README

June 19, 2014

Contents

Introduction

The Arabic Corpus of Auditory Dictation Errors (ArCADE) is a corpus of Arabic words as transcribed by 62 native English speakers learning Arabic. This corpus is designed to assist researchers in investigating non-native spelling errors in Arabic, and particularly for spelling errors due to listening difficulties. Unlike error corpora collected from non-native Arabic writing samples, it is designed to elicit spelling errors arising from perceptual errors. A principal purpose for creating the corpus was to aid in the development and evaluation of tools for detecting and correcting listening errors to aid in dictionary lookup of words learners encountered in spoken language (cf. Rytting et al., 2010).

The ArCADE corpus was created through an elicitation experiment, similar in structure to an American-style spelling test. The principal difference (other than the language) is that in this case, the participants are expected to be unfamiliar with the words, and thus forced to rely on what they hear in the moment, rather than their lexical knowledge. Participants listened to 261 words presented over headphones and wrote their responses to the audio stimuli on a response sheet that contained numbered boxes. They were asked to use Arabic orthography with full diacritics and short vowels (*fatha*, *damma*, *kasra*, *shadda,* and *sukun*).

While the stimuli words were specifically chosen to facilitate the study of non-glide consonants (for which the mapping between orthography and phonology is relatively straightforward), we hope that the corpus will prove useful for studies beyond its original design.

Copyright

License

Please see the LICENSE file for how to license this corpus. If you have any questions, contact the University of Maryland (UMD) Office of Technology Commercialization (OTC):

Office of Technology Commercialization
2130 Mitchell Building
University of Maryland
College Park, MD 20742
Phone: 301-405-3947 | Fax: 301-314-9502
Email: umdtechtransfer@umd.edu
http://www.otc.umd.edu/

How to cite

When referencing this corpus, please cite the following paper:

- MLA:

Rytting, C. Anton, Paul Rodrigues, Tim Buckwalter, Valerie Novak, Aric Bills, Noah H. Silbert, and Mohini Madgavkar. "ArCADE: An Arabic Corpus of Auditory Dictation Errors." *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*. 2014.

- APA:

Rytting, C.A., Rodrigues, P., Buckwalter, T., Novak, V., Bills, A., Silbert, N.H., & Madgavkar, M. (2014, June). ArCADE: An Arabic Corpus of Auditory Dictation Errors. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*.

- BibTeX:

```
@inproceedings{rytting_etal:2014bea,
  title={ArCADE: An Arabic Corpus of Auditory Dictation Errors},
  author={C. Anton Rytting and Paul Rodrigues and Tim Buckwalter and
Valerie Novak and Aric Bills and Noah H. Silbert and Mohini Madgavkar},
  booktitle={Proceedings of the Ninth Workshop on Innovative Use of NLP
for Building Educational Applications},
  year={2014}
}
```

List of files in the ArCADE data set

- README -- this file

- `LICENSE` – a copy of the license for this corpus
- `Arcade.xml` – the digitized corpus of stimuli, responses, and alignments
- `audio/*.wav` – audio files for the stimuli
- `scanned_responses/*.pdf` – participants' handwritten responses

Description of the Files

XML database (Arcade.xml)

The XML database is the main file for the ArCADE data set. It contains information about the original stimuli (including links to the audio files), the orders the stimuli were presented, the participants in the study, and the written responses the participants gave to the stimuli (digitized by the collectors). It also contains some interpretations of those responses, including hypothesized phonological representations of what the participants likely heard (given their written responses) and segment- or character-level alignments between their responses and the original stimuli.

Types of elements

1) <stimulus> -- a stimulus word with descriptive information (e.g., orthography, pronunciation), a list of "target consonants" relevant to the original experimental design, and a link to the corresponding audio file used in the experiment.
2) <stimulus_set> -- a particular ordering of the stimuli as administered to the participants.
3) <subject> -- information about a particular participant (ID and stimulus_set) and a link to the PDF of the participant's handwritten responses.
4) <response> -- a participant's response to a particular word, with the digitized orthography, an automatically generated "transcription" of the most likely pronunciation the student had in mind, and edit operations of selected segments (the above-mentioned "target consonants").
5) <alignment> -- four different alignments between stimuli and responses:
   a) Phoneme-based alignment by Foma (Hulden, 2009), used for aligning target consonants and calculating edit operations in the <response> elements above. Since this alignment was not deterministic, multiple alignments exist for each stimulus-response pair.
   b) Three additional alignments using Phonetisaurus (Novak, 2011-2014): Phoneme-based, full orthographic, and orthographic without diacritics.

