# Testing Listeners' Use of Acoustic Detail in an Artificial Language

Michael Hahn

LINGUIST 245, Spring 2017

Humans constantly make predictions about upcoming linguistic input, integrating diverse knowledge sources such as grammar, world knowledge, frequency, and acoustic detail. This research aims to address the question how humans develop this ability. In this project, I pilot an experiment that investigates how people acquire the ability to make real-time predictions when they learn a new linguistic system. More specifically, the goal of this experiment is to determine whether participants learning an artificial language will learn to incrementally recognize words based on coarticulatory information.

## 1 Introduction

A prominent example for linguistic predictions is the use of coarticulatory information in the speech stream, which can give information about the next segment before its onset [Hardcastle, 2006]. A large body of research shows that human speech perception make use of coarticulation (e.g., Martin and Bunnell [1981], Whalen [1984]).

This research attempts to address the question of how humans acquire the ability to predict based on fine acoustic detail by exposing participants to an artificial language and using eye-movement data to detect predictions made based on acoustic detail.

The experiment uses the visual-world paradigm. Participants see four images, displaying four distinct objects, and hear an auditory stimulus, naming one of the four objects in the artificial language. In one condition, the auditory stimulus contains coarticulatory information that disambiguates the

referent before the onset of the identifying word. In the other condition (baseline condition), this acoustic information is absent. Each participant is assigned to one of the conditions. Eye movements are tracked to detect when participants start looking at the named object.

**Linking Assumption and Predictions** Following Salverda et al. [2014], I assume that participants, across the conditions, attempt to identify the target as soon as possible. Furthermore, I assume that, across both conditions, there is a fixed amount of time after which identification of the target lead to increased fixations on the target. In particulary, this amount of time is expected to be about 200 ms, as Salverda et al. [2014] found.

Based on this, I assume that participants utilize anticipatory coarticulation if and only if fixations to the target occur earlier in the condition with coarticulatory information than in the baseline condition.

I predict that, given sufficient exposure to the artificial language and the task, listeners will indeed be able to utilize anticipatory coarticulation for predicting the target. Therefore, it is predicted that, given enough exposure to the language and task, fixations to the target will occur later in the baseline condition.

**Relation to Literature** The experimental paradigm and setup is based on Salverda et al. [2014], who studied the use of anticipatory coarticulation in English speech perception. Previously, Dahan et al. [2001] already used the visual-world paradigm to demonstrate the use of anticipatory coarticulation for early disambiguation.

The experiment also relates to previous work where participants acquired phonetic or phonological patterns when being exposed in an artificial language (Peperkamp and Dupoux [2007], Finley and Badecker [2009], Cristia et al. [2013], Gallagher [2013]). The experiment described here seems to be the first study of the acquisition of fine-grained speech perception and online processing in an artificial language learning setting. The study also relates to work on *adaptation* in speech perception, such as work on perception of foreign accents [Sumner and Samuel, 2009].

# 2 Methods

## 2.1 Participants

Two male native-speakers of American English participated in this pilot. Both were graduate students at Stanford.

## 2.2 Materials

**Artificial Language** The artificial language was generated by applying the following substitutions to initial consonants of English nouns:

```
/r/ -> /s/
/l/ -> /f/
/s/ -> /r/
/f/ -> /l/
```

That is, liquids were exchanged with fricatives. As Salverda et al. [2014] explain, coarticulatory information distinguishes strongly between liquids and fricatives. The substitutions were applied to phonemes of American English, independently of the orthographic form. No words with initial consonant clusters or initial vowels appeared in the experiment.

**Stimuli** To create **visual stimuli**, I used the images described by Rossion, B. and Pourtois, G. [2004], downloaded from `http://wiki.cnbc.cmu.edu/images/SVLO.zip`. These are a color adaptation of the classical Snodgrass-Vanderwart images used by Salverda et al. [2014].

For the **auditory stimuli**, a list of 52 words of the artificial language was created. All words were chosen from the list of words accompanying the available visual stimui. To the extent that this was compatible with the other constraints on the stimulus set, only words for which Rossion, B. and Pourtois, G. [2004] reported an agreement rating greater than 0.9 were used.

For these words, auditory recordings were created by a female native speaker of English. Each word was read preceded by the definite article, e.g. 'the sench'. As explained by Salverda et al. [2014], the unstressed schwa of the article is likely to show coarticulatory effects.

