Pronunciation Assessment Proposal
Jim Salsman
April 2019

Item (1): Nakagawa et al's 2011 intelligibility vs.
SRI's 1995 "Goodness of Pronunciation" assessments:

82% agreement with the accuracy of crowdworkers' transcriptions, up from 75% reported by the inventor
— arxiv.org/abs/1709.01713

State of the art:

Educational Testing Service's SpeechRater 5.0
"System-Human agreement" is **58.4%** (Chen *et al.*, 2018.)

Terms

Speech recognition

Pronunciation assessment

Fluency assessment
(Usually requires accent and dialect adaptation)

Intelligibility prediction (Nakagawa et al., 2011)

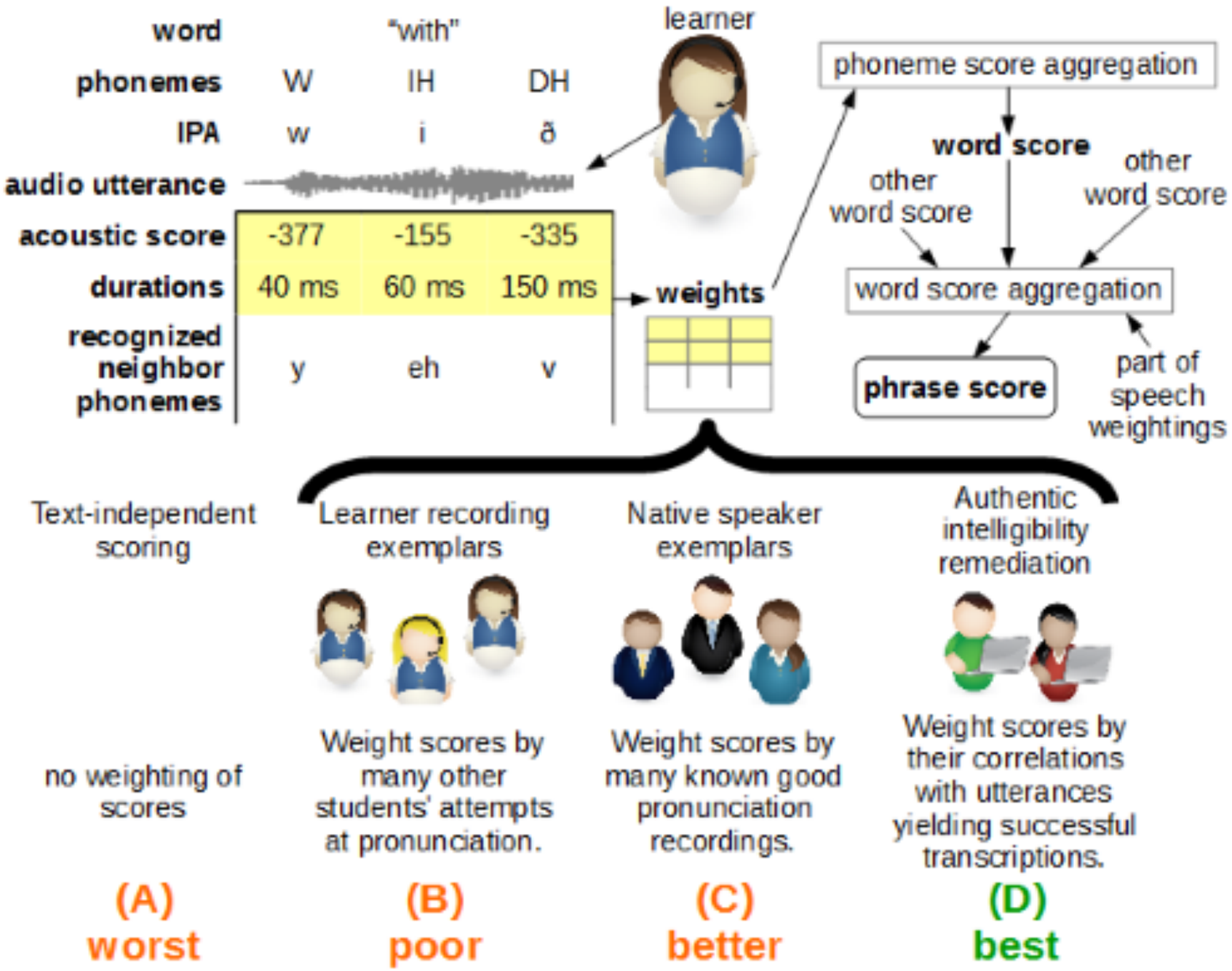
Remediation

Feedback -- natural speech or visual?

Interactions -- web app or download delay?

Sequencing -- good learner analytics?

Motivation
Intelligibility assessment



Why accent adaptation?

Fig. 1. Accents of the British Isles.

