Karlsruhe Institute of Technology

Cognitive Systems Lab

Methods for Efficient Semi-Automatic Pronunciation Dictionary Bootstrapping

Tim Schlippe, 18 August 2014

Interspeech 2014 – The 15th Annual Conference of the International Speech Communication Association
Singapore

Outline

1. Introduction

2. Our semi-automatic pronunciation generation strategy

3. Experiments
   1. Analysis of intervals for G2P (grapheme-to-phoneme) model updates
   2. Use of single G2P converters
   3. Grapheme-to-phoneme converter combination
   4. Methods for initial grapheme-to-phoneme converter training data

4. Conclusion and future work

Motivation

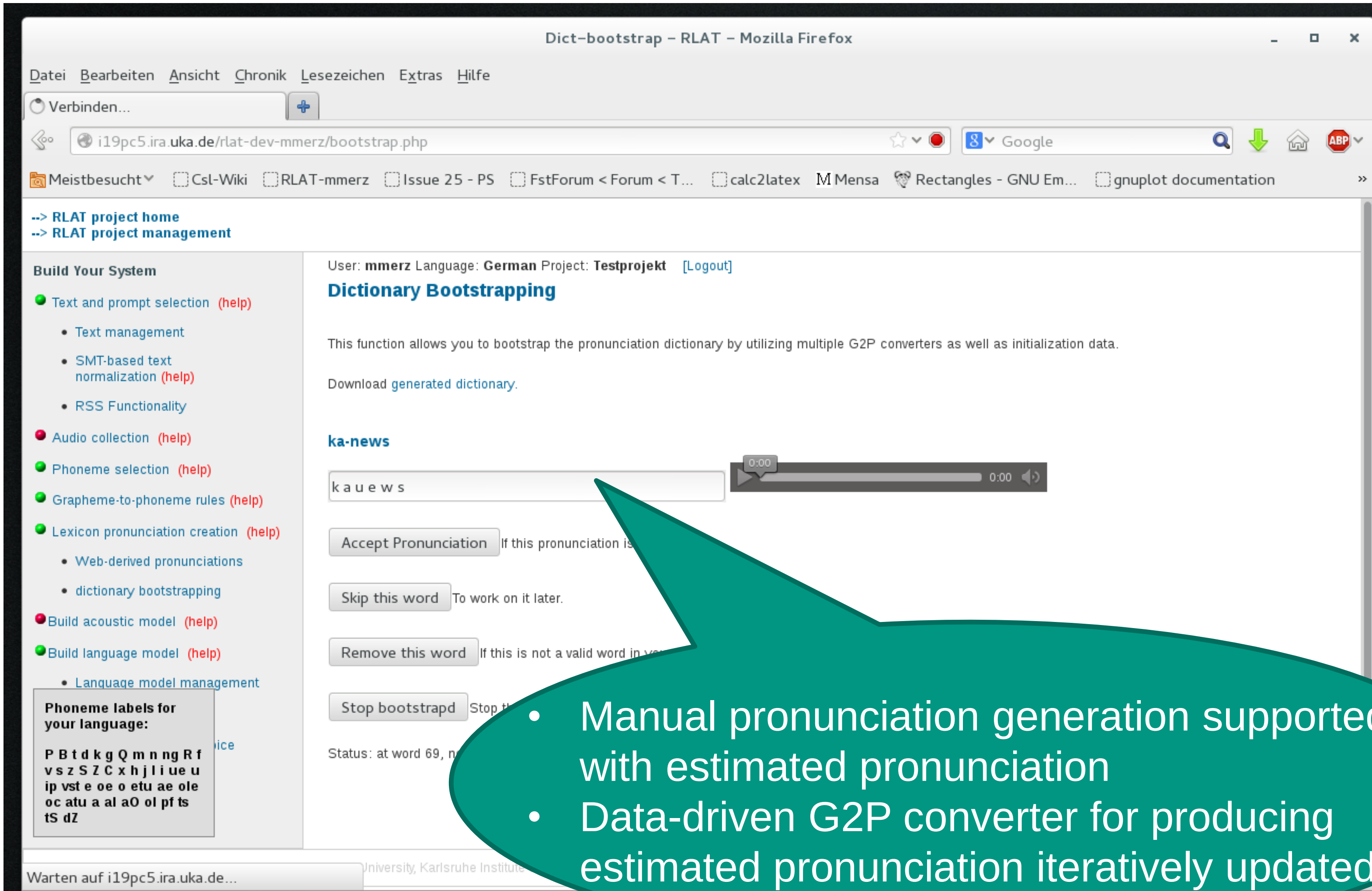
- Pronunciation dictionaries needed for text-to-speech and automatic speech recognition (ASR)
- Manual production of pronunciations slow and costly
  - e.g. 19.2–30s / word for Afrikaans (*Davel and Barnard, 2004*)

➔ Semi-automatic approaches reduce human effort but

  - are still time-consuming.
  - start with an empty dictionary.
    … and initial grapheme-to-phoneme (G2P)
    training data is generated manually
  - usually one single favorite G2P converter is used.