Examples and Explanations of the XML

<Stimulus>

There are 261 <stimulus> elements, one for each stimulus word in the original experiment. Each <stimulus> element contains an ID, the name of the corresponding audio file, and the orthography (<orth>) and pronunciation (<phon>) of the word. The orthography is given in Arabic script (utf8) with full diacritics; the pronunciation field is given using an ASCII transliteration based on the transcription used by Nizar Habash and Owen Rambow in their MAGEAD system (2006). Details of this transcription system are given in Appendix A.

Finally, each <stimulus> element contains additional details about selected consonants, or *target segments*, within each word. These target segments were chosen due to their appearance in particular contexts, descriptions of which are given in Appendix B. The positions of each target segment within the stimulus word are indicated in the "tagged" <orth> and <phon> fields; details about the phonological context are given in the <targetinfo> element associated with each target segment. In the following example, two consonants in the word /'aEViba/ (IPA: /ʔaʕðiba/) are targeted for analysis: /E/ (IPA: /ʕ/) and /V/ (IPA: ð/).

```
<stimulus id="159" audio="audio/aCdhiba.wav">
  <orth type="untagged">أَعْذِبَة</orth>
  <orth type="tagged">أَ<target_segment number="0">ع</target_segment>ْ<target_segment
        number="1">ذ</target_segment>ِبَة</orth>
  <phon type="untagged">'aEViba</phon>
  <phon type="tagged">'a<target_segment number="0">E</target_segment><target_segment
        number="1">V</target_segment>iba</phon>
  <targetinfo number="0" target_orth="ع" target_phon="E" context="V_C" emphasis="NoEmph" \
        preceding_environment="V" following_environment="C" gemination="false" \
        conditioning_emphasis="NA" conditioning_emphasis_is_coronal="NA"/>
  <targetinfo number="1" target_orth="ذ" target_phon="V" context="C_V" emphasis="EmphRightBefore" \
        preceding_environment="C" following_environment="V" gemination="false" \
        conditioning_emphasis="E" conditioning_emphasis_is_coronal="Non-coronal"/>
</stimulus>
```

<stimulus_set>

Each of the nine stimulus sets (numbered 2-10) specifies the order in which the participants heard the stimuli in the original experiment. The **number** attribute indicates the order of presentation; the **stimulus** attribute refers back to the ID of a particular <stimulus> element as described above.

<subject>

Each of the 62 <subject> elements gives the participant's ID code, the path to the participant's scanned responses, and the stimulus set ID which corresponds to the order in which the participant heard the stimuli.

<response>

The response elements contain the responses of the participants to the stimuli they heard.

Attributes:

- **stimulus**: the ID of the stimulus to which this entry corresponds
- **subject:** the ID of the participant who produced the response
- **stim_orth:** the orthographic representation of the stimulus (equal to the <orth> field of the corresponding <stimulus> element, provided here for ease of reference).
- **stim_phon:** the phonological representation of the stimulus' pronunciation (equal to the <phon> field of the corresponding <stimulus> element, provided here for ease of reference).

Child elements:

- **<orth>:** The orthography of the response dictation, as written down by the participant. Note that because not all participants supplied all diacritics as instructed, diacritics are not available for every entry. Only those diacritics which the participants actually wrote are given here; where no diacritic was indicated, none is recorded. Cases of non-response or missing data are coded as '~~~'.
- **<phon>:** The reconstructed phonological representation of what the participant heard, as derived from the orthographic representation. Details of how this was derived are given in Appendix A. Note that since the exact pronunciation of an Arabic word is only fully predictable (out of context) when the orthography is written with the full complement of diacritics, there were many instances where the phonological representation could only be partially reconstructed. In these instances, special symbols were used to indicate the ambiguities that arose, as follows:
    - The symbol '`[aiu]`' represents an instance where the participant's orthography implies a short vowel (/a/, /i/, or /u/) but provides no indication *which* short vowel was perceived. These are associated with 'bare alif' (Unicode u+0627). Instances where a short vowel is possible, but not clearly implied (e.g., the absence of a *sukun* between two consonants), are simply left blank.
    - The symbol '`[au]`' represents an instance where the participant's orthography implies either /a/ or /u/ but provides no indication *which* of these two short vowels was perceived. These are associated with 'Hamza on alif' (Unicode u+0623).
    - The symbols '`Y`' and '`W`' indicates the participant wrote a letter (yaa 'ي' or waw 'و', respectively) but did not provide sufficient diacritics to indicate whether the letter represented a long vowel, a diphthong, or a glide consonant.