14 accent loci in UK and Ireland,

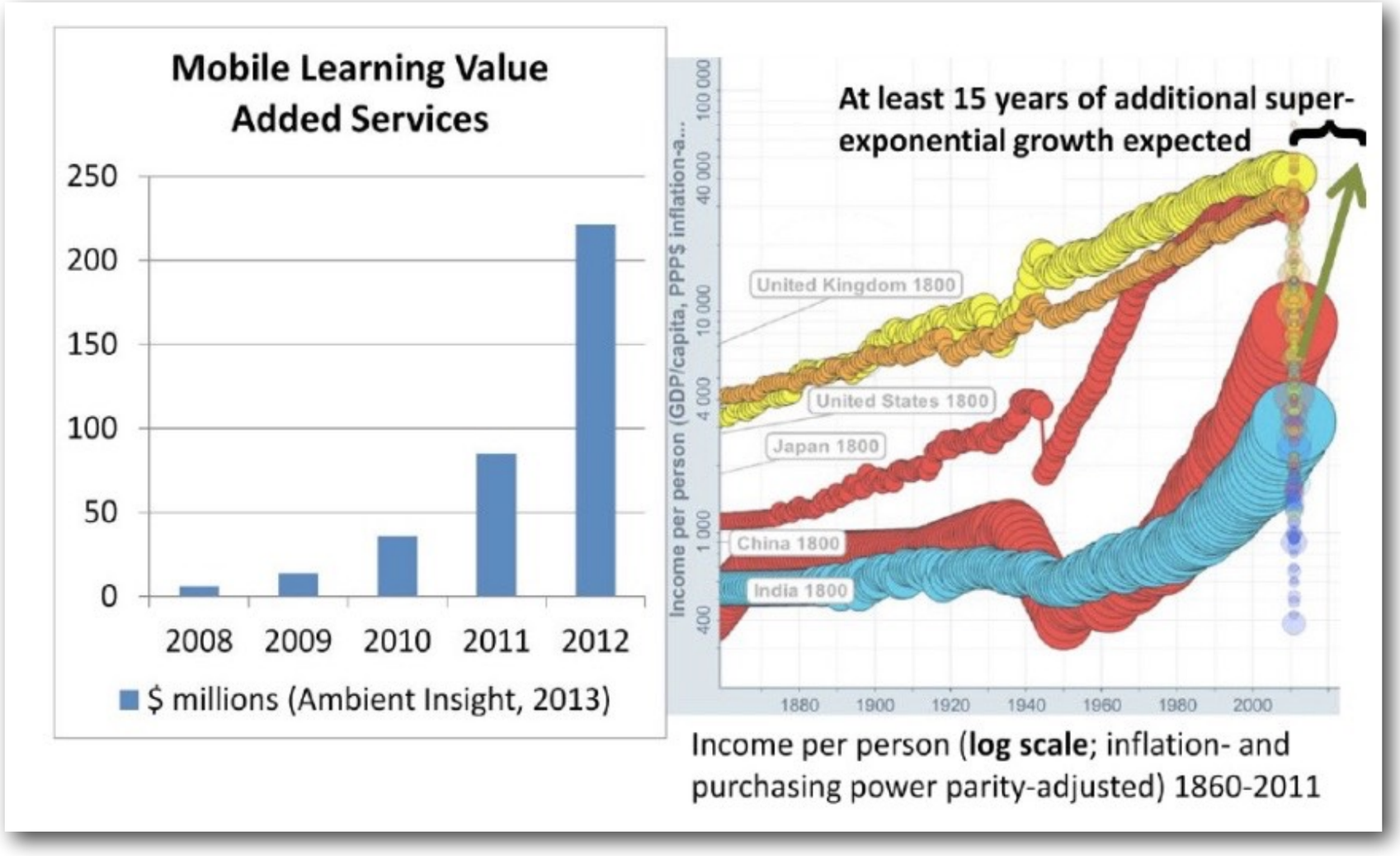
3-4 in the USA and Canada,

Australia, South Africa, New Zealand, and

English as a Foreign Language everywhere

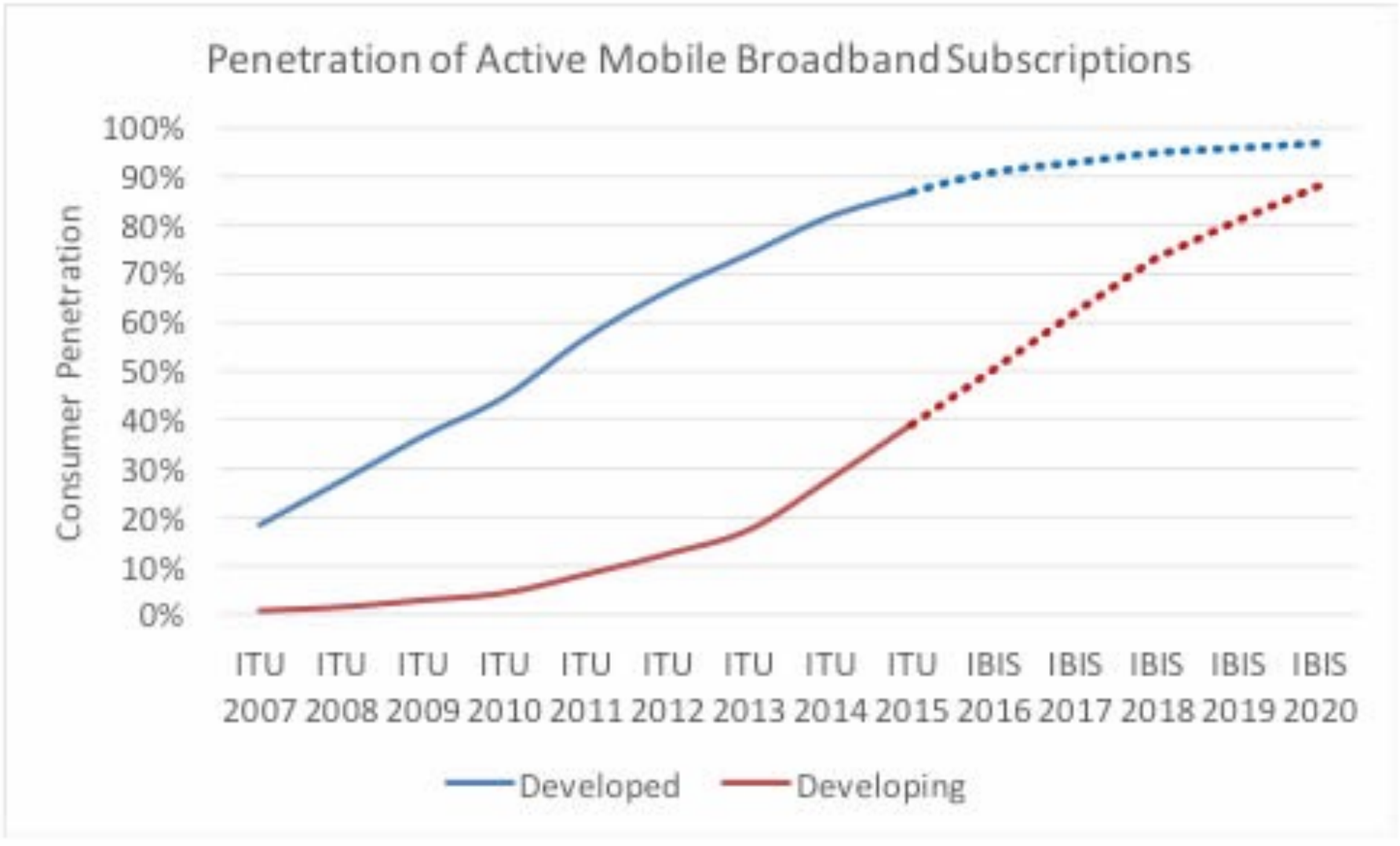
At least 150 feasibly discernible accents.

