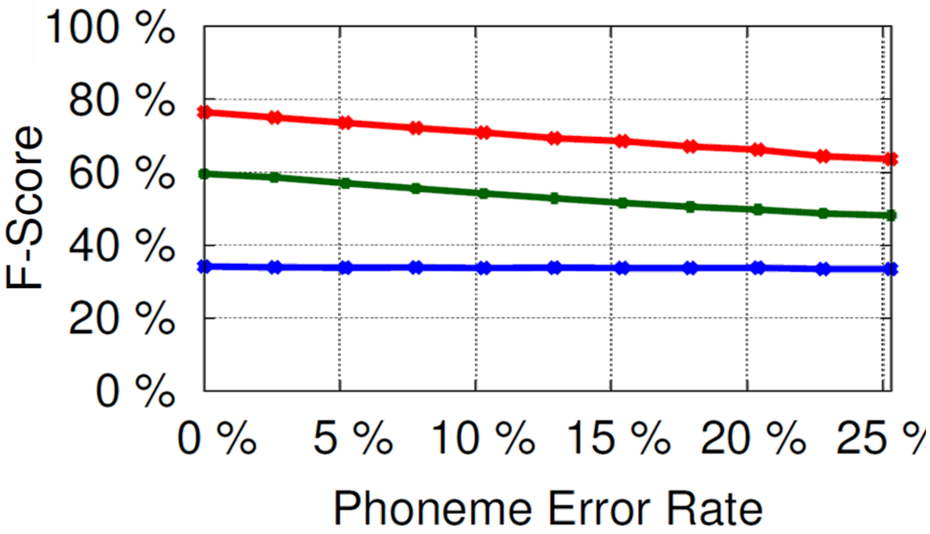
http://code.google.com/p/pisa/

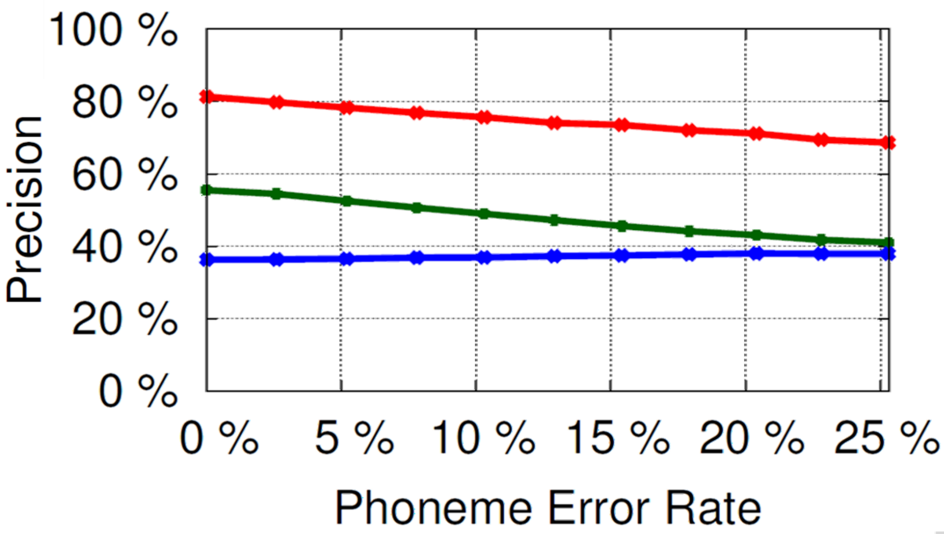


Adaptor Grammars (Monolingual) 🟦

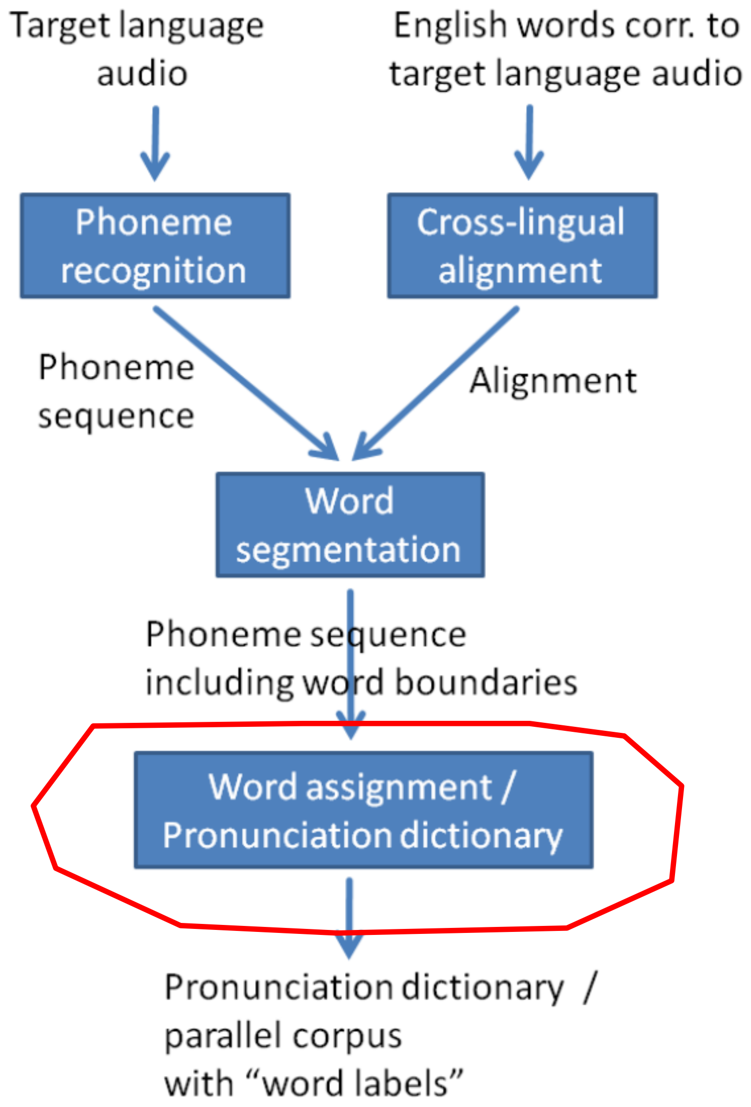
GIZA++ word-to-phoneme alignments 🟩

Model 3P 🟥





Roadmap



Target language audio

English words corr. to target language audio

Phoneme recognition

Cross-lingual alignment

Phoneme sequence

Alignment

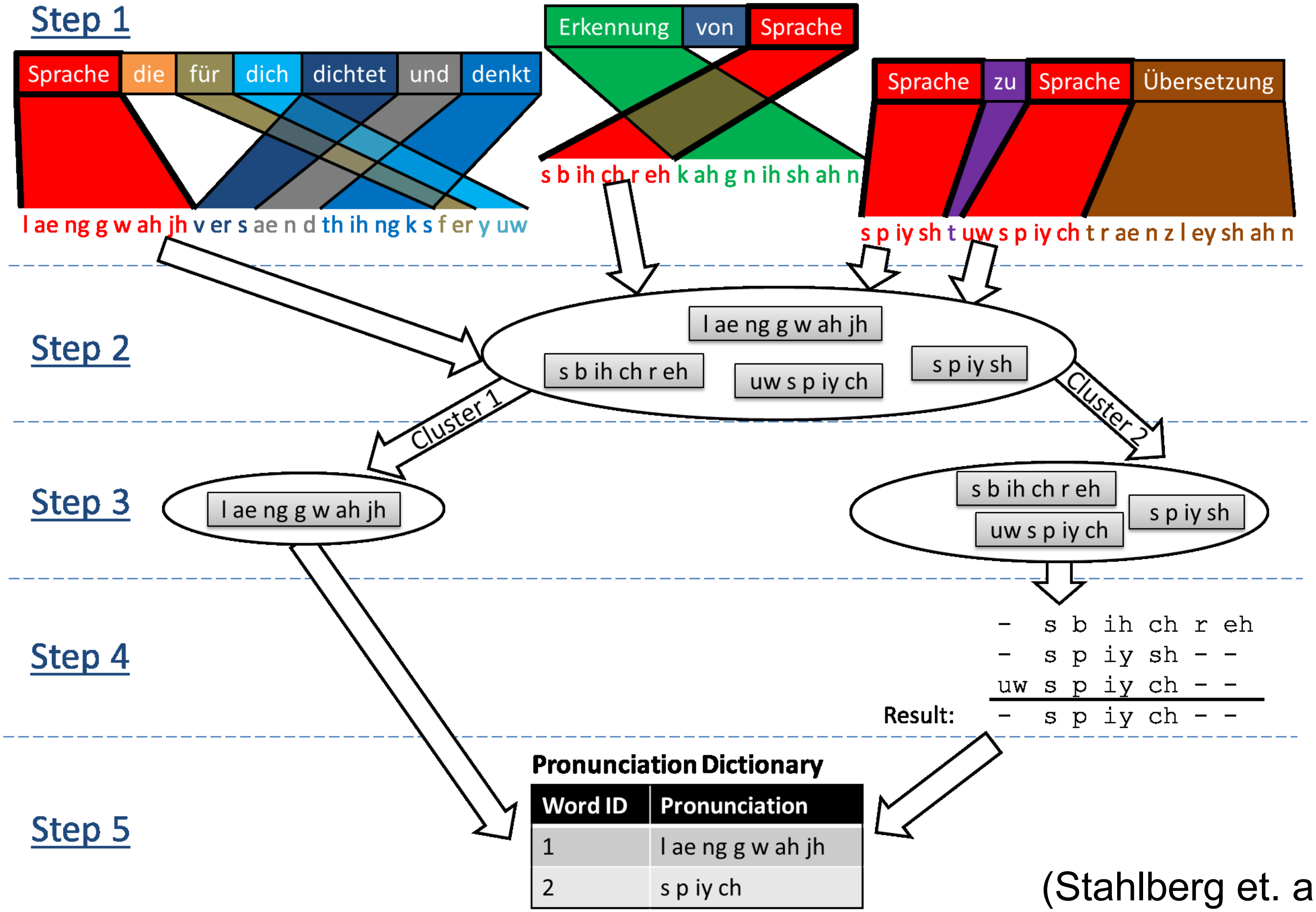
Word segmentation