Semi-automatic pronunciation generation

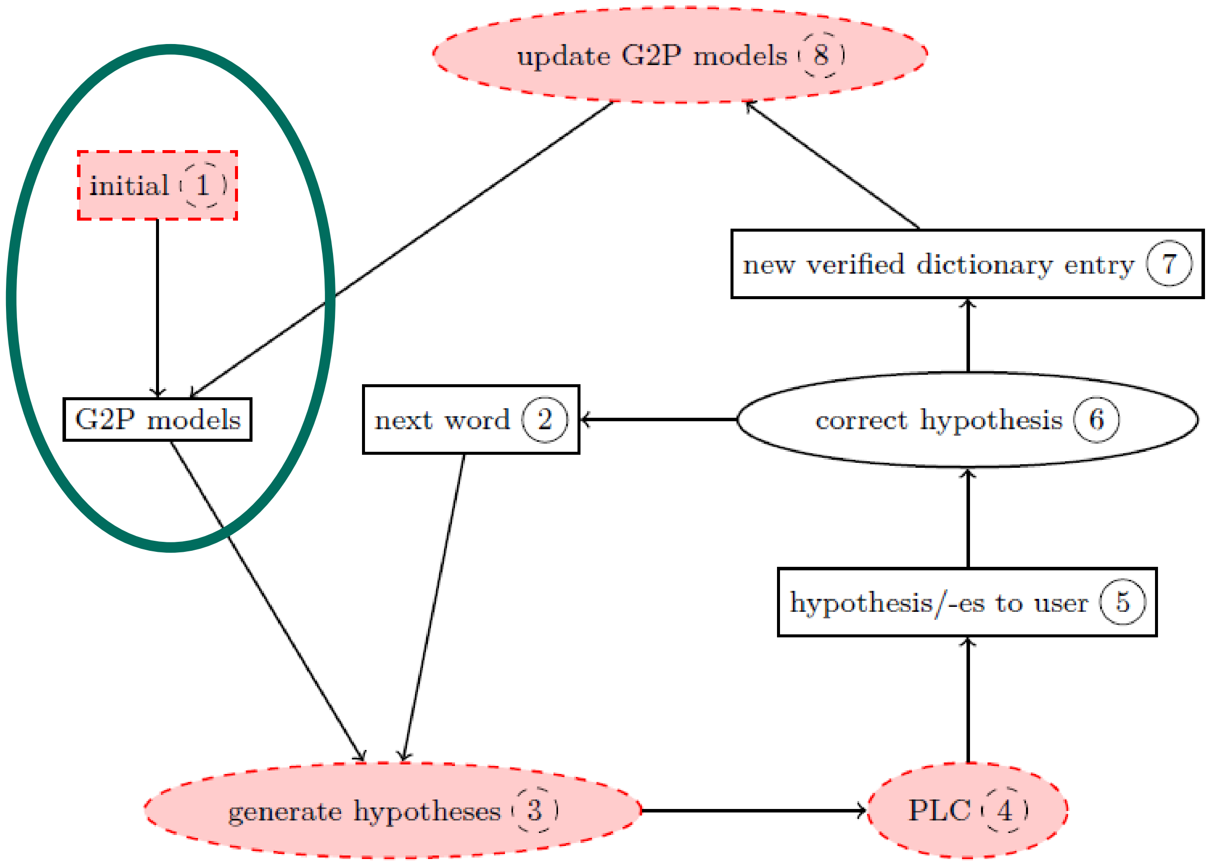
Cognitive Systems Lab



- Manual pronunciation generation supported with estimated pronunciation
- Data-driven G2P converter for producing estimated pronunciation iteratively updated