Market size and opportunity

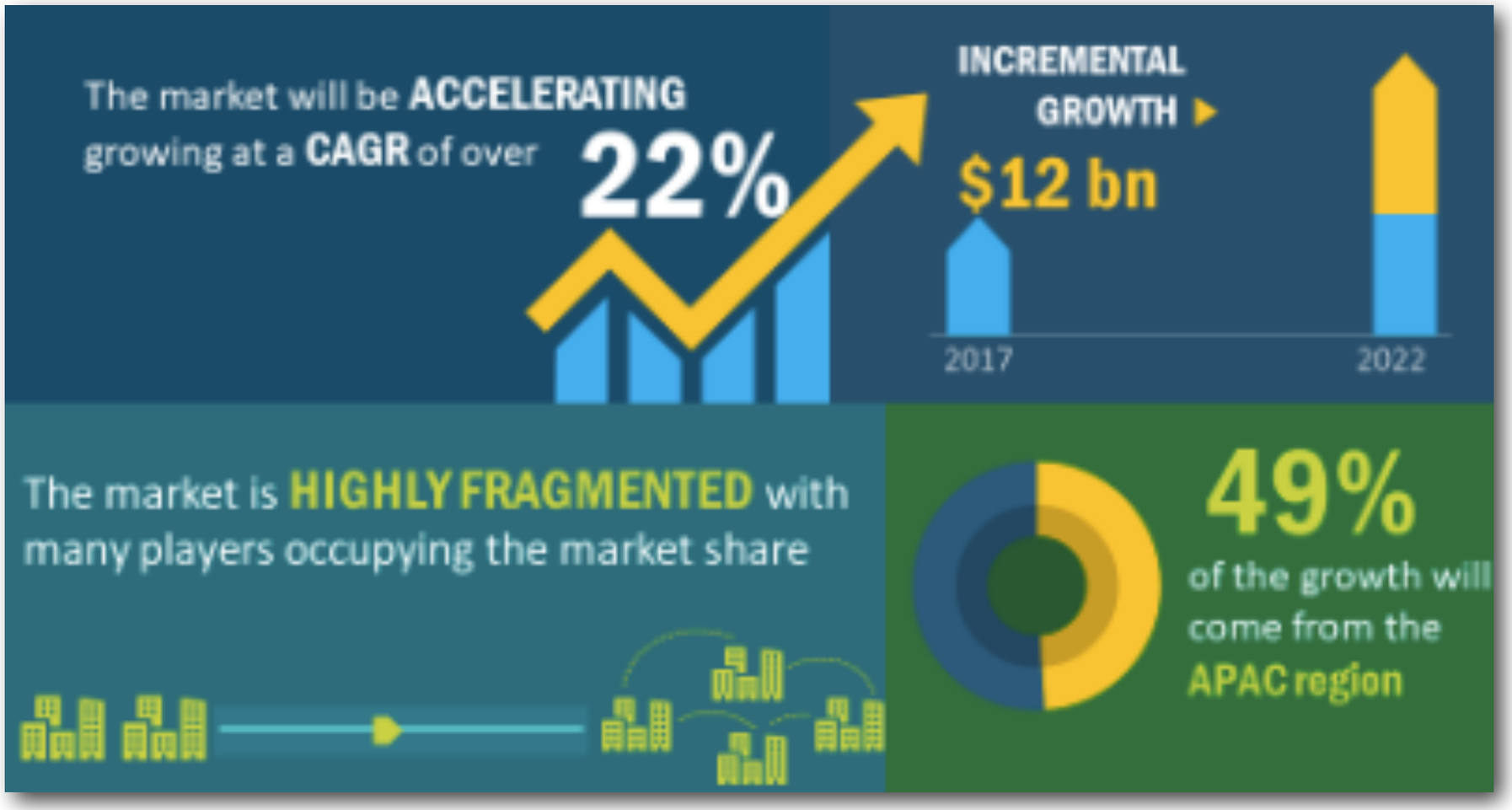


- Online language learning was a $6 billion market in 2011. The global market for digital English language learning products reached $1.8 billion in 2013, and exceeded $3 billion in 2018.