Phoneme sequence including word boundaries

Word assignment / Pronunciation dictionary

Pronunciation dictionary / parallel corpus with "word labels"

Word-Pronunciation Extraction



```
         -   s  b  ih  ch  r  eh
         -   s  p  iy  sh  -  -
         uw  s  p  iy  ch  -  -
Result:  -   s  p  iy  ch  -  -
```

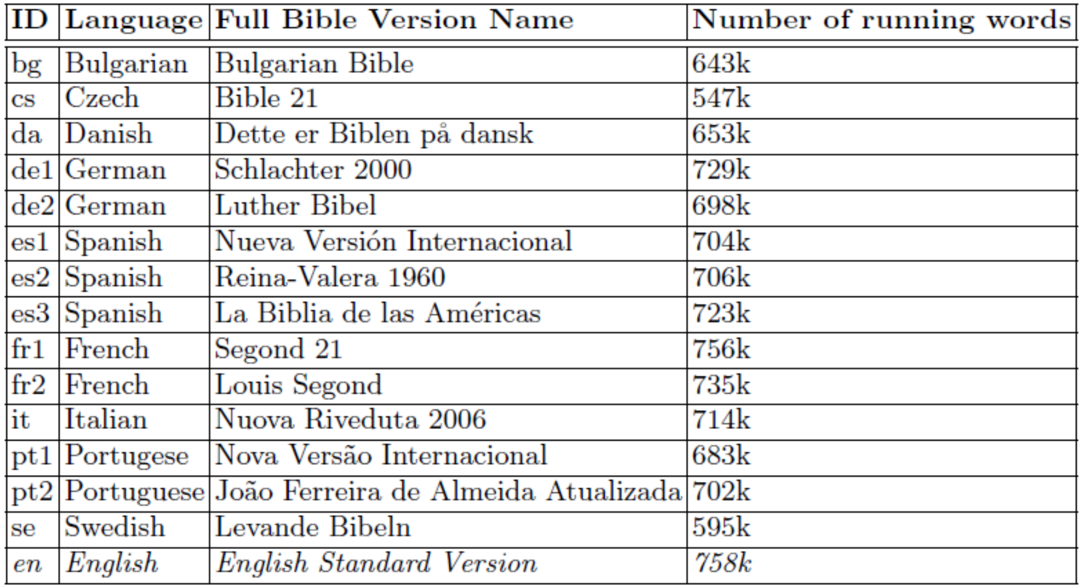
Pronunciation Dictionary

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

(Stahlberg et. al, 2013)

Experiments – Corpus

- Parallel data from the Christian Bible (30.6k verses,14 written translations)
- Variety of linguistic approaches to Bible translation (dynamic equivalence, formal equivalence, and idiomatic translation)
- English as "under-resourced target language"
(deeper insight in strengths and weaknesses of our algorithm) → ESV Bible
- "Perfect phoneme recognizer":
Replaced words in ESV Bible and removed word boundaries