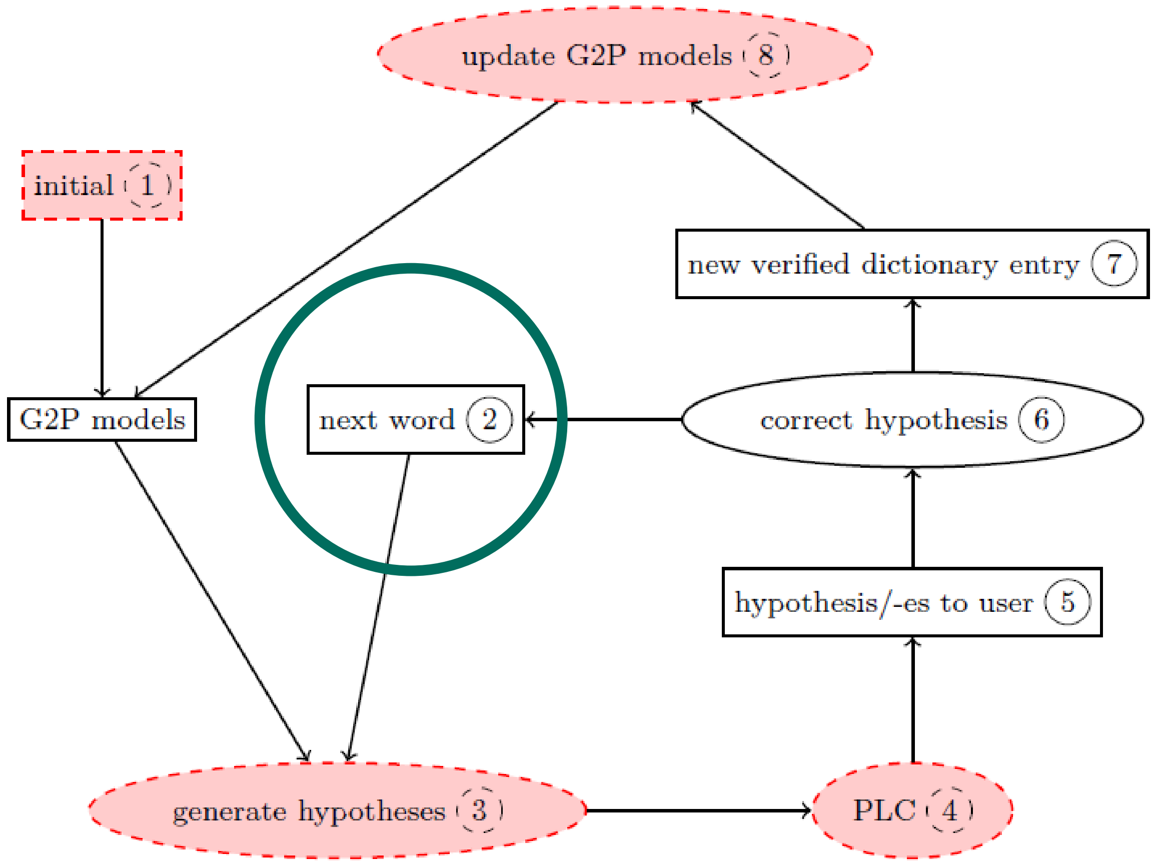
Karlsruhe Institute of Technology

Semi-automatic pronunciation generation



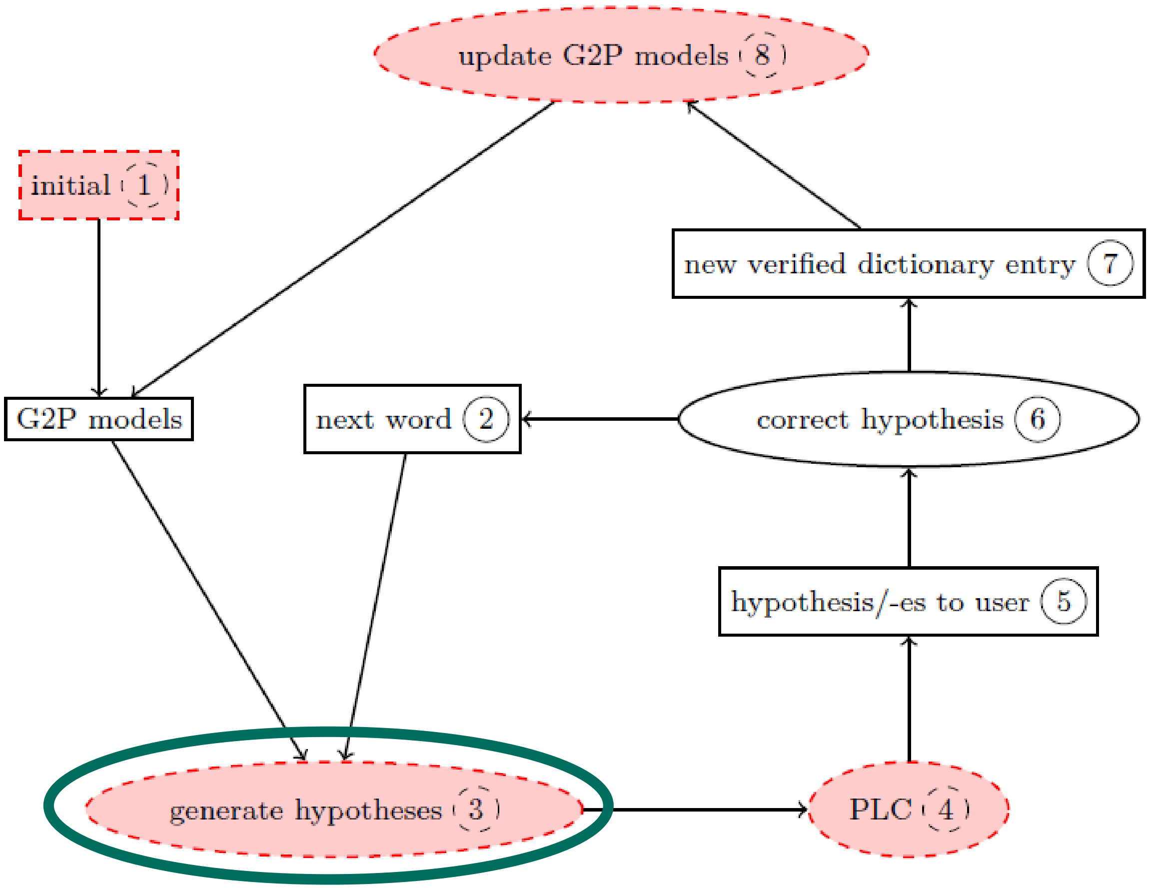
1. Word-pronunciation pairs to train initial G2P converters.