Market
English language instruction



Market size and opportunity



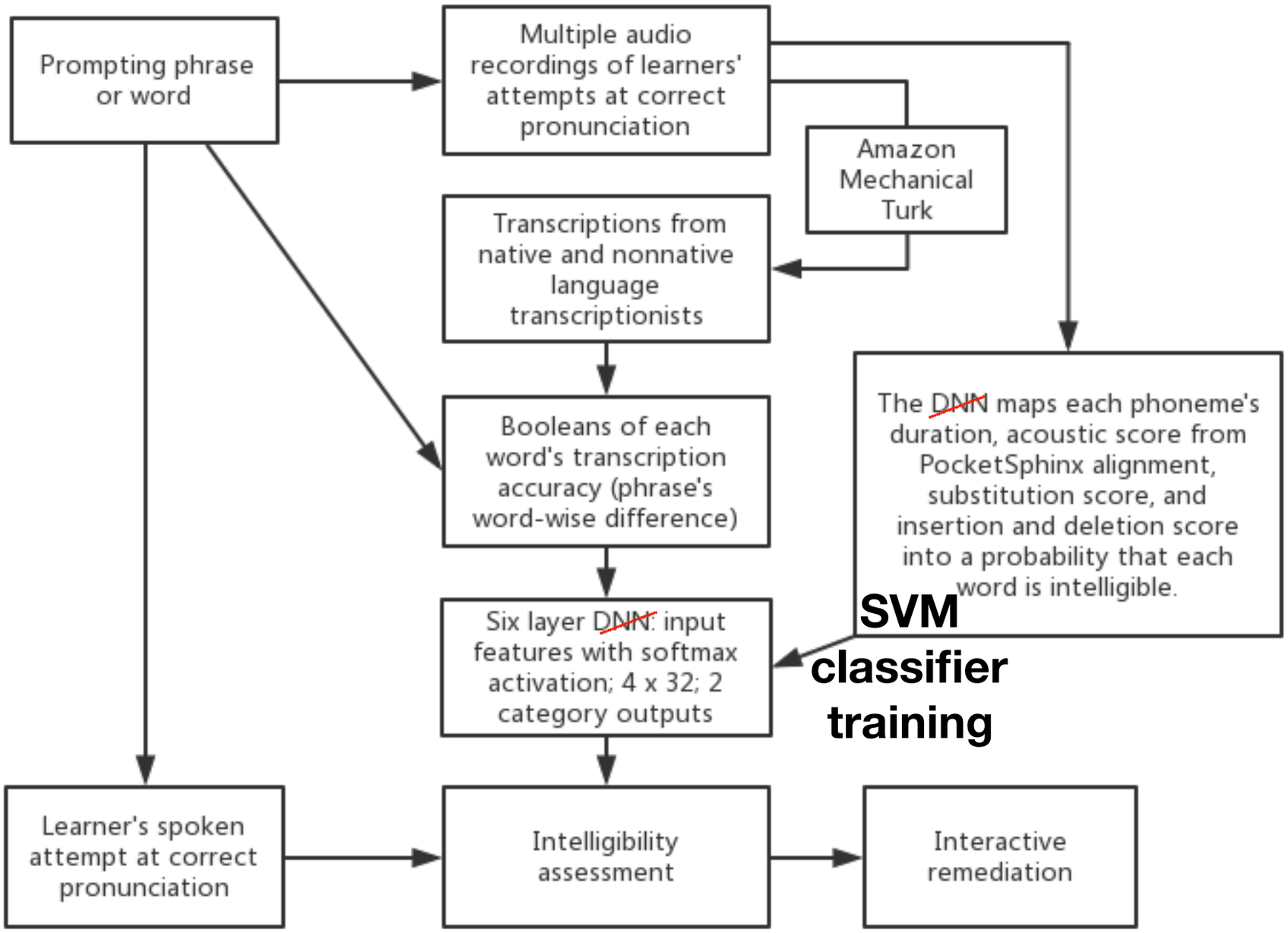
- The year-over-year growth rate was 20% in 2018. *WiseGuyReports* projects the global digital English language learning market will surpass $18 billion by 2022.

Goal

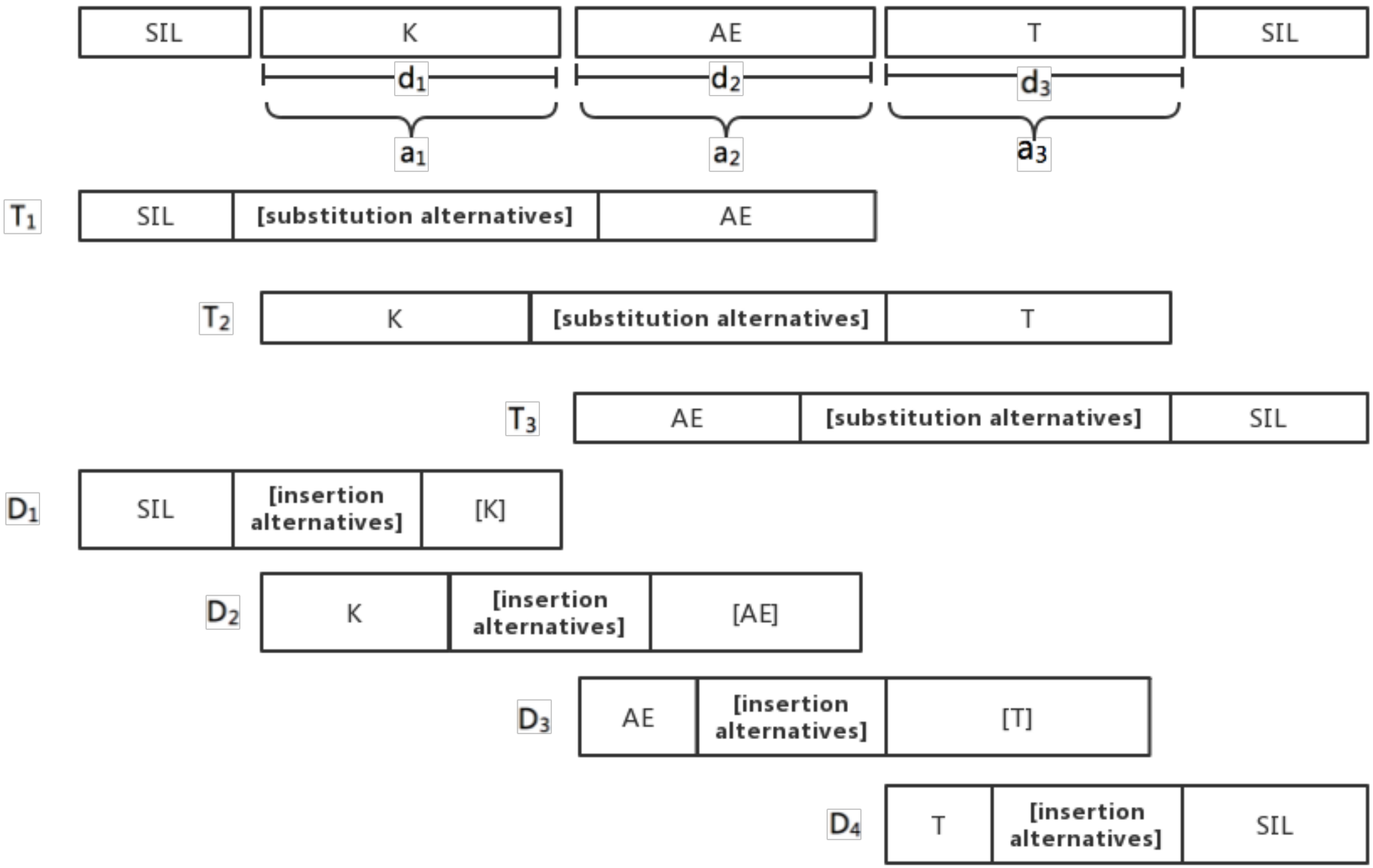
The goal is to ask people to try to pronounce words, and some day phrases, in a way that speech recognition features predict will be correctly transcribed by those who hear the audio utterance. This technique can correctly adapt to spoken accents like vowel shifts, but not dialect.