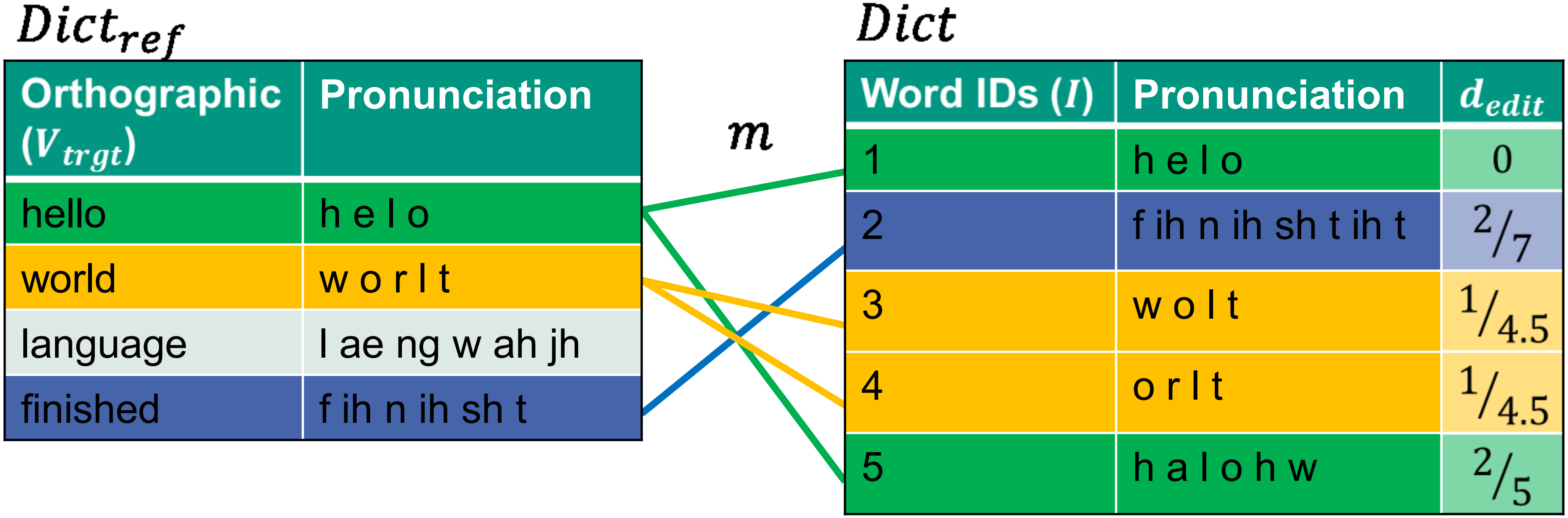


ID	Language	Full Bible Version Name	Number of running words
bg	Bulgarian	Bulgarian Bible	643k
cs	Czech	Bible 21	547k
da	Danish	Dette er Biblen på dansk	653k
de1	German	Schlachter 2000	729k
de2	German	Luther Bibel	698k
es1	Spanish	Nueva Versión Internacional	704k
es2	Spanish	Reina-Valera 1960	706k
es3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

Evaluation Measures (1)

$$m(n) = \arg\min_{v \in V_{trgt}} d_{edit}(Dict(n), Dict_{ref}(v))$$

$n \in I$ = Set of word IDs



$Dict_{ref}$

Orthographic (V_{trgt})	Pronunciation
hello	h e l o
world	w o r l t
language	l ae ng w ah jh
finished	f ih n ih sh t

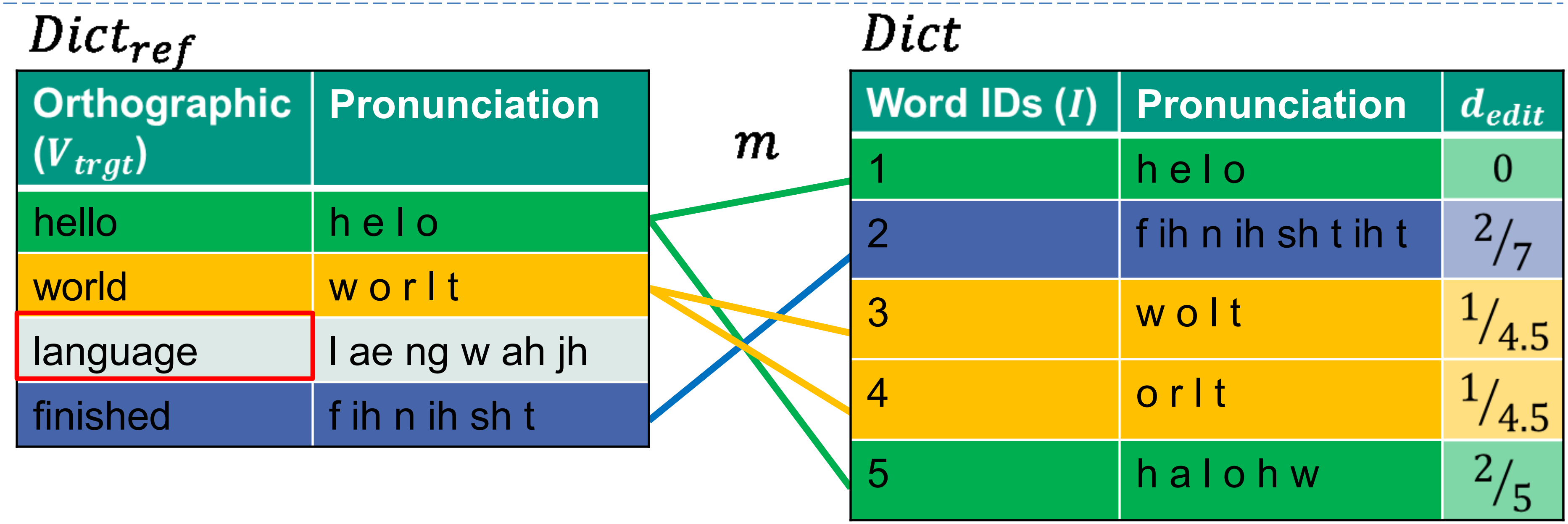
$Dict$

Word IDs (I)	Pronunciation	d_{edit}
1	h e l o	0
2	f ih n ih sh t ih t	$^{2}/_{7}$
3	w o l t	$^{1}/_{4.5}$
4	o r l t	$^{1}/_{4.5}$
5	h a l o h w	$^{2}/_{5}$