In the following example, the participant failed to add any diacritics, with the result that the word's initial short vowel is unknown (thought the initial alif implies the participant heard one). The lack of diacritics on the yaa (ي) prevents assignment of the letter as a glide (/y/ -- IPA /j/), a diphthong (/ai/) or a long vowel (/i:/), so we use the special symbol 'Y' to encode the ambiguity explicitly. This results in the compact representation /`[aiu]ZYba`/ for a number of possible pronunciations.

```
<response stimulus="209" subject="60" stim_orth="أَعْذِبَة" stim_phon="'aEViba">
  <orth>اظيبة</orth>
  <phon>[aiu]ZYba</phon>
  <target_segment number="0" stim_phon="E">
    <possible_alignment id="0.1" response_phon="Z" operation="sub" alignment_probability="1.0"/>
  </target_segment>
  <target_segment number="1" stim_phon="V">
    <possible_alignment id="1.1" response_phon="Y" operation="sub" alignment_probability="1.0"/>
  </target_segment>
</response>
```

Note that none of these ambiguities affect non-glide consonants, which is why target segments are restricted to non-glide consonants.
As with the <orth> element, cases of non-response or missing data are coded as '~~~'.

- **<target_segment>:** the segments in the participant's response which correspond to the target_segments in the original stimulus. The determination of which segments in the response correspond to which target segments in the stimulus was made by a custom-built automatic alignment system, described in the section on alignments (below). Since the alignment system was non-deterministic, some target segments have multiple possible alignments. In these cases, a probability for each alignment is given, corresponding to the number of alignments in which this particular segment was aligned with the target segments, divided by the total number of alignments posited. The following example shows a case where the aligner returned four possible alignments. Two of the four alignments aligned the target segment /E/ (IPA: /ʕ/) in the stimulus with the corresponding /E/ in the response (a "match" edit operation), one alignment aligned the /E/ with the /Z/ (IPA: /$d^{ʕ}$/) in the response (a "substitution" edit operation), and one alignment found no corresponding segment, but aligned it to the empty string [epsilon] (a "deletion" edit operation).

```
<response stimulus="159" subject="52" stim_orth="أَعْذِبَة" stim_phon="'aEViba">
 <orth>عَظبة</orth>
 <phon>EaZba</phon>
 <target_segment number="0" stim_phon="E">
  <possible_alignment id="0.1" response_phon="E" operation="match" alignment_probability="0.5"/>
  <possible_alignment id="0.2" response_phon="Z" operation="sub" alignment_probability="0.25"/>
  <possible_alignment id="0.3" response_phon="[epsilon]" operation="del"
       alignment_probability="0.25"/>
 </target_segment>
 <target_segment number="1" stim_phon="V">
  <possible_alignment id="1.1" response_phon="a" operation="sub" alignment_probability="0.25"/>
  <possible_alignment id="1.2" response_phon="Z" operation="sub" alignment_probability="0.5"/>
  <possible_alignment id="1.3" response_phon="[epsilon]" operation="del"
       alignment_probability="0.25"/>
 </target_segment>
</response>
```

The possible values for the edit operations are as follows:

- "match" – meaning the stimulus and response symbols match exactly, both in type and length.
- "gem[ination]" – meaning the stimulus and response match in type, but the stimulus consonant was short and the response is long (i.e., the participant wrote a *shadda*).
- "degem[ination]" -- meaning the stimulus and response match in type, but the stimulus consonant was long (geminate) and the response is short (i.e., the participant failed to write a *shadda*).
- "sub[stitution]" – meaning the reference symbol and the response symbol do not match. Note that whenever two phonemes do not match in type (i.e., differ in any

feature other than gemination), they are scored simply as a <substitution>, even if they also differ in length.

- "del[etion]" – meaning a stimulus symbol is aligned with the empty string (i.e., a gap between two segments) in the response.
- "NA" – In the case of an entire word missing, a word that had no reasonable alignment or some other source of missing data, 'NA' is used to indicate that no edit operation can be assigned.

<alignment>

The alignment elements make explicit the alignments between stimuli and responses. There are four kinds of alignment elements: two based on the phonemic transcription and two based on orthography.