Learners provide needed transcriptions of student (peer) speech.

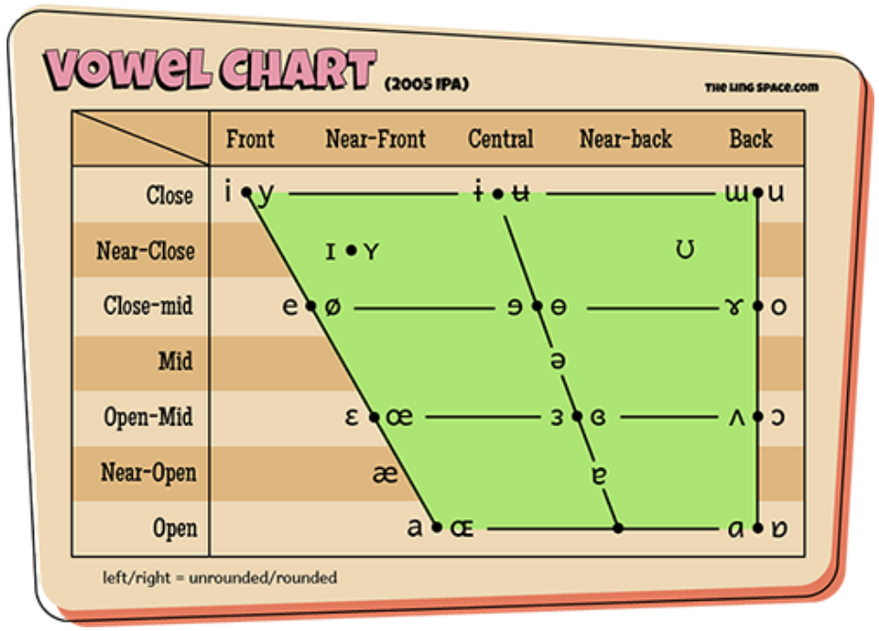
Intelligibility remediation Data flow diagram



Intelligibility remediation
Scoring: 4 features/phoneme



Intelligibility remediation
Scoring: 10 features/phoneme



Pulmonic consonants [edit source]

Place → / Manner ↓	Labial			Coronal					Dorsal			Laryngeal		
	Bilabial	Labio-dental	Linguo-labial	Dental	Alveolar	Palato-alveolar	Retro-flex	Alveolo-palatal	Palatal	Velar	Uvular	Pharyn-geal	Epi-glottal	Glottal
Nasal	m̥ m		ɱ		n̥ n		ɳ̊ ɳ		ɲ̊ ɲ	ŋ̊ ŋ	ɴ			
Stop	p b	p̪ b̪	t̪ d̪		t d		ʈ ɖ		c ɟ	k g	q ɢ		ʡ	ʔ
Sibilant affricate					ts dz	tʃ dʒ	tʂ dʐ	tɕ dʑ						
Non-sibilant affricate	pɸ bβ	pf bv		t̪θ d̪ð	t̪s̪ d̪z̪	t̠ʃ d̠ʒ			cç ɟʝ	kx gɣ	qχ		ʡʜ	ʔh
Sibilant fricative					s z	ʃ ʒ	ʂ ʐ	ɕ ʑ						
Non-sibilant fricative	ɸ β	f v	θ̼ ð̼	θ ð	θ̠ ð̠		ɻ̊˔ ɻ˔		ç ʝ	x ɣ	χ ʁ	ħ ʕ		h ɦ
Approximant		ʋ̥ ʋ			ɹ̥ ɹ		ɻ̊ ɻ		j̊ j	ɰ̊ ɰ				ʔ̞
Flap/tap		ⱱ̟ ⱱ	ɾ̼		ɾ̥ ɾ		ɽ̊ ɽ				ɢ̆		ʡ̮	
Trill	ʙ̥ ʙ		r̼		r̥ r		ɽ̊r̥ ɽr				ʀ̥ ʀ		ʜ ʢ	
Lateral affricate					tɬ dɮ		ꞎ̊	cʎ̝		kʟ̝ gʟ̝				

nsonant.

help · **full chart** · template

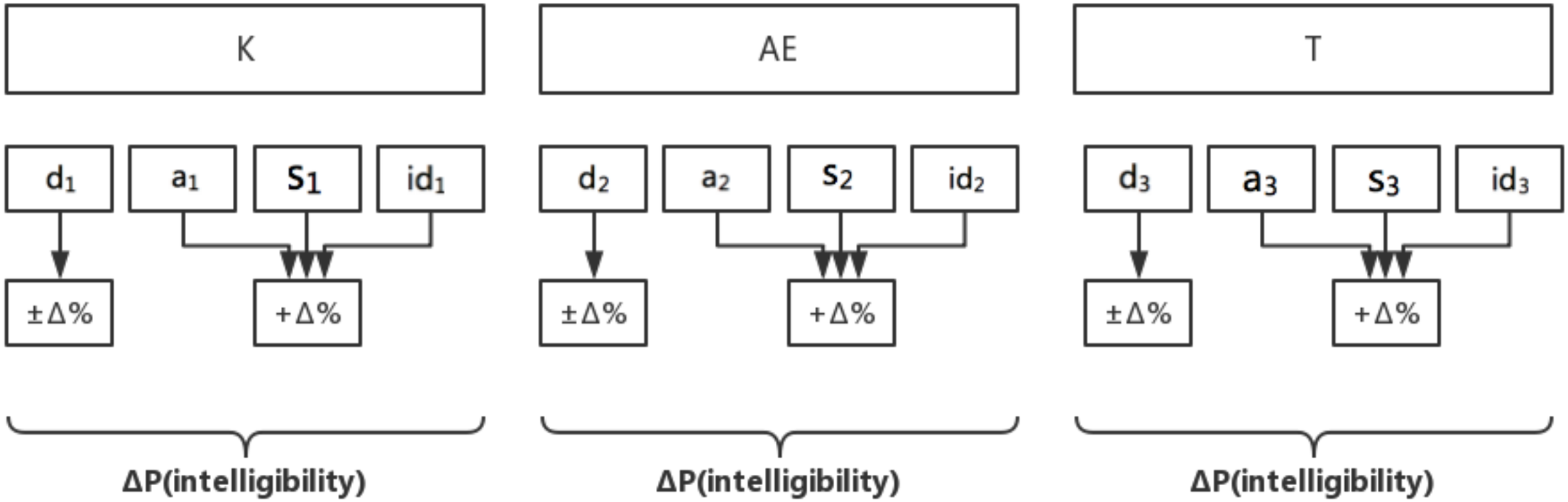
Phoneme (CMUBET)	IPA (International Phonetic Alphabet)	Place: 1.0-5.0	Closedness: 1.0-7.0	Roundedness: no=-1.0, yes=1.0	Voiced: unvoiced consonant=-1.0, voiced consonant=+0.5, vowel=+1.0	PNLL (Proportion Neighbors Less Likely)
AA	**ɑ**	5	1	-1	1	
AE	**æ**	1	2	0	1	
AH	**ʌ**	5	3	-1	1	These
AO	**ɔ**	5	3	1	1	values
AW	**ɑʊ**	4.5	3.5	-0.5	1	are
AY	**ɑɪ**	3.5	3.5	-1	1	calculated
B	**b**	1	1	-1	0.5	from
CH	**ʧ**	3.5	4.0	-0.5	-1	the
D	**d**	3	4.0	-0.5	0.5	recognition
DH	**ð**	2.5	4.0	0	0.5	results.
EH	**ɛ**	1	3	-1	1	
ER	**ɜɹ**	3	3	-1	1	