Evaluation Measures (2)

Out-Of-Vocabulary Rate (OOV-Rate)

- Calculated on a subset of the English Bible using the set of matched vocabulary entries $m(I)$



$Dict_{ref}$

Orthographic (V_{trgt})	Pronunciation
hello	h e l o
world	w o r l t
language	l ae ng w ah jh
finished	f ih n ih sh t

$Dict$

Word IDs (I)	Pronunciation	d_{edit}
1	h e l o	0
2	f ih n ih sh t ih t	$^2/_7$
3	w o l t	$^1/_{4.5}$
4	o r l t	$^1/_{4.5}$
5	h a l o h w	$^2/_5$

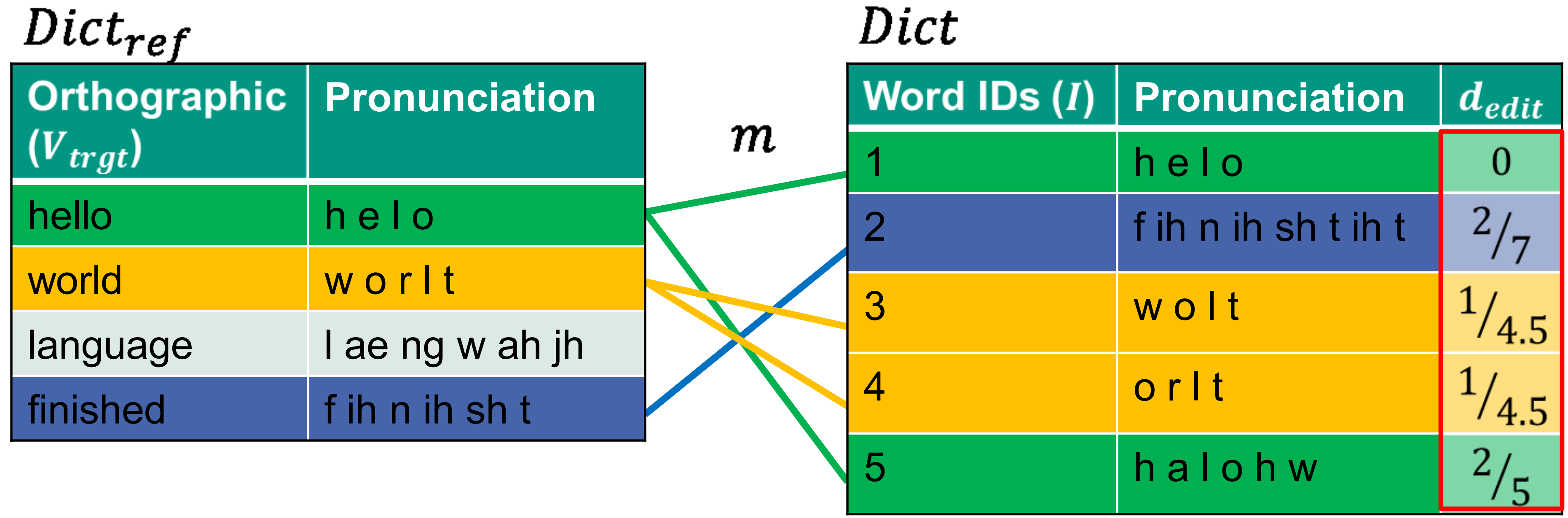
$$m(n) = \arg\min_{v \in V_{trgt}} d_{edit}(Dict(n), Dict_{ref}(v))$$

$n \in I =$ Set of word IDs

Evaluation Measures (3)

Phoneme Error Rate (PER)

$$PER = \frac{\sum_{n \in I} d_{edit}(Dict(n), Dict_{ref}(m(n)))}{|I|}$$



Orthographic (V_{trgt})	Pronunciation
hello	h e l o
world	w o r l t
language	l ae ng w ah jh
finished	f ih n ih sh t

Word IDs (I)	Pronunciation	d_{edit}
1	h e l o	0
2	f ih n ih sh t ih t	$^2/_7$
3	w o l t	$^1/_{4.5}$
4	o r l t	$^1/_{4.5}$
5	h a l o h w	$^2/_5$