1) Phoneme-based alignments:
   a) The first kind of alignment (type= "phon" aligner= "foma") is a phoneme-based alignment created with a custom-built weighted edit distance model using Foma. This is the alignment used for aligning target consonants and calculating edit operations in the <response> elements, as shown in the example above. It was explicitly designed to favor the alignment of vowels and glides with other vowels and glides (or with empty string), so as to facilitate the analysis of non-glide consonants, without introducing other biases. Since this alignment was not deterministic, multiple alignments exist for each stimulus-response pair. Here is an example:

```
<alignment id="3539" reference="'aEViba" elicited="EaZba" type="phon" aligner="foma" stimulus="159">
 <substring reference="'" elicited="[epsilon]"/>
 <substring reference="a" elicited="[epsilon]"/>
 <substring reference="E" elicited="E" target_num="0"/>
 <substring reference="[epsilon]" elicited="a"/>
 <substring reference="V" elicited="Z" target_num="1"/>
 <substring reference="i" elicited="[epsilon]"/>
 <substring reference="b" elicited="b"/>
 <substring reference="a" elicited="a"/>
</alignment>
```

   b) The other phoneme-based alignment was induced automatically by training Phonetisaurus on the stimuli-response pairs of the <phon> representation. The symbols /`[aiu]`/, /`[au]`/, /`ai`/, /`au`/, and all geminate consonants and long vowels (i.e., a colon with its preceding character – e.g., /`i:`/, /`k:`/) were treated as single (multi-character) symbols. Any other groupings of characters found in the alignments are the result of Phonetisaurus inducing those groupings automatically.
2) Orthographic alignments
   a) The "full" orthographic alignment aligns the full orthography of the stimulus word (with all its diacritics) with the original orthographic response from the participant (with whatever diacritics the participant wrote on that word). No characters are grouped into multi-character symbols, except as automatically induced by Phonetisaurus.
   b) The "no diacritics" orthographic alignment strips all diacritics (short vowels, *sukun*, and *shadda*—Unicode range u+064B to u+0652) from the stimulus word and the orthographic

response from the participant. Characters with hamzas are left unchanged. As with the full orthographic alignment, no characters are grouped into multi-character symbols, except as automatically induced by Phonetisaurus.

In all four kinds of alignments, the special value "`[epsilon]`" indicates an empty string.

Audio Files (audio/*.wav)

In the audio directory we provide the audio files used as stimuli to elicit the dictations from non-native listeners. We include the 261 main items, but not the four practice items used at the beginning of the session.

The audio data used in the dictation study was recorded in a sound-proof booth with a unidirectional microphone (Earthworks SR30/HC) equipped with a pop filter, and saved as WAV files (stereo, 44.1kHz, 32-bit) with Adobe Audition.

The stimuli were recorded at four delivery speeds so that the general difficulty level of the task could be evaluated in the pilot phase. The recording selected for the study was the "medium fast" version, or third fastest speed, because it sounded the most natural and fluent. The audio files were segmented and normalized with respect to peak amplitude with Matlab.

The native Arabic speaker in the audio recording is of Egyptian and Levantine background, but was instructed to speak with a neutral ("BBC Arabic") accent. The speaker was selected for this task because of her demonstrated abilities to speak a neutral accent.

Handwritten Responses (scanned_responses/*.pdf)

Because not all students practice typing Arabic, we designed a handwritten response sheet. Participants handwrote their responses to the stimuli, and all response sheets were typed up for analysis purposes. Due to the complexity of reading unfamiliar handwriting and general typing errors, some data entry errors were introduced into the data. A review of the errors introduced in typing up the response sheets was completed. These data are available upon request.

Participants 54, 55, 58, and 59 apparently left off the last page worth of responses, and so nine responses from each of them are missing. (In arcade.xml, these are marked with '~~~' and 'NA' as described above.)

References

Habash, Nizar, & Rambow, Owen. 2006. MAGEAD: a morphological analyzer and generator for the Arabic dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 681-688). Association for Computational Linguistics.

Hulden, Mans. 2009. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session* (pp. 29-32). Association for Computational Linguistics.

Novak, Jozef. 2011-2014. Phonetisaurus: A WFST-driven Phoneticizer. http://code.google.com/p/phonetisaurus.

Brustad, Kristin, Al-Batal, Mahmoud, and Al-Tonsi, Abbas. 2004a. *Alif baa*. 2nd ed. Georgetown University Press, Washington, DC.

Brustad, Kristin, Al-Batal, Mahmoud, and Al-Tonsi, Abbas 2004b. *Al-kitaab fii ta'allum al-'arabiyya.* 1st ed., volume 1. Georgetown University Press, Washington, DC.