Solution

- Natural spoken remediation feedback produces authentic skill improvements without the distraction of visual feedback.

Demo on YouTube

(2) Diphone with most room for improvement: "worst" phoneme(s)



The additional four vocal tract articulation features are all set to zero, and the ninth, proportion of neighbors less likely, is set to 1.0 for this step.

Intelligibility remediation
Manifest and plans

Speech collection
Transcript collection
Transcript integration
Balancing
Sufficiency
Scoring
    Feature extraction
    SVM classifier
    Phoneme with-most-room-for-improvement isolation
ID (email) -- adaptivity and payment processing integration: TBD
Exemplary flag for data collection: DONE
-- collecting transcripts from learners: IN PROGRESS
Multiple choice support tool done but not used yet

Web server API

Testing pronunciation intelligibility assessment

Please try the live microphone demo!

Audio utterance URL:

or upload: Choose File no file selected (reload form to clear upload file)

Format: MP3, WAV at 16,000 per second of 16 bit signed little endian monophonic samples, or raw samples in that format, or M4A.

Evaluate Note: audio will be stored.

Word: airport, argue, ask, baby, banana, bowl, branch, breakfast, brother, bus, chair, coach, contest, december, eighth, enjoy, enough, expensive, family, father, feature, flower, forget, fourth, future, grandma, hall, happy, hill, holiday, homework, hospital, idea, inside, invent, july, lesson, library, litter, live.

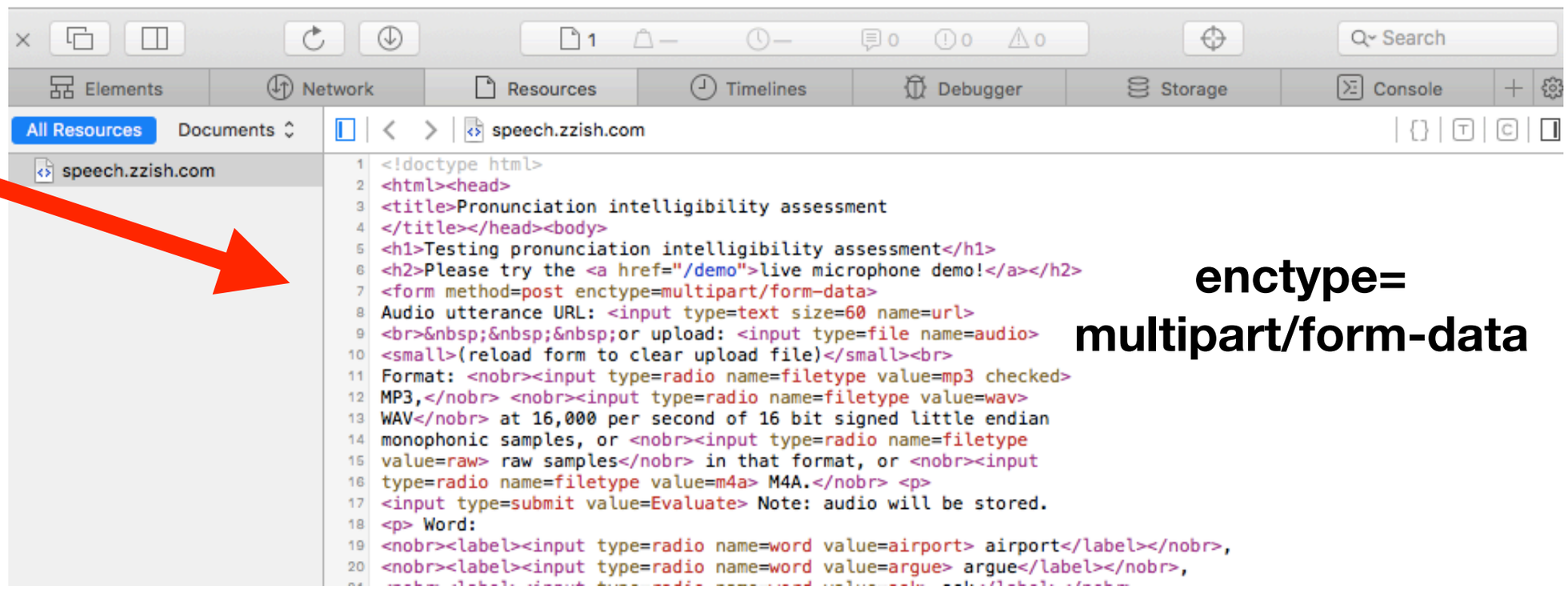
Evaluate

For JSON check here:

JSON

<form>

enctype= multipart/form-data



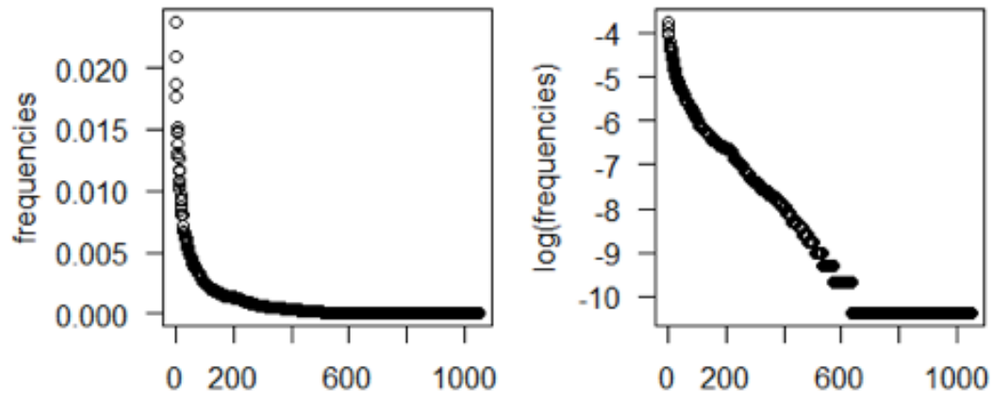
Web pages
API utilization in Javascript

```
mp3blob = new Blob(buffer, {type: 'audio/mp3'});

(window.XMLHttpRequest) ? req = new XMLHttpRequest() :
    (window.ActiveXObject) ? req = new ActiveXObject("Microsoft.XMLHTTP") :
    req = false; // cross platform for IE7 or something

req.open("post", "/");

formdata = new FormData();
formdata.append('url', '');
formdata.append('filetype', 'mp3');
formdata.append('word', document.getElementById("word").innerHTML);
formdata.append('email', document.getElementById('email').value);
formdata.append('exemp', document.getElementById('exemp').checked);
formdata.append('audio', mp3blob);
formdata.append('json', 'checked')

req.addEventListener('load', function(event) {
  document.getElementById('upload').disabled = true;
  var resp = JSON.parse(req.responseText);
  var fUrl = '/rec/index.html?email='
    + escape(document.getElementById('email').value);
  if (resp.prob_good < 50 && resp.feedback != "") {
    fUrl += '&word=' + escape(word) + '&feedback=' + escape(resp.feedback);
  }
  alert('Data sent, parsed response:\n' + req.responseText
        + 'Visiting: ' + fUrl);
  document.location = fUrl;
});
req.addEventListener('error', function(event) {
  alert('Unable to upload.');
});

req.send(formdata);
```

(3) ~650 non-diphthong diphones for speech skill learner analytics, instead of phonemes or words

A *diphone* is the last part of one phoneme followed by the first part of another. Either phoneme could be silence, and they can be the same phoneme. Diphthongs include diphones in them.

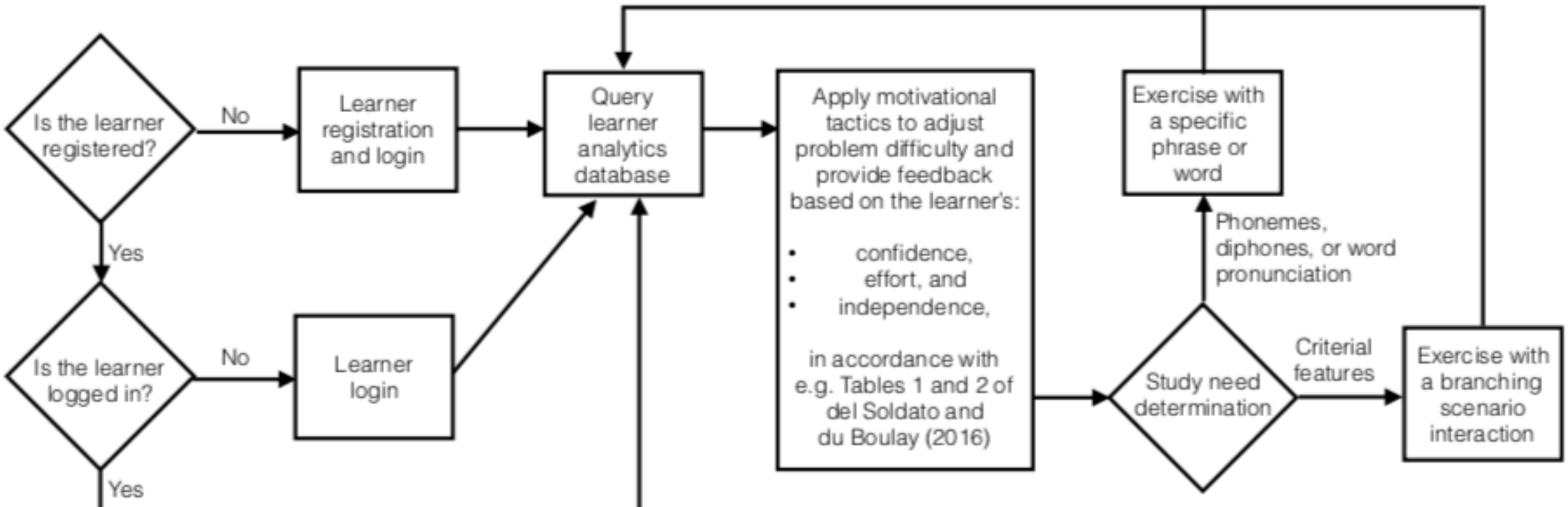
This list of the top 4,800 words by frequency in English speech was used with CMUDICT to create the following list of 1,052 diphones without dipthongs by approximate prevalence.