Semi-automatic pronunciation generation



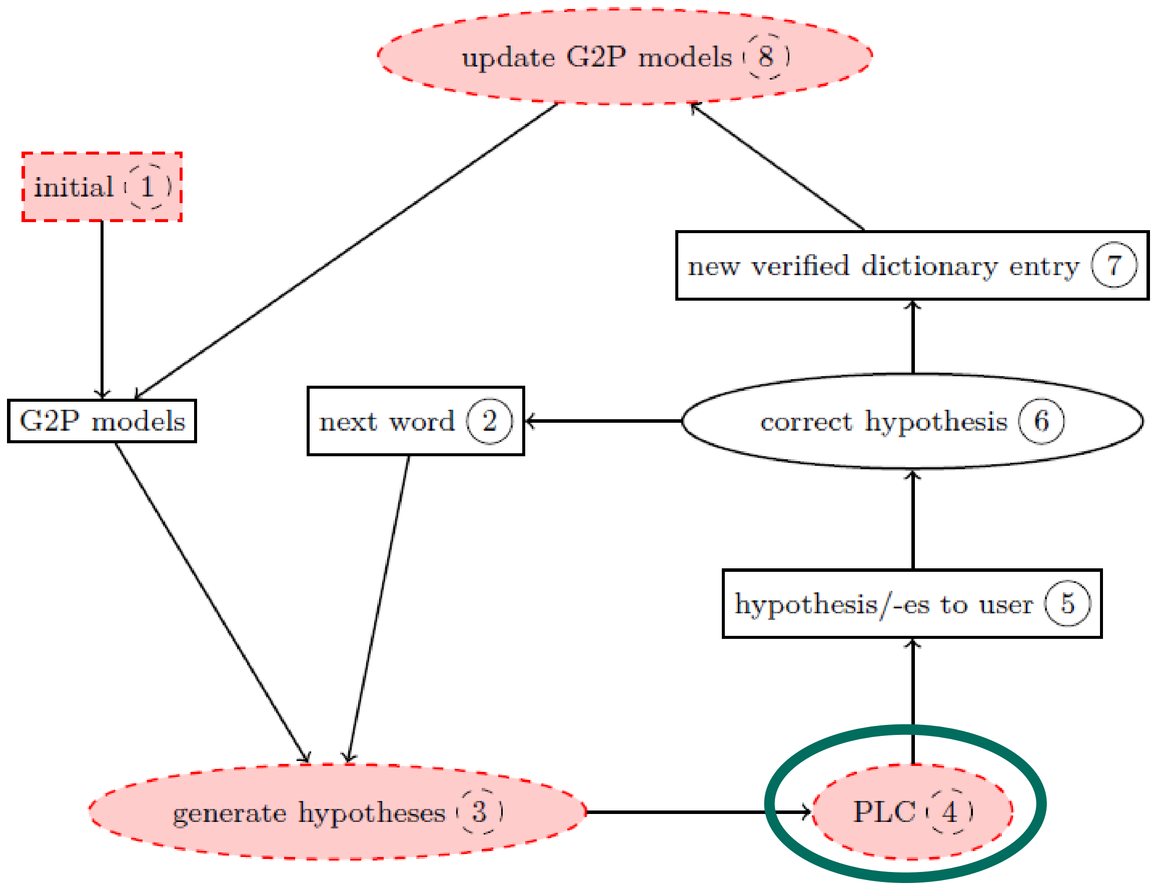
2. A word is determined for pronunciation generation.

Semi-automatic pronunciation generation



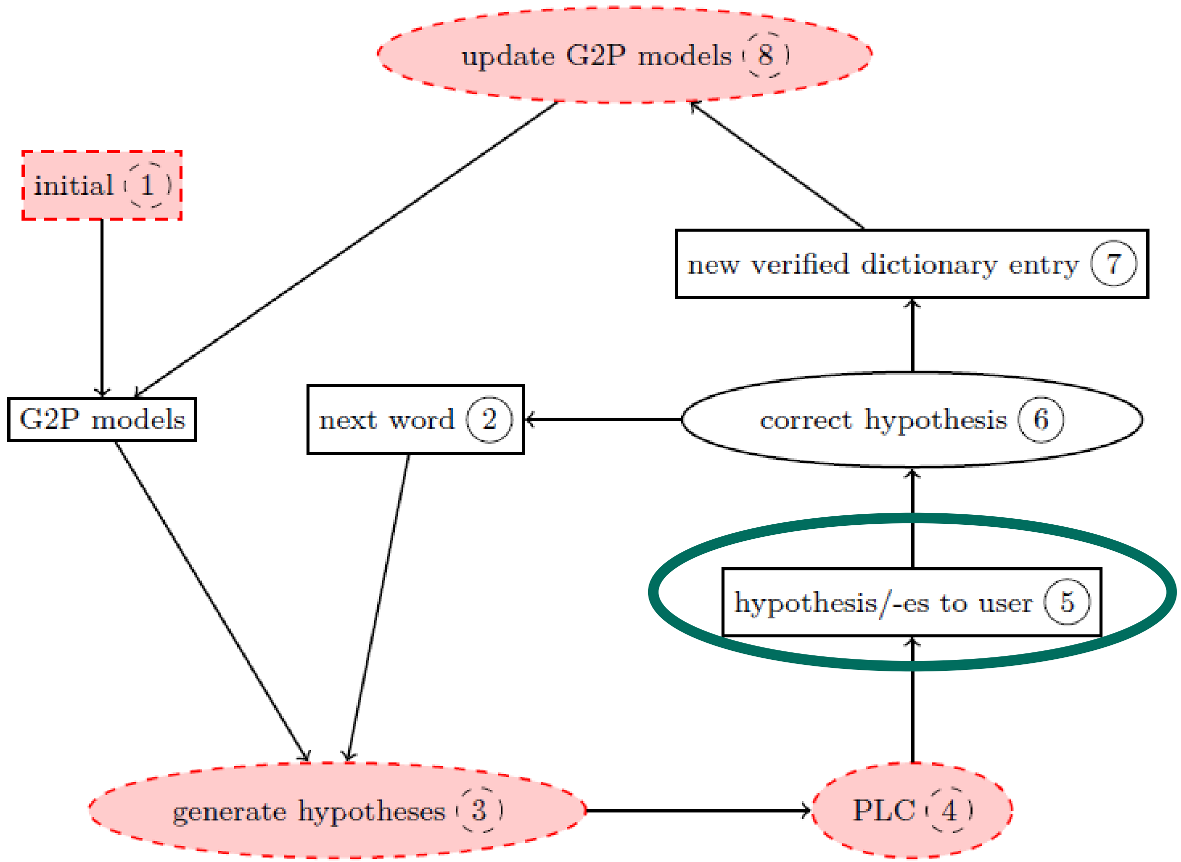
3. Pronunciations of that word are generated (by multiple G2P models).