Rytting, C. Anton, Rodrigues, Paul, Buckwalter, Tim, Zajic, David M., Hirsch, Bridget, Carnes, Jeff, Lynn, Nathanael, Wayland, Sarah, Taylor, Chris, White, Jason, Blake, Charles, Browne, Evelyn, Miller, Corey, & Purvis, Tristan. 2010. Error Cor-rection for Arabic Dictionary Lookup. In *Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

Appendix A: Transcriptions of the Pronunciation (phon) fields

(IPA)	Magead	Arabic	Alternate characters
ʔ	'	ء	آ,إ,أ,ؤ,ئ
aː	A	ا	ى
a	a	*fatha* (◌َ)	ة
b	b	ب	
t̪	t	ت	ة followed by *tanwin*
θ	v	ث	
ʤ	j	ج	
ħ	H	ح	
x	x	خ	
d̪	d	د	
ð	V	ذ	
r	r	ر	
z	z	ز	
s	s	س	
ʃ	c	ش	
sˤ	S	ص	
dˤ	D	ض	
tˤ	T	ط	
ðˤ	Z	ظ	
ʕ	E	ع	
ɣ	g	غ	
f	f	ف	
q	q	ق	
k	k	ك	
l	l	ل	
m	m	م	
n	n	ن	*tanwin* (◌ً, ◌ٌ, ◌ٍ)
h	h	ه	
w	w	و	
u	u	*damma* (◌ُ)	
uː	U	◌ُو	
i	i	*kasra* (◌ِ)	
iː	I	◌ِي	word finally ◌ِيّ, ي, يّ
a͡ɪ	ai	◌َي	
a͡u	au	◌َو	
Abstract Symbols			
a, i, or u	[aiu]	initial ا	
a or u	[au]	initial أ	
geminate consonant	:	*shadda* (◌ّ)	ال before a sun letter[1]
j, iː, or a͡ɪ	Y	ي	
w, uː, or a͡u	W	و	
Special combinations			
ʔaː	'A	آ	
ʔa	'a	أ	
ʔu	'u	أ	ؤُ
ʔi	'i	إ	ئِ
iːj[2]	Iy	◌ِيّ	
a͡ɪj[2]	aiy	◌َيّ	
uːw[2]	Uw	◌ُوّ	
a͡uw[2]	auw	◌َوّ	

[1] In words الناس, التعاطف, التَّشجيع .

[2] Following ALA-LC, we interpret geminate glides as long vowels (or diphthongs) plus a single glide: thus `/Iy/` (IPA /iːj/) rather than `/iy:/` (IPA /ijː/).

Appendix B: Description of the Stimuli Selection Process and Target Segment Contexts

In order to create a corpus of auditory errors across a wide range of Arabic consonants as efficiently as possible, we devised an elicitation experiment similar to a spelling test of the type one might encounter in elementary school. The principal difference between our methodology and a spelling test is that in this case, the participants are expected to be unfamiliar with the words, and thus forced to rely on what they hear in the moment, not what they have studied beforehand. To that end, we selected words from a commonly-used dictionary of Modern Standard Arabic such that the set of words would contain a complete set of non-glide consonants in various phonetic contexts. The selection of words (and target consonants within those words) was subject to several constraints and preferences. In general, we used only citation forms of words (i.e., words found in the dictionary without additional morphological construction). In order to keep the stimulus list as short as possible while maintaining coverage of the full set of target stimuli in each targeted context, we chose words with multiple suitable target consonants over words with fewer suitable target consonants.

Furthermore, consonants that were morphologically predictable (e.g., non-root consonants in words of Semitic origin, when roots were known; also consonants participating in a reduplicative pattern such as *tamtam* and *zilzala*) were excluded from the list of target segments. Doubled Roots (e.g., where the second radical is the same as the third radical) were allowed as target segments if they were realized as a single geminate consonant, but not if the two consonants surfaced separately (e.g., in broken plurals such as أسنان /*asnan*/). In addition, we excluded words from our stimuli list if we anticipated the spelling to be too easy of a task for an introductory Arabic student. Items found in vocabulary lists associated with two commonly-used introductory textbooks *(Alif-Baa* and *Al-Kitaab*) were excluded (Brustad, Al-Batal & Al-Tonsi, 2004; Brustad, Al-Batal & Al-Tonsi, 2004). Items that were obvious loanwords from Western languages, and place-names known to most American students (e.g., *Scotlandia* = "Scotland") were also excluded. Terms that might be offensive or otherwise distracting were removed, as well. These exclusions were evaluated manually with the help of a native speaker.