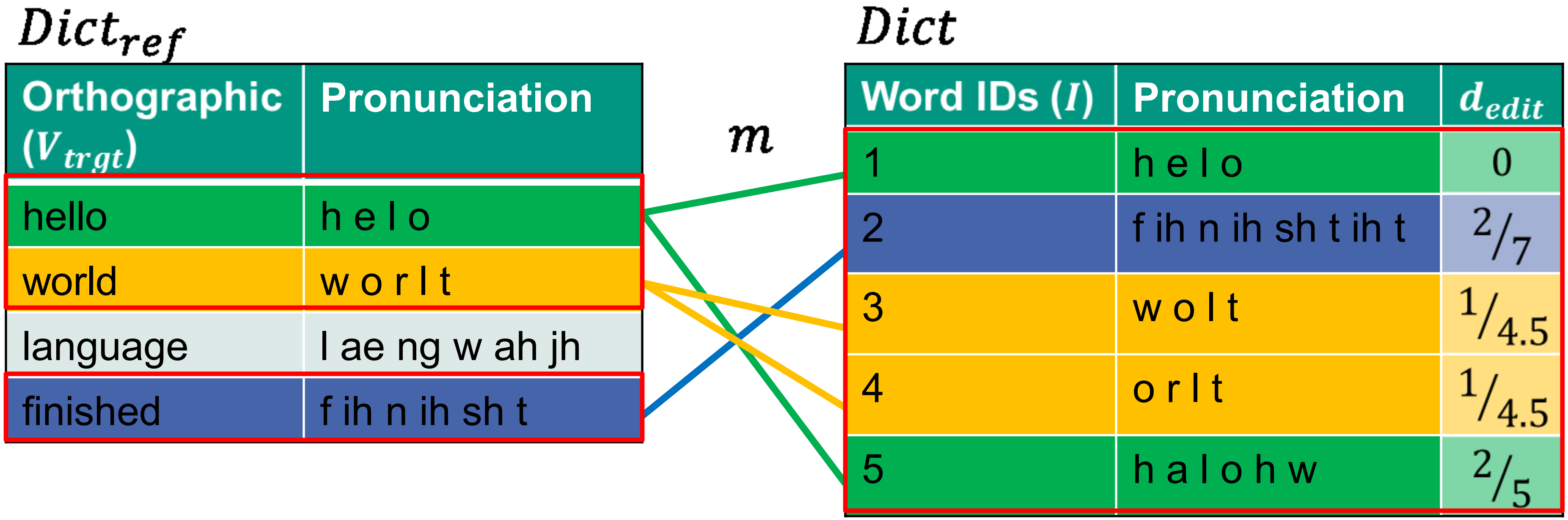
$$m(n) = \arg\min_{v \in V_{trgt}} d_{edit}(Dict(n), Dict_{ref}(v))$$

$n \in I$ = Set of word IDs

Evaluation Measures (4)

Hypo/Ref Ratio

$$HypoRefRatio = \frac{|I|}{|m(I)|}$$



$Dict_{ref}$

Orthographic (V_{trgt})	Pronunciation
hello	h e l o
world	w o r l t
language	l ae ng w ah jh
finished	f ih n ih sh t

m

$Dict$

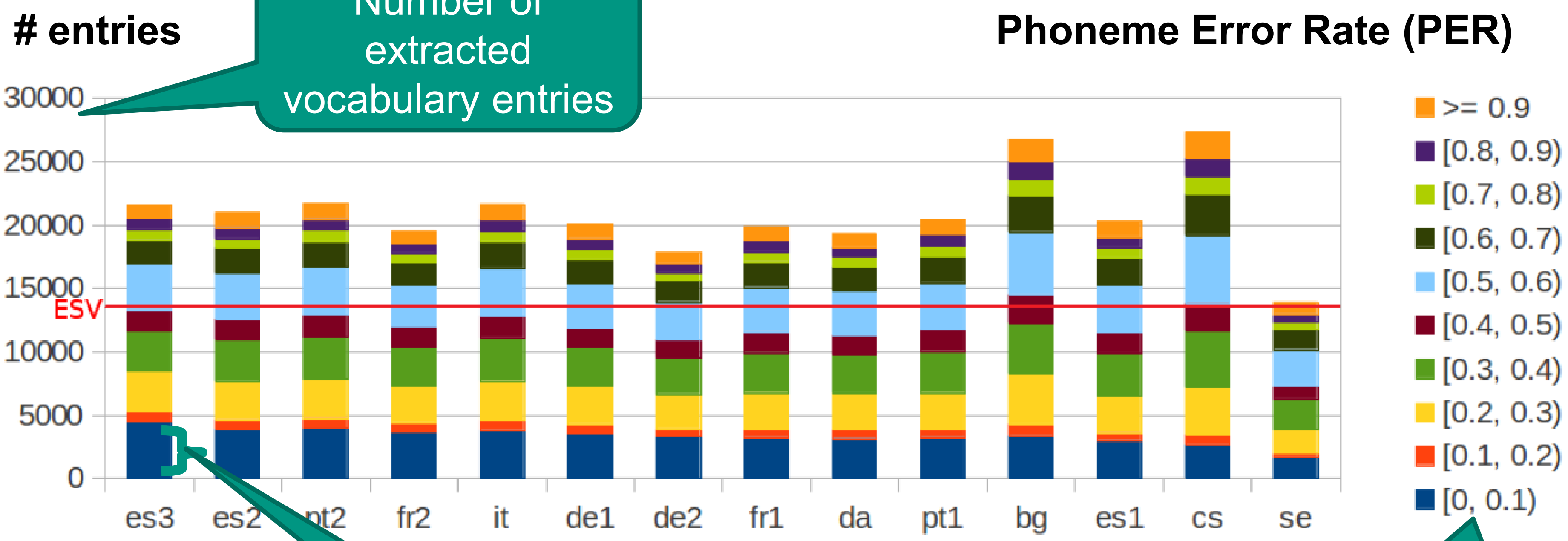
Word IDs (I)	Pronunciation	d_{edit}
1	h e l o	0
2	f ih n ih sh t ih t	$^{2}/_{7}$
3	w o l t	$^{1}/_{4.5}$
4	o r l t	$^{1}/_{4.5}$
5	h a l o h w	$^{2}/_{5}$