Artificial language words were presented in a form were the orthographic form representing the first consonant was replaced with the letter corresponding to the replacing phoneme (e.g., 'wrench' became 'sench'). The subject

was instructed to read each word as if it were a natural English word, and to change only the initial part but leave the pronunciation rest of the word unchanged. Each word was presented together with the intending corresponding English form. The subject went through the list twice and read each word at least two times.

From the recordings from the second session, one recording was selected for each word occurring in the first block, and a second recording was selected for each word also occurring in the second block. That is, no recording occurred in both blocks. This ensures that, in the critical trials, participants cannot identify the target based on the identity of the recording. Care was taken to ensure that selected stimuli had no additional noise and were similar in volume and speed of pronunciation.

The onset of the word-initial consonant after the article was annotated using Praat [Boersma, 2001]. Following Salverda et al. [2014], fricatives were annotated based on the beginning of high-frequency frication, and liquids as the point where spectral energy of F3 and F4 drops. As they highlight, annotation can only be approximate. After this pilot stage of my experiment, it will therefore be necessary to have two other researchers annotate onsets and assess interannotator agreement. All sound files were trimmed at the beginning so that the consonant begins 500ms after the beginning of the sound file.

The presence and degree of coarticulation in the article is illustrated in Figure 1. It is expected that this acoustic information is prominent enough to be utilized by participants.

**Items** There is a total of 128 items. This includes 104 practice items, 16 critical items, and 8 filler items. Each item consists of four English words with associated pictures.

The **practice items** were selected such that each of the 52 items from the list is a target in one item and a non-target in one other item. Each practice item contains two elements from the list and two elements not in the list. No pair of two words of the list occurs in two items.

The four words in the **critical items** started with a liquid, a fricative, a velar stop, and a bilabial stop, respectively. Each combination of consonants ((l, f, g, p), etc.) occurs in exactly one item. The target varies between subject and is either the word with a liquid or the word with a fricative. Each critical word is a target for one of the two subjects. For each subject,

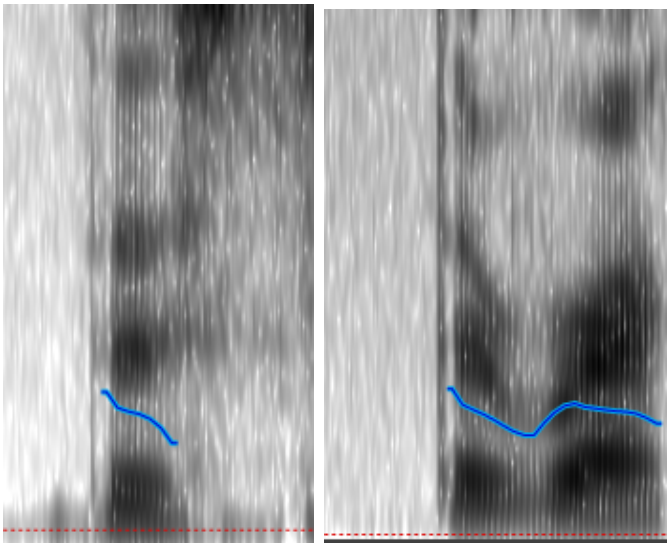

Figure 1: Extracts from recordings of ‘the saccoon’ (left) and ‘the ruitcase’ (right). Both extracts show the article and the first consonant of the noun, which the schwa of the article in the center. It can be seen that the F3 of the schwa is constant before /s/ but goes down in front of /r/, in line with what Salverda et al. [2014] report.

| | | | |
|---|---|---|---|
| ladder | fish | grapes | pineapple |
| lamp | fox | kite | potato |
| leaf | foot | glove | bird |
| leg | seahorse | gun | pipe |
| leopard | sailboat | gorilla | bowl |
| lightbulb | sandwich | comb | beetle |
| lips | football helmet | kettle | bee |
| lobster | cigar | kangaroo | penguin |
| rabbit | fork | cow | bike |
| raccoon | scissors | corn | pear |
| record player | football | carrot | pepper |
| roller skates | fence | glass | basket |
| ring | celery | couch | bike |
| rolling pin | finger | garbage can | bee |
| ruler | cigarette | goat | pants |
| wrench | suitcase | guitar | ball |

Figure 2: Critical Items. Each row denotes an item. The first two columns consist of words starting with a fricative or liquid in English and in the artificial langueg. The second two columns consist of distractor words, whose form is identical in English and the artificial language.

each fricative-liquid combination is the target in exactly two out of four items. No word appears in two critical items. No pair of two critical words appeared both in a critical item and in a practice item.