Each Arabic consonant other than the glides /w/ and /j/ occurred as a target consonant for the analysis of perceptual confusions. The stimulus list was constructed such that target consonants occurred in each possible combination of six consonant/vowel/word-boundary contexts (hereafter referred to as "C/V contexts") and three contexts with respect to neighboring emphatic consonants (hereafter referred to as "emphatic contexts"). The six C/V contexts are as follows, with C = consonant, V = vowel, # = word boundary, and '_' (underscore) = location of target consonant:

- C_V,
- V_C,
- V_C,
- #_V,
- V_#,
- C_#.

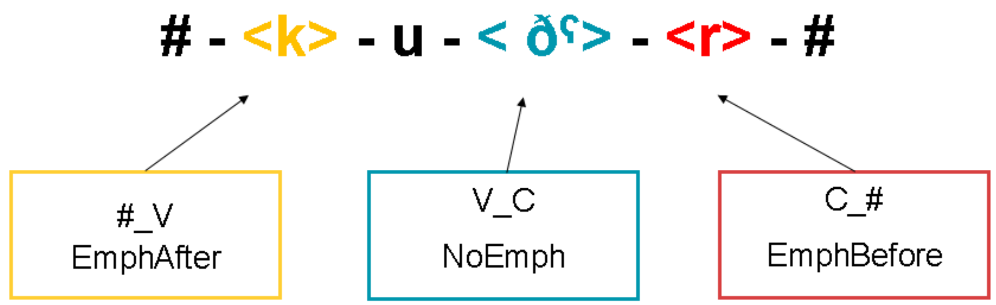
The three emphatic contexts are as follows, with C_E = emphatic consonant, V = a vowel (long or short) or a diphthong, C_P = plain (non-emphatic) consonant, a question mark indicates optionality, and [^C_E]* indicates any number of non-emphatic phonemes:

- EmphBefore or 'B' context: /C_EV?_V?C_P?/
- EmphAfter or 'A' context: /C_P?V?_V?C_E/
- NoEmph or 'N' context: /[^C_E]*_ [^C_E]*/ (i.e., no additional emphatic consonants in the word).

No other emphatic consonants occurred in words with contexts in which an emphatic consonant preceded or followed a target consonant, and no target consonants were both preceded and followed by emphatic consonants. The four standard emphatic consonants – i.e., /sˤ/, /dˤ/, /tˤ/, /ðˤ/ (ص, ض, ط, ظ)– as well as /ʕ/, /ħ/, and /q/ (ح, ع, and ق) were all considered emphatic consonants for the purposes of defining the three emphatic contexts. Consonants outside of one of these three specific contexts were excluded from analysis (i.e., not allowed as a target segment).

Note that the six C/V contexts and the emphatic contexts cross-cut each other, such that there are 15 possible combinations of the two context types (6×3 contexts minus the three logically impossible combinations: namely, [#_V × 'B']; [V_# × 'A']; and [C_# × 'A']). The geminate consonants constitute a sixteenth context, always occurring as [V_V × 'N']. The following table shows these contexts in a single grid, with the example of the stimulus word /kuZr/ (IPA: /kuðˤr/) with its three target segments:

	#_V	C_V	V_V	V_C	V_#	C_#
EmphAfter	#_VC_E …	…C_P _VC_E…	… V_VC_E …	… C_P V_C_E…	N/A	N/A
NoEmph	#_VC_P …	…C_P _V…	…V_V…	…C_P V_C_P…	…C_PV_#	…C_P_#
NoEmph (Gem)	N/A	N/A	…V_:V…	N/A	N/A	N/A
EmphBefore	N/A	…C_E _V…	…C_E V_V…	…C_E V_C_P…	…C_E V_#	…C_E_#


- <k> - u - < ðˤ> - <r> -

#_V
EmphAfter

V_C
NoEmph

C_#
EmphBefore

This method of stimuli selection yielded a word list of 261 words including a total of 649 instances of target non-glide consonants: one instance of each geminate consonant (except for /x:/, which was inadvertently omitted) and between 17 and 50 instances of each singleton consonant – at least one instance for each of the 15 contexts, with a few missing contexts: (1) [C_V × 'A'] for /ʔ/ and /θ/; (2) [C_V × 'B'] for /j/ and /z/; (3) [C_# × 'B'] for /h/, (4) [V_# × 'N'] for /f/, and (5) [#_V × 'A'] for /z/.