Semi-automatic pronunciation generation



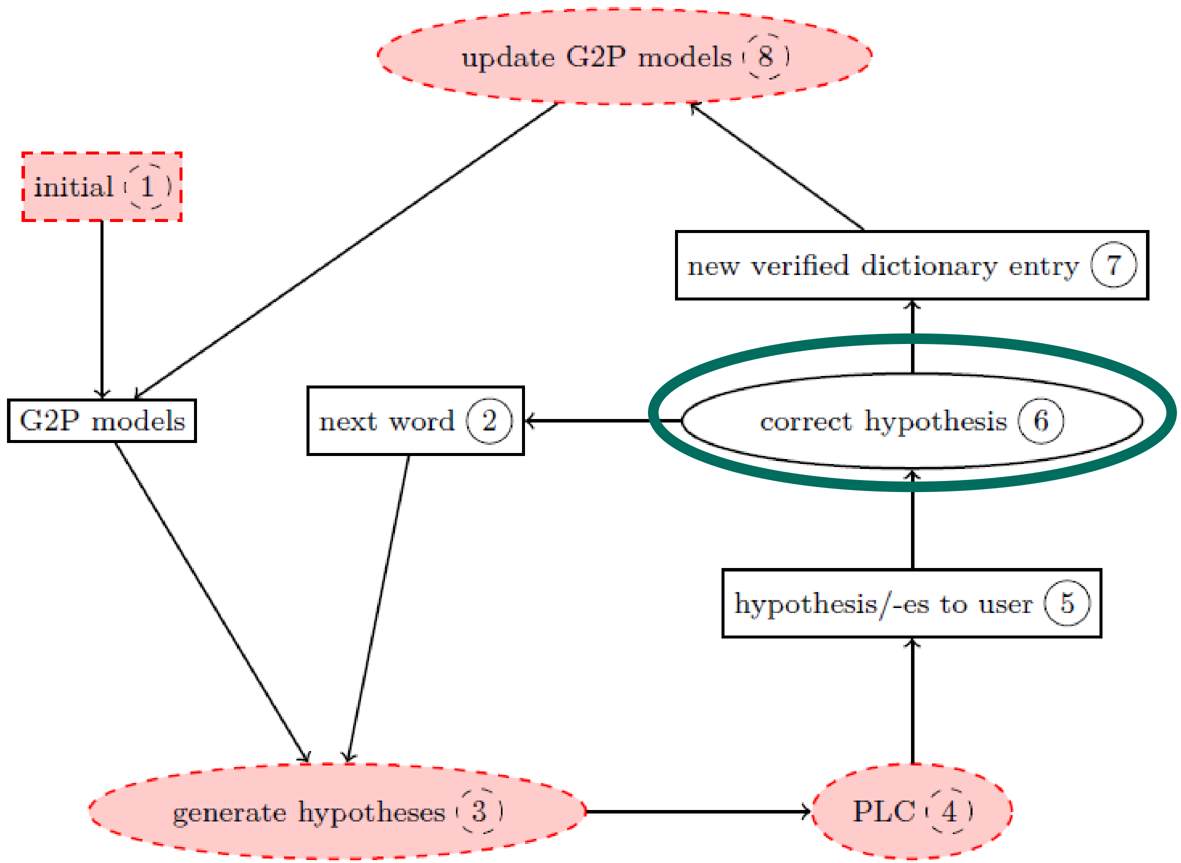
4. Phoneme-level combination (PLC) of different G2P converter outputs.

Semi-automatic pronunciation generation



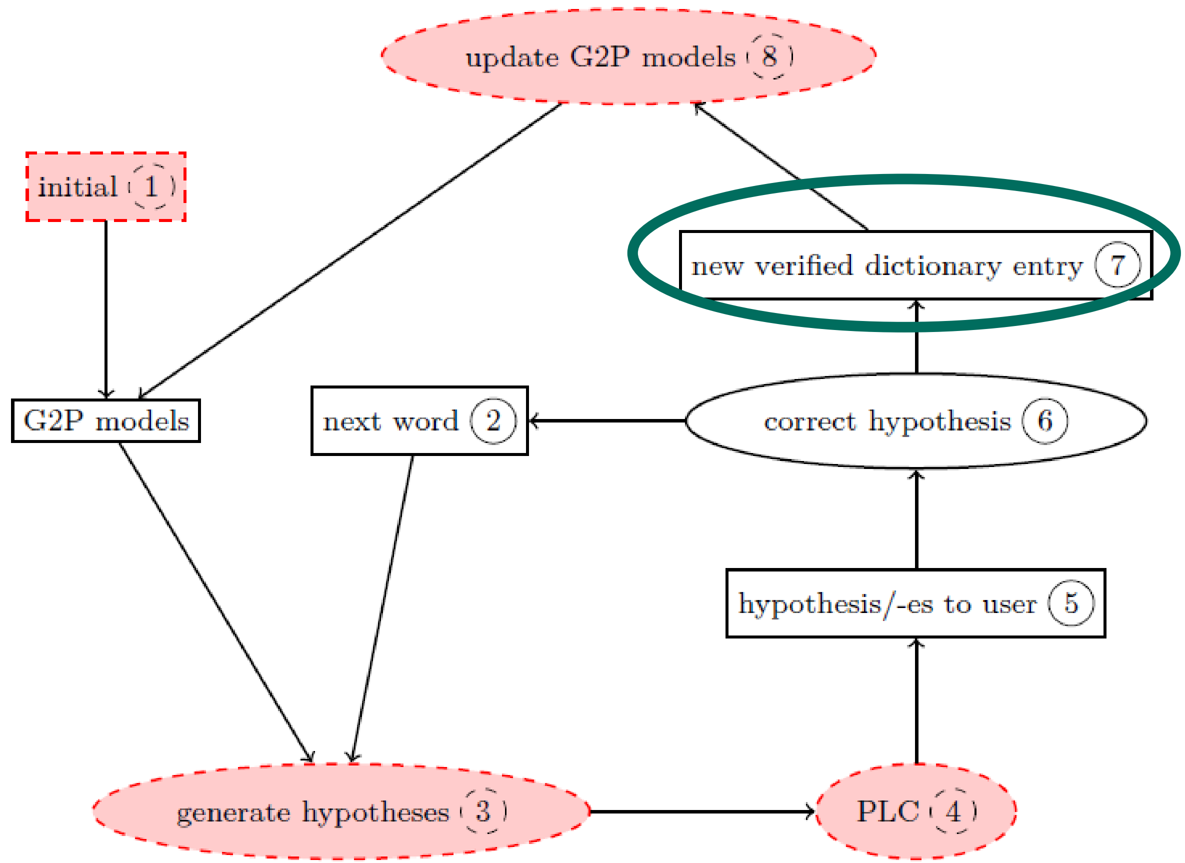
5. Display pronunciation to the user.