Each **filler item** consists of four other words, two of them from the list and two not in the list. None of the words start with any of the consonants from the critical items. No word appears in two filler items, or a filler item and a critical item.

Each occurrence of a word in the training block used the same auditory stimulus. However, the second block uses different recordings. This ensures that subjects cannot make use of the identity of the recording (e.g., the exact articulation of the article).

### 2.3 Procedure

Participants were told that they were part of a crew of scientists and that their task was to learn an alien language. They were then instructed that they would see four images and had to click on the image that they thought the alien was referring to.

On each trial, the screen displayed the four images belonging to the item. After 1s, an auditory stimulus was played. Participants were informed that they should click on the image showing the item named by the auditory stimulus. The screen went blank as soon as the participant clicked on the screen. After 2.5s, the next trial started. This procedure follows Salverda et al. [2014].

After an initial block of 104 training items, the eyetracker was calibrated and participants entered the second block, which consisted of the critical and the filler items. Training items were presented in randomized order, identical for both participants, subject to the constraint that no two successive items shared a word. Critical and filler items were presented in random order, randomized for each participant, within a single block. In the critical items, the position of the four pictures was counterbalanced such that, across displays, the target and the distractor each appeared in each place the same number of times. In the other items, the position of the four pictures was randomized, identical across participants.

## 3 Results

For each trial, fixations to the four regions were aggregated into bins of 10ms each, starting at the onset of the initial consonant. Figure 3 shows the proportion of trials with fixations on targets, competitors, and the two distractors in each of these bins. On the whole, fixations to the target appear to increase over time, while fixations to other regions appear to decrease at least after 600 ms.

Figure 4 shows a logistic mixed-effects model predicting for each trial, each participant, and each bin, whether a fixation to the target occurred, with item as random effect. The model shows that fixations to the target did indeed increase significantly over time. Furthermore, the two participants differed both in total number of bins with fixations and in the speed with which fixations to the target increased.

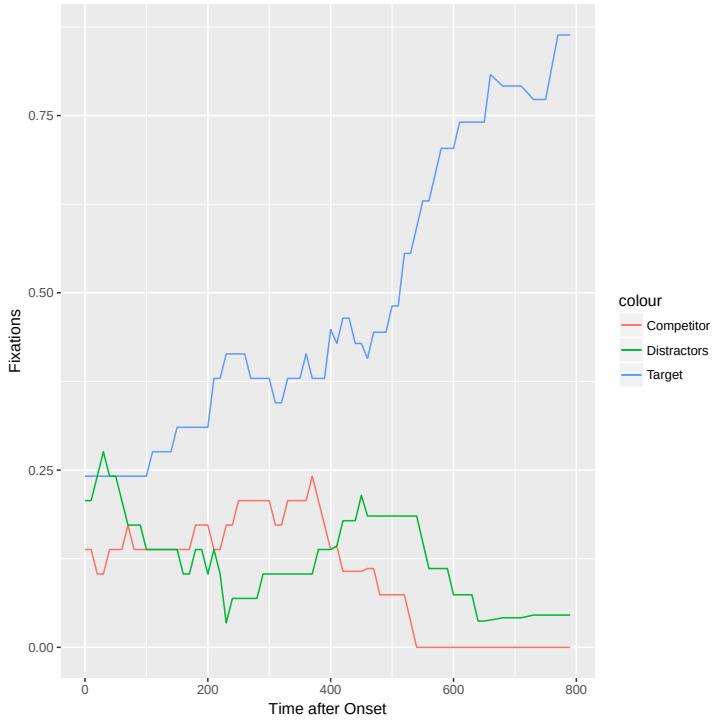


Figure 3: Proportion of trials with fixations on target, competitor, and other images after onset of the initial consonant.

| | $\beta$ | SE | z value | Pr($>\|z\|$) |
|---|---|---|---|---|
| (Intercept) | -0.09308 | 0.33756 | -0.276 | 0.78276 |
| StartOfBin | 4.59892 | 0.24445 | 18.814 | $<$ 2e-16 |
| Participant | 0.05867 | 0.01906 | 3.079 | 0.00208 |
| StartOfBin:Participant | -0.85627 | 0.07294 | -11.740 | $<$ 2e-16 |

Figure 4: Mixed-Effects Model predicting fixations to the target over time. Predictors are centered. The model was computed using lme4 [Bates et al., 2015].