$$m(n) = \arg\min_{v \in V_{trgt}} d_{edit}(Dict(n), Dict_{ref}(v))$$

$n \in I =$ Set of word IDs

Which Translation Is Favorable? – Distribution of edit distances

Cognitive Systems Lab



Number of extracted vocabulary entries

Distribution of the edit distances between the extracted pronunciations and the nearest entry
in the reference dictionary for all 14 source translations

Number of extracted vocabulary entries close to real target language words (<0.1 edit distance)

Edit distances of extracted vocabulary entries to the next reference vocabulary entry

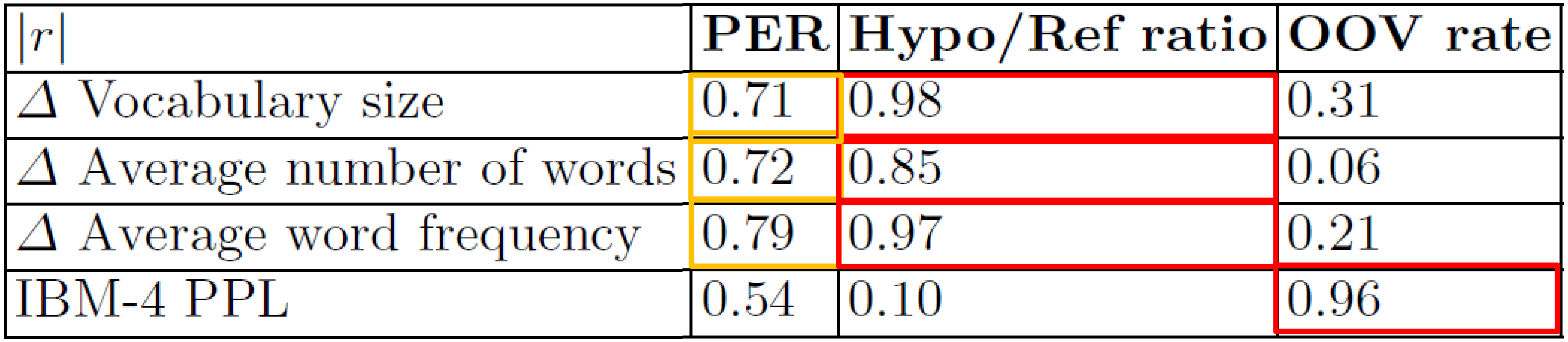
Karlsruhe Institute of Technology

Which Translation Is Favorable? – Impact of 4 factors to our evaluation measures

- **∆ vocabulary size:**
  Difference between vocabulary size of the source translation and size of the ESV Bible
- **∆ average number of words per verse:**
  Difference between average verse length in the source translation and in the ESV Bible
- **∆ average word frequency:**
  Difference between the average number of word repetitions in the source translation and in the ESV Bible
- **IBM-4 PPL:**
  To measure the general correspondence of the translation to IBM-Model based alignment models, we run GIZA++ with default configuration at the word level and use the final perplexity of IBM-Model 4