UH_R 2.376%, AH_N 2.083%, T_SIL 1.863%, Z_SIL 1.758%, SIL_S 1.514%, IY_SIL 1.486%, EH_IY 1.465%, D_SIL 1.387%, N_SIL 1.313%, S_SIL 1.270%, SIL_K 1.264%, R_SIL 1.168%, AA_IY 1.156%, SIL_P 1.078%, IH_NG 1.063%, S_T 1.048%, AH_L 1.017%, NG_SIL 0.973%,

Adaptivity

Learner analytics sequencing and branching scenario transitions



Reference: du Boulay, B. and del Soldato, T. (2016) "Implementation of Motivational Tactics in Tutoring Systems: 20 years on," *International Journal of Artificial Intelligence in Education*, 26(1):170-182,

`http://users.sussex.ac.uk/~bend/papers/motivation-revised2.pdf`

(4) JavaScript PocketSphinx.js recognition on the client web browser

load_prompts_utt.py | pocketsphinxjs-config.h.in | psRecognizer.cpp | load_words.py | combo-new.dict

```cpp
#include "psRecognizer.h"
#include "pocketsphinxjs-config.h"


namespace pocketsphinxjs {
  typedef std::map<std::string, std::string> StringsMapType;
  typedef std::map<std::string, std::string>::iterator StringsMapIterator;

  // Implemented later in this file
  ReturnType parseStringList(const std::string &, StringsSetType*, std::string*);

  Recognizer::Recognizer(): is_fsg(true), is_recording(false), current_hyp(""),
  grammar_index(0) {
    Config c;
    if (init(c) != SUCCESS) cleanup();
  }

  Recognizer::Recognizer(const Config& config) : is_fsg(true), is_recording(false),
  current_hyp(""), grammar_index(0) {
    if (init(config) != SUCCESS) cleanup();
  }

  ReturnType Recognizer::reInit(const Config& config) {
    ReturnType r = init(config);
    if (r != SUCCESS) cleanup();
    return r;
  }

  ReturnType Recognizer::addWords(const std::vector<Word>& words) {
    if (decoder == NULL) return BAD_STATE;
```

(5) Data collection: Words, speech and transcripts

700 words (for comparison the Cambridge/EC English Profile has 6,500 words in levels A1-C2) and phrases;

30-60 recordings per word;

4-12 transcripts per recording; and

4 numeric features per phoneme, upgraded to 10.

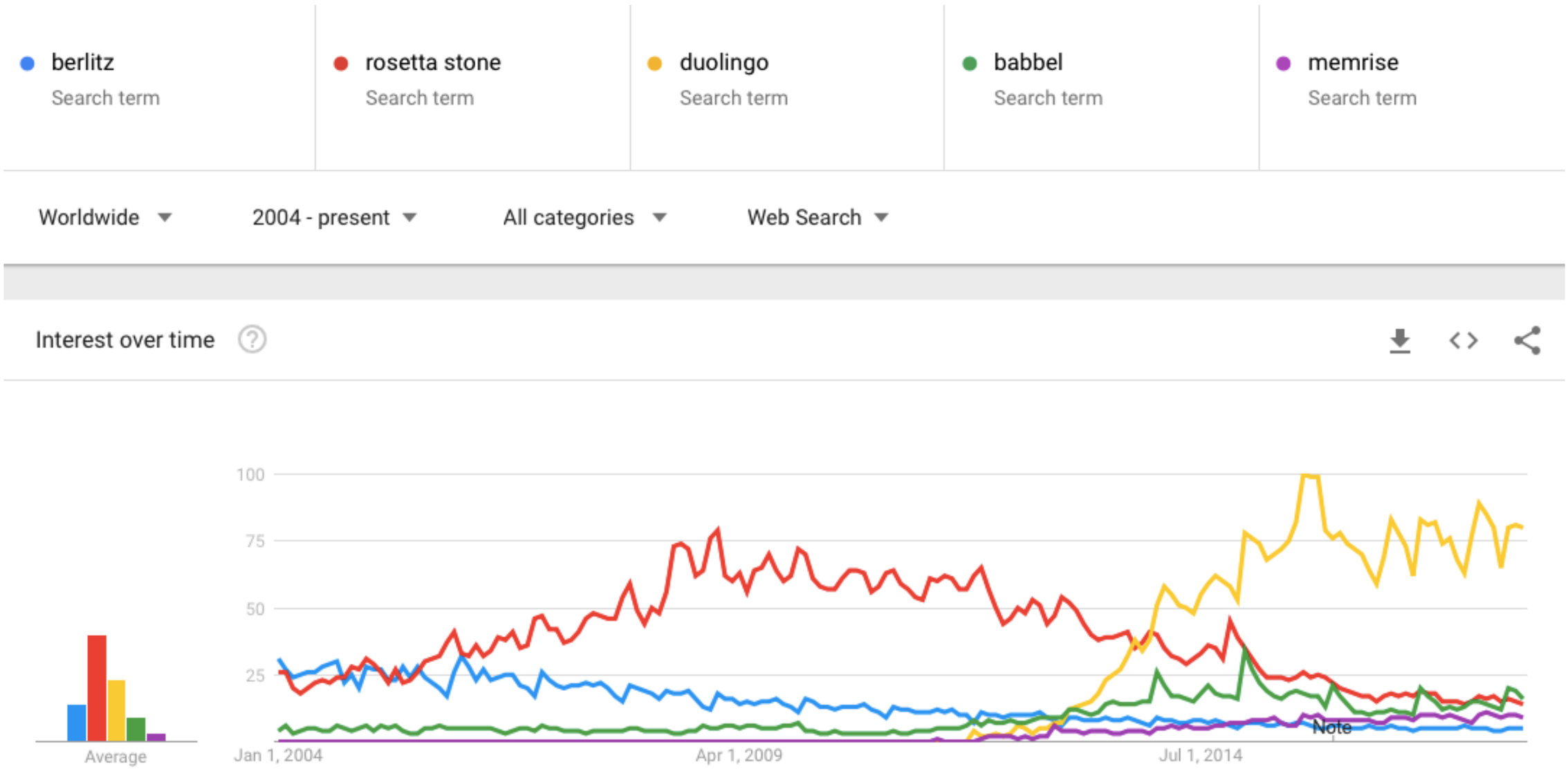
Data collection
Balancing

	Current	Goal	Balancing
Prompts (word or phrase)	700	7000	Vocabulary grade level (eg. A1, A2, B...)
Recordings	30 per prompt	60 per prompt	Requires both good, completely wrong, and marginal
Transcripts	4 per recording	8 per recording	Beware of corruption from lazy and other defectors
Exemplary recordings	15 per prompt (40 words)	*s**4 per prompt? (2 gender x 2 age)	*s* needs to be large enough for balancing recordings

Schema:

Users,
Authenticators,
Words,
Utterances,
Prompts,
Topics,
Choices,
Lessons,
Schools.

(6) Proposal TBD, e.g. third party restricted market sale free from ongoing cost center resource drains.



Questions? Email:

jim@talknicer.com