| | $\beta$ | SE | z value | Pr($>|z|$) |
|---|---|---|---|---|
| (Intercept) | -1.0456 | 0.4595 | -2.276 | 0.0229 |
| First Contrast (after 100ms) | 0.7267 | 0.5822 | 1.248 | 0.2120 |
| Second Contrast (after 200ms) | 0.5834 | 0.6311 | 0.924 | 0.3553 |

Figure 5: Results of logistic model predicting early fixations to target

To assess whether fixations to the target occurred early after the onset of the consonant, I fitted a model comparing the time-windows [0ms–100ms], [100ms–200ms], and [200ms–300ms], predicting, for each trial, whether a fixation would fall on the target. The second time window corresponds to the expected time in which fixations to the target would occur above chance when using anticipatory coarticulation, but not in the baseline condition, based on the findings of Salverda et al. [2014]. The third time window corresponds to the earliest time when fixations to the target should occur above chance in both conditions. The model has a random intercept for the item. The windows are Helmert-coded, with the first contrast comparing [0ms–100ms] to [100ms–300ms], and the second contrast comparing [100ms–200ms] to [200ms–300ms].

The resulting logistic model is shown in Figure 5. The intercept is negatively significant, showing that throughout the time windows, fixations to the target are reliably rarer than 50 %. No significant difference was shown between the time windows. Thus, nothing can be concluded about whether fixations to the target increase reliably before 300ms. Similar models with more windows, comparing later times, did not converge in the algorithm implemented by lme4.

I also conducted a moving-window analysis following Salverda et al. [2014]. For each sequence of four successive bins (i.e., spans of length 40 ms) I counted the number of trials with fixations on the target and on the competitor. For each such 40ms-window, I then computed the $\chi^2$-statistic. The result is shown in Figure 6. The critical statistic is $\chi^2 = 3.84$ at $p = 0.05$. The earliest bin in which the statistic exceeds this value starts at 520ms after the onset of the consonant. Thus, this analysis also does not show that fixations to the target reliably occurred in the expected early time windows.

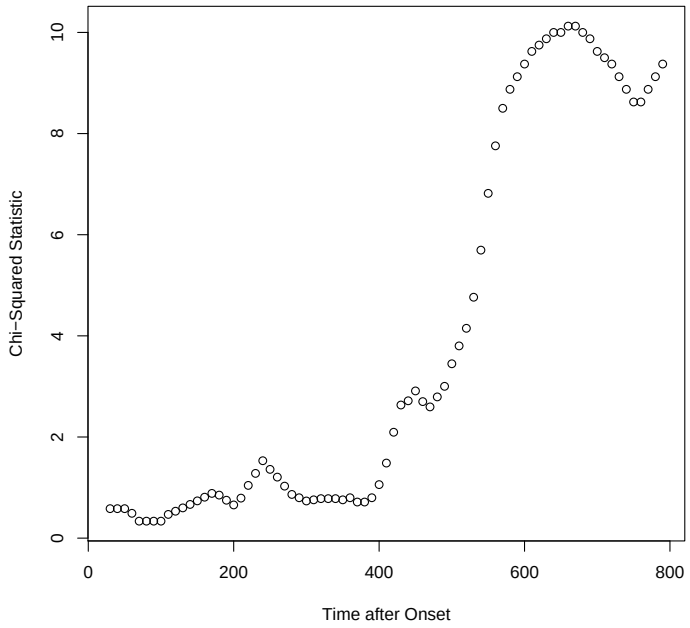


Figure 6: $\chi^2$-statistic for the numbers of fixations on the target and on the distractor after onset.

# 4 Discussion

As this is only a pilot with a very small number of participants, no decisive conclusions can be drawn from the results.

## 4.1 Need to Measure Baseline

This project only included a pilot study to assess doability of the experiment and whether an effect might possibly be observed in the time-window described by Salverda et al. [2014]. In future research, it will be necessary to also obtain a baseline for the first time when auditory information affects eye movements given the stimulus set used. This is necessary in particular given that the speed of processing in an artificial language might be much slower than speed of processing the native language.

The procedure will follow Salverda et al. [2014]. I will create a second set of auditory stimuli of the form 'The... (NOUN)', with a prosodic pause after the article, pronounced by the same speaker. This speaker will also produce several instances of the article 'the' in isolation. One of these instances will then be cross-spliced on the stimuli. This ensures that the article includes no acoustic information about the noun.