Semi-automatic pronunciation generation



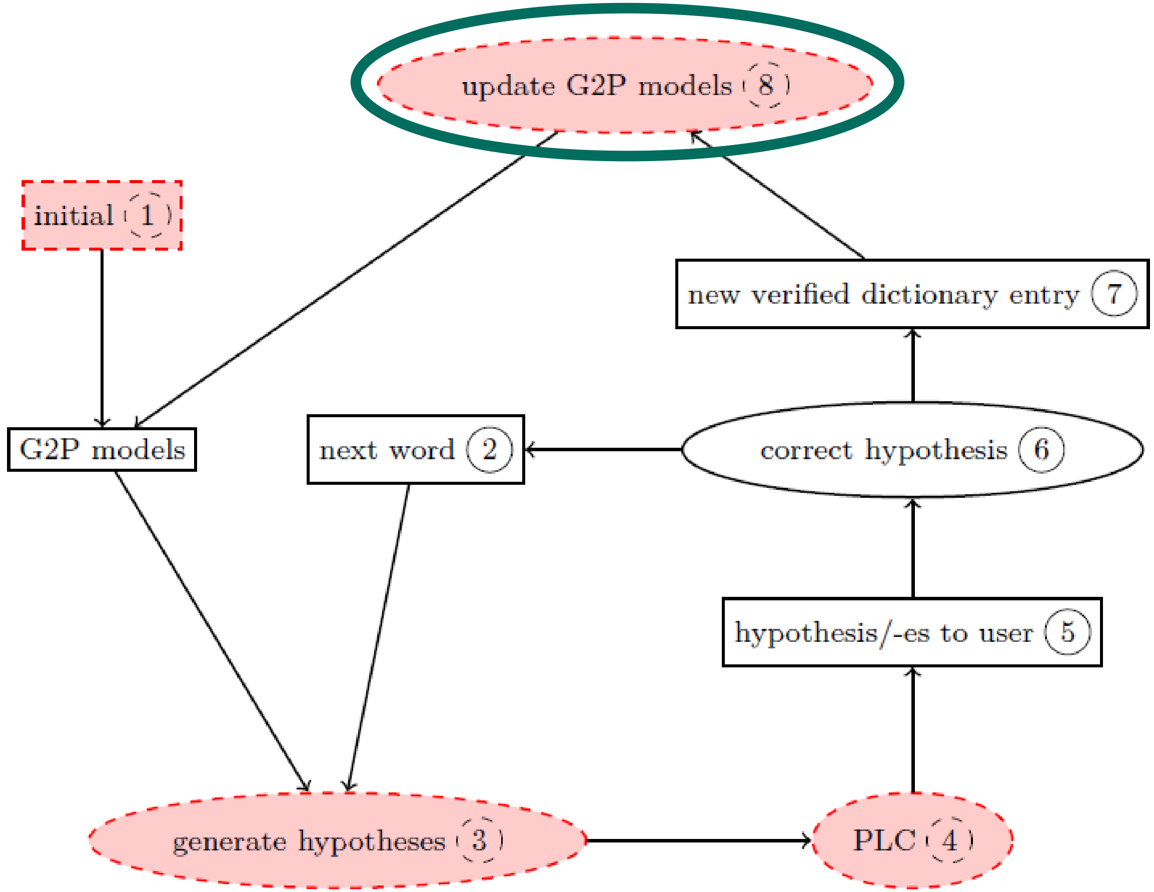
6. User edits displayed pronunciation to the correct pronunciation.

Semi-automatic pronunciation generation



7. Word and corrected pronunciation are added to the dictionary.

Semi-automatic pronunciation generation



8. After a certain number of words, the G2P converters update their grapheme-to-phoneme models.

Experimental Setup

- Languages
  - English, German, Spanish, Vietnamese, Swahili, Haitian Creole
  - Differing in their grade of G2P relationship
  - Reference dictionaries from CMU, GlobalPhone (*Schultz&Schlippe, 2014*)

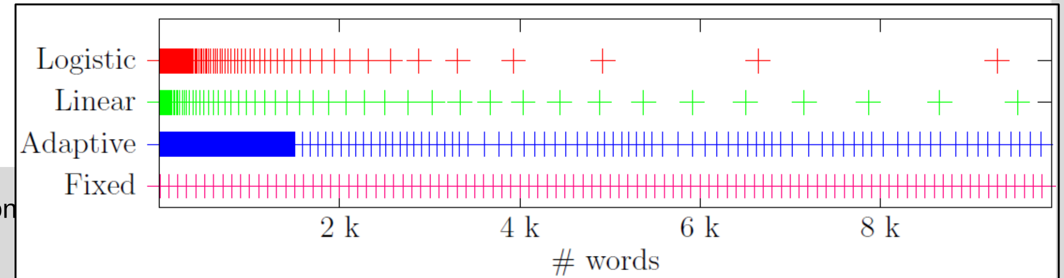
- G2P converters
  - Sequitur G2P (*Bisani&Ney, 2008*),
  - Phonetisaurus (*Novak et al., 2012*),
  - Default&Refine (*Davel, 2005*),
  - CART tree (*Lenzo, 1997*)