Which Translation Is Favorable? – Correlation of evaluation measures

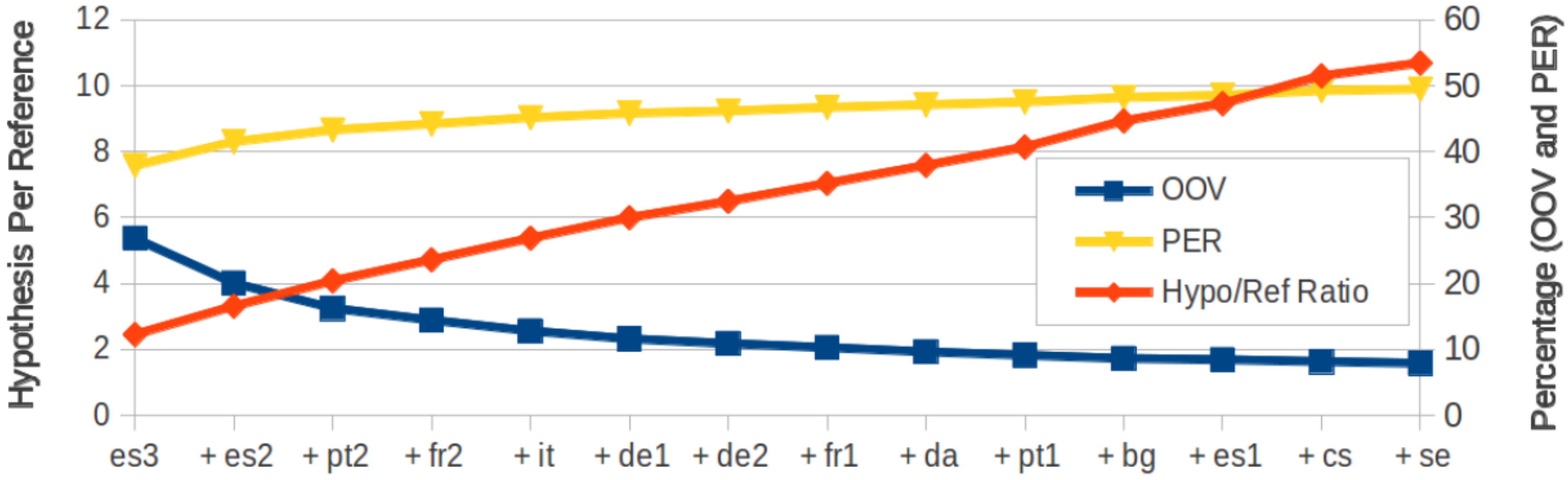
High $0 \leq |r| \leq 1$ – high linear correlation



$\|r\|$	**PER**	**Hypo/Ref ratio**	**OOV rate**
Δ Vocabulary size	0.71	0.98	0.31
Δ Average number of words	0.72	0.85	0.06
Δ Average word frequency	0.79	0.97	0.21
IBM-4 PPL	0.54	0.10	0.96

Combining multiple translations

- Concatenate pronunciations and remove homophones



Evaluation measures over the number of combined source translations

- Combining all 14 translations results in a dictionary with only 7.9% OOV rate,
- But more than 9 of 10 dictionary entries are extracted unnecessarily (Hypo/Ref ratio 10.7:1)

Common Errors (1)

- Off-by-one alignment errors

Extracted (incorrectly)	Correct
z f ih s t s	f ih s t s (fists)
ih k s t	**f** ih k s t (fixed)
ih z r ey l **ah**	ih z r ey l (israel)

➡ Context information may be helpful

Common Errors (2)

- Different words with the same stem are merged together

Extracted (incorrectly)	Correct
s ih d uw s ih t	s ih d uw s t (seduced) or s ih d uw s i ng (seducing)
ih k n aa l ih jh m	ih k n aa l ih jh (acknowledge) or ih k n aa l ih jh m ah n t (acknowledgement)

➡ Clustering issue

Common Errors (3)

- Missing word boundaries between words often occurring in the same context

Extracted (incorrectly)	Correct
w er ih n d ih g n ah n t	were indignant
f ih n ih sh t ih t	finished it

➡ Cross-lingual information of multiple languages may help

Summary

- Speech processing in non-written and under-resourced languages or dialects
- Cross-lingual information helps to find word boundaries
- Proposed steps for extracting a pronunciation dictionary with word IDs from these segmentations and alignments
- Pronunciation quality is still not good enough for productive use
  - Need better compensation for alignment and phoneme recognition errors when extracting pronunciations
  - Initial approach for combining dictionaries from multiple translations drops OOV rate, but increases number of unnecessary entries

Possible Next Steps

- Iterative extraction
- Better clustering
  - Analysis for different cluster algorithms
  - Add contextual information
- Use information from multiple source languages
- Integrate monolingual word and syllable segmentation
- Real phoneme recognizer
  - How to bootstrap the phoneme recognizer? – maybe multilingual voting and adaptation techniques based on confidence score

¡Muchas gracias!
¡Moltes gràcies!

References

Stahlberg, F., Schlippe, T., Vogel, S., Schultz, T.: Word Segmentation Through Cross-Lingual Word-to-Phoneme Alignment. In: SLT (2012)

Besacier, L., Zhou, B., Gao, Y.: Towards Speech Translation of Non-Written Languages. In: SLT (2006)

Stüker, S., Waibel, A.: Towards Human Translations Guided Language Discovery for ASR Systems. In: SLTU (2008)

Stüker, S., Besacier, L., Waibel, A.: Human Translations Guided Language Discovery for ASR Systems. In: Interspeech (2009)