## 4.2 Future Work

One participant remarked that he was puzzled that the 'alien' language was surprisingly close to English. In future experiments, it might perhaps be preferable to instruct participants that they are listening to a person or an alien who is speaking English but mixes up some sounds.

A potential confounding factor might be that anticipatory coarticulation might contain information about the first vowel in the noun in addition to the first consonant. In this pilot, the critical items were designed so that target and competitor never have the same first vowel. As an alternative, it might be necessary to cross-splice recordings or synthesizing recordings to ensure that only information about consonants is available.

An important feature of the critical items is that liquids and fricatives are *exchanged*, relative to English. This means that prediction using knowledge of English might constitute a strong bias counteracting an effect of prediction in the artificial language. On the one hand, this means that an effect in the predicted direction could not be explained by participants' knowledge

of English, as it would go into the oppoite direction in that case. On the other hand, this design might put the bar for detecting the predicted effect unnecessarily high. Future work might use critical items where influence from English would neither reinforce nor counteract predictions made based on the artificial language.

# References


Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

Paul Boersma. Praat, a system for doing phonetics by computer. *Glot International*, pages 341–345, 2001.

Alejandrina Cristia, Jeff Mielke, Robert Daland, and Sharon Peperkamp. Similarity in the generalization of implicitly learned sound patterns. *Laboratory Phonology*, 4(2), January 2013. ISSN 1868-6354, 1868-6346. doi: 10.1515/lp-2013-0010. URL http://www.degruyter.com/view/j/lp.2013.4.issue-2/lp-2013-0010/lp-2013-0010.xml.

Delphine Dahan, James S. Magnuson, Michael K. Tanenhaus, and Ellen M. Hogan. Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, 16 (5-6):507–534, October 2001. ISSN 0169-0965, 1464-0732. doi: 10.1080/01690960143000074. URL http://www.tandfonline.com/doi/abs/10.1080/01690960143000074.

Sara Finley and William Badecker. Artificial language learning and feature-based generalization. *Journal of Memory and Language*, 61(3): 423–437, October 2009. ISSN 0749-596X. doi: 10.1016/j.jml.2009.05.002. URL http://www.sciencedirect.com/science/article/pii/S0749596X09000564.

Gillian Gallagher. Learning the identity effect as an artificial language: bias and generalisation. *Phonology*, 30(2):253–295, August 2013. ISSN 0952-6757, 1469-8188. doi: 10.1017/S0952675713000134. URL https://www.cambridge.org/core/journals/phonology/article/

learning-the-identity-effect-as-an-artificial-language-bias-and-generalisation/A902BA44C31EBBBC4E994202B26A406C.

W. Hardcastle. Coarticulation. In Keith Brown, editor, *Encyclopedia of Language & Linguistics (Second Edition)*, pages 501–505. Elsevier, Oxford, 2006. ISBN 978-0-08-044854-1. URL http://www.sciencedirect.com/science/article/pii/B0080448542005654. DOI: 10.1016/B0-08-044854-2/00565-4.

J.G. Martin and H.T. Bunnell. Perception of anticipatory coarticulation effects. *Journal of the Acoustical Society of America*, 69(2):559–567, 1981. ISSN 0001-4966. doi: 10.1121/1.385484.

Sharon Peperkamp and Emmanuel Dupoux. Learning the mapping from surface to underlying representations in an artificial language. *Laboratory phonology*, 9:315–338, 2007. URL http://www.lscp.net/persons/dupoux/papers/Peperkamp-Dupoux-2007-Learning-mapping-surface-underlying-artificial-language.LabPhon9.pdf.

Rossion, B. and Pourtois, G. Revisiting snodgrass and vanderwart's object set: The role of surface detail in basic-level object recognition. *Perception*, 33:217–236, 2004.

Anne Pier Salverda, Dave Kleinschmidt, and Michael K. Tanenhaus. Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language*, 71(1):145–163, February 2014. ISSN 0749596X. doi: 10.1016/j.jml.2013.11.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0749596X13001113.

Meghan Sumner and Arthur G. Samuel. The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language*, 60(4):487–501, May 2009. ISSN 0749-596X. doi: 10.1016/j.jml.2009.01.001. URL http://www.sciencedirect.com/science/article/pii/S0749596X09000096.

D.H. Whalen. Subcategorical phonetic mismatches slow phonetic judgments. *Perception & Psychophysics*, 35(1):49–64, 1984. ISSN 0031-5117. doi: 10.3758/BF03205924.