- Phoneme-level combination (PLC) (*Schlippe et al., 2014*)

Experimental Setup

- G2P model update interval (CPU time vs. G2P quality)
  - Fixed (Update after fixed number of new pronunciations)
  - Linear (linearly growing intervals)
  - Adaptive
    - First: after each word,
      later: after certain number of predicted pronunciations needed corrections

- Our tradeoff: Logistic growth
  - Build G2P models frequently for small dictionaries
    … where generation is cheap and gain is high
  - Stable G2P models for larger dictionaries are updated less often
    … while still obeying an upper bound for updates



Experimental Setup

- Evaluated different sources for initial pronunciations
  - Pronunciations derived from 1:1 G2P mapping
    (most commonly associated phoneme for each grapheme)

Word	enact	balls	moans
Reference	EH N AE K T	B AO L ZH	M OW N ZH
Grapheme-based	e n a c t	b a l l s	m o a n s
Grapheme-based (mapped)	EH N AX K T	B AX L L S	M AW AX N S

  - Web-driven G2P converters / Web-derived Pronunciations (Wiktionary)

Evaluation Metric

- Human editing effort measured by
  - Total number of edits
    … not comparable between different dictionary sizes or languages
  - Cumulated phoneme error rate as a normed value:

$$cPER(n) := \frac{\sum_{i=1}^{n} sub(w_i) + ins(w_i) + del(w_i)}{\sum_{i=1}^{n} phonemes(w_i)}$$

… from beginning of editing process to current word w_i
… comparable across languages and dictionary sizes

- Human effort is analyzed in simulated experiments
  - … assuming that the developer changes the displayed phoneme sequence to the reference

Experiments: Single G2Pconverters

- Single G2P converters on 10k dictionaries (cPER)

	en	de	vi	es	ht	sw
Sequitur	**15.24**	**11.02**	4.83	**2.19**	0.39	0.25
Phonetis.	15.28	11.10	**4.42**	2.28	0.43	**0.21**
D&R	16.80	12.85	5.12	2.23	0.42	**0.21**
CART	20.01	13.89	5.20	2.56	**0.36**	0.26

  - Languages with closer G2P relationship → less effort
  - Sequitur performs best in three of six cases, closely followed by Phonetisaurus
  - Phonetisaurus, Default&Refine and CART tree perform best in one case each

Experiments: G2P converter combination

- Phoneme-level combination (PLC)
  - cPER

	en	de	es	vi	sw	ht	average
Best single G2P	15.24	11.02	2.19	4.42	0.21	0.36	
PLC	14.47	10.81	2.00	3.78	0.13	0.28	
Relative to single	5.05^s	1.91^s	8.68^s	14.48^s	38.10^s	22.22^s	+15.07

  - Always lower user editing effort
  - On average: 15% relative reduction
  - Between 1.9% and 38.2% relative depending on the language

Experiments: Initial Pronunciations

- Initial pronunciations
  - derived from 1:1 G2P mapping, Web-derived pronunciations (WDP)
  - included in phoneme level combination (PLC)
  - leave out after performance without them is better (cross-over based on average of all tested languages)

  - 1:1 G2P mapping provides benefit in the beginning
  - Wiktionary speeds up whole generation

Experiments: Initial Pronunciations

- Initial pronunciations
  - derived from 1:1 G2P mapping, Web-derived pronunciations (WDP)
  - included in phoneme level combination (PLC)

	en	de	es	vi	sw	ht	average
Best single G2P	15.24	11.02	2.19	4.42	0.21	0.36	
PLC	14.47	10.81	2.00	3.78	0.13	0.28	
Relative to single	5.05^s	1.91^s	8.68^s	14.48^s	38.10^s	22.22^s	+15.07
1:1 G2P mapping + PLC	14.45	10.81	1.97	3.73	0.12	0.26	
Relative to PLC	0.14	0.00	1.50	1.32^s	7.69^s	7.14^s	+2.97
WDP + PLC	13.64	10.22	1.87				
Relative to PLC	5.74^s	5.46^s	6.50^s				+5.90

  - Still 6% reduction on top of PLC with WDP
  - 1:1 G2P mapping only 3% on top of PLC

Conclusion and Future Work

Cognitive Systems Lab

- Semi-automatic pronunciation generation strategy
- Evaluated with languages with different G2P relationship
- 15% effort reduction with phoneme-level combination
- 6% on top with Web-driven G2P converters
- Cross-lingual G2P conversion (*Schlippe at al., 2013*) for initial pronunciations gave only slight improvement

- Future work:
  - Integrate additional G2P converters
  - Improve phoneme-level combination
  - Testing our semi-automatic pronunciation generation strategy with real human developers

Thank you!

References

[1] Sameer R. Maskey, Alan W. Black, and Laura M. Tomokiyo, "Bootstrapping Phonetic Lexicons for New Languages," in *8th International Conference on Spoken Language Processing (ICSLP)*, Jeju, Korea, 4-8 October 2004.

[2] Marelie Davel and Olga Martirosian, "Pronunciation Dictionary Development in Resource-Scarce Environments," in *10th Annual Conference of the International Speech Communication Association (Interspeech)*, Brighton, UK, 6-10 September 2009.

[3] Tanja Schultz, Alan W. Black, Sameer Badaskar, Matthew Hornyak, and John Kominek, "SPICE: Web-based Tools for Rapid Language Adaptation in Speech Processing Systems," in *The 8th Annual Conference of the International Speech Communication Association (Interspeech 2007*, Antwerp, Belgium, 27-31 August 2007.

[4] John Kominek, *TTS From Zero: Building Synthetic Voices for New Languages*, Ph.D. thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 2006.

[5] Ngoc Thang Vu, Tim Schlippe, Franziska Kraus, and Tanja Schultz, "Rapid Bootstrapping of five Eastern European Languages using the Rapid Language Adaptation Toolkit," in *The 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, Makuhari, Japan, 26-30 September 2010.

[6] John Kominek and Alan W. Black, "Learning Pronunciation Dictionaries: Language Complexity and Word Selection Strategies," in *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, New York, New York, USA, 4-6 June 2006.

[7] Marelie Davel and Etienne Barnard, "Efficient Generation of Pronunciation Dictionaries: Human Factors During Bootstrapping," in *8th International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, Korea, 4-8 October 2004.

[8] Marelie Davel and Etienne Barnard, "The Efficient Creation of Pronunciation Dictionaries: Machine Learning Factors in Bootstrapping," in *8th International Conference on Spoken Language Processing*, Jeju Island, Korea, 4-8 October 2004.

[9] Marelie Davel and Etienne Barnard, "Bootstrapping Pronunciation Dictionaries: Practical Issues," in *The Annual Conference of the International Speech Communication Association (Interspeech 2005)*, Lisbon, Portugal, 4-8 September 2005.

[10] Steven Lawrence Davis, Shane Fetters, Beverly Gustafson, Louise Loney, and David Eugene Schulz, "System and Method for Preparing a Pronunciation Dictionary for a Text-to-Speech Voice," Tech. Rep. US Patent 7630898 B1, AT&T, September 2005.

[11] Arnab Ghoshal, Martin Jansche, Sanjeev Khudanpur, Michael Riley, and Morgan Ulinski, "Web-Derived Pronunciations," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 19-24 April 2009.

[12] Tim Schlippe, Sebastian Ochs, and Tanja Schultz, "Wiktionary as a Source for Automatic Pronunciation Extraction," in *The 11th Annual Conference of the International Speech Communication Association (Interspeech)*, Makuhari, Japan, 26–30 September 2010.

[13] Tim Schlippe, Wolf Quaschningk, and Tanja Schultz, "Combining Grapheme-to-Phoneme Converter Outputs for Enhanced Pronunciation Generation in Low-Resource Scenarios," in *The 4th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2014)*, St. Petersburg, Russia, 14-16 May 2014.

References

[14] Ngoc Thang Vu and Tanja Schultz, "Vietnamese Large Vocabulary Continuous Speech Recognition," in *2009 IEEE Automatic Speech Recognition and Understanding (ASRU)*, Merano, Italy, 13-17 December 2009.

[15] Tanja Schultz and Tim Schlippe, "GlobalPhone: Pronunciation Dictionaries in 20 Languages," in *The 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, 26–31 May 2014.

[16] Tanja Schultz, Ngoc Thang Vu, and Tim Schlippe, "GlobalPhone: A Multilingual Text & Speech Database in 20 Languages," in *The 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, Vancouver, Canada, 26-31 May 2013.

[17] Maximilian Bisani and Hermann Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," in *Speech Communication*, May 2008, vol. 50, issue 5, pp. 434–451.

[18] Josef R. Novak, "Phonetisaurus: A WFST-driven Phoneticizer," http://code.google.com/p/phonetisaurus/, 2011.

[19] Josef R. Novak, Nobuaki Minematsu, and Keikichi Hirose, "WFST-based Grapheme-to-Phoneme Conversion: Open Source Tools for Alignment, Model-Building and Decoding," in *The 10th International Workshop on Finite State Methods and Natural Language Processing*, Donostia–San Sebastin, July 2012, Association for Computational Linguistics.

[20] Marelie Davel, "The Default & Refine Algorithm, A Rule-based Learning Algorithm," http://code.google.com/p/defaultrefine/, August 2005.

[21] Marelie Davel and Etienne Barnard, "A Default-and-Refinement Approach to Pronunciation Prediction," in *Symposium of the Pattern Recognition Association of South Africa*, South Africa, 2004.

[22] Marelie Davel and Etienne Barnard, "Pronunciation Prediction with Default & Refine," *Computer Speech and Language*, vol. 22, no. 4, pp. 374–393, October 2008.

[23] Kevin Lenzo, "t2p: Text-to-Phoneme Converter Builder," http://www.cs.cmu.edu/afs/cs.cmu.edu/user/lenzo/html/areas/t2p/, 1997.

[24] Andreas Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *International Conference on Spoken Language Processing (ICSLP)*, Denver, Colerado, 16-20 September 2002.

[25] Tim Schlippe, Sebastian Ochs, and Tanja Schultz, "Web-based Tools and Methods for Rapid Pronunciation Dictionary Creation," *Speech Communication*, vol. 56, no. 0, pp. 101 – 118, 2014.

[26] Tim Schlippe, Sebastian Ochs, and Tanja Schultz, "Grapheme-to-Phoneme Model Generation for Indo-European Languages," in *The 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 25-30 March 2012.

[27] Tim Schlippe, Sebastian Ochs, and Tanja Schultz, "Automatic Error Recovery for Pronunciation Dictionaries," in *The 13th Annual Conference of the International Speech Communication Association (Interspeech 2012)*, Portland, Oregon, 9–13 September 2012.

[28] Tim Schlippe, Mykola Volovyk, Kateryna Yurchenko, and Tanja Schultz, "Rapid Bootstrapping of a Ukrainian Large Vocabulary Continuous Speech Recognition System," in *The 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, Vancouver, Canada, 26-31 May 2013.

[29] Dilek Hakkani-Tür, Giuseppe Riccardi, and Allen Gorin, "Active Learning For Automatic Speech Recognition," in *The 27th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, Orlando, Florida, 13 – 17 May 2002.