Sten Knutsen

Application for Interdisciplinary Thesis

Proposed Title: “Isolating the Elusive Verb Stem Vowel”

Primary Faculty Advisor: Karin Stromswold, Professor, Psychology

Secondary Faculty Advisor: Paul de Lacy, Professor, Linguistics

Introduction

As efficient speech processors, we are continually updating our understanding of utterances as new information rapidly streams towards us, encoded in a linear fashion as an acoustic signal. From this relentless acoustic stream we automatically and often unconsciously cull whatever pertinent information or cues are available to us and model complicated and often non-linear syntactic constructions from them at a subconscious level. These constructions form the basis of our mental representation of the utterance and our understanding of what it means. Each time we refresh our model based upon the appearance of some new explicit or implicit cue, predictions about upcoming syntactic structure may be made, and an initial construction may undergo significant revision based upon our new expectations. Ultimately, these revisions alter our mental representation and understanding of the sentence.

Recently, several studies have explored how listeners might use low-level acoustic cues to update their expectations regarding sentence structure. Stromswold, Kharkwal, Sorkin, & Zola (2013, under review) conducted an eye-tracking study to explore the use of durational cues in predicting the syntactic structure of temporarily ambiguous sentences. Participants were presented with sound recordings of both active sentences, e.g. *The girl was pushing the boy* and passive ones, e.g. *The girl was pushed by the boy*. At the same time, participants were shown two illustrations that might possibly relate to the sound recording: one that matched the correct understanding of the sentence, and a complement illustration which did not. Eye-tracking data revealed that participants were actually “choosing” the correct picture as they heard the verb stem — for example *push* — as opposed to “waiting” for what one would expect to be the disambiguating cue: the verb’s inflectional morphology (*-ing* or *-ed*). Carefully analyzing the acoustic nature of the recorded sentences revealed that verb stems of passive sentences were significantly longer than those in active sentences, leading the researchers to conclude that verb stem duration might provide listeners with a significant cue as to the ultimate syntactic structure of a sentence. Furthermore, post hoc analyses of this study's data revealed that the verb stem vowel itself was the source of this overall increase in verb duration. However, as this study was performed with linguistically trained participants (undergraduate linguistics students), a follow-up study was conducted by Rehrig et al. (2015) on linguistically naïve adults. This study confirmed the findings of Stromswold et al. (2013).

Following up on both Rehrig et al. (2015) and Stromswold et al. (2013), the production study Mayro et al. (2016) was carried out to further establish the existence of the lengthening effect found on verb stems of passive sentences and to confirm the verb stem vowel as the probable locus of this effect. Carefully constructed truncated passives, e.g. *The red apple was picked a couple of hours ago* (as well as several other syntactic constructions that were prepared by the author of the study) were produced by linguistically naïve, English-speaking participants. Analysis of the data showed that the passive verb stems participants produced were significantly longer than past active verb stems with vowel duration again emerging as the source of this lengthening.

Study Hypothesis, Goals

The aforementioned studies all converged upon the duration of the verb stem vowel as being primarily responsible for the passive/active verb stem duration difference. It follows from this that listeners could utilize this subtle acoustic cue in updating their mental model of an unfolding utterance. However, studies thus far have not determined whether listeners might also be gathering acoustic cues that are occurring “upstream” of the verb stem vowel. If such cues do exist, they might also provide a substantial contribution to the listener’s expectations (and therefore predictive judgments); and perhaps the verb stem vowel duration is merely the final, critical trigger. Therefore, the goal of this study is to isolate verb stem vowel duration as the *only* acoustic cue that listeners use to predict upcoming syntactic structure. In order to pin down the verb stem vowel as such, artificially manipulating naturally produced verb stem vowels and then presenting them in a controlled comprehension task might prove to be a fruitful venture.

Although current software makes the accurate editing of digital sound files in a computer easier than ever before, artificial manipulation of sound recordings (and more specifically, manipulations targeting verb stem vowels) for linguistic research is a decades-old practice. In their study of listener expectations regarding segmental duration, prepared their stimuli by first digitizing analog recordings of speech. Then, using a digital waveform editing program, they either deleted or reduplicated waveform segments to adjust the duration of a vowel before returning the modified speech sample to an analog tape for presentation to the participant. Current technology (e.g. Praat phonetic analysis software) allows similar control over segment duration manipulation. The software automatically employs the lengthening/shortening strategies employed in Klatt and Cooper (duplication/deletion of waveform segments) without the need to repeatedly convert between analog and digital recording mediums.

The comprehension study I am proposing seeks to artificially elongate recordings of naturally produced English active verb stem vowels and shorten passive ones. The altered sentences would then be presented to native English speaking participants (who are naïve to the purpose of the experiment) in their entirety. At approximately the same time, simple color illustrations corresponding to the possible passive and active constructions would be presented to the subject. An eye-tracking device would then capture one of two outcomes: the temporary confounding of the subject based upon the false cue (t); or the subject simply focusing on the correct picture (meaning some other cues are present in the acoustic stream).

METHODS

Participants

A total of 30 monolingual native English speakers will participate: a control group of 15 will perform the experiment with unaltered sound files; the balance will be presented with the artificially altered sentence stimuli. All participants will be screened for hearing, speech, language, learning or other neurological disorders. A record of my completion of the CITI human subjects certification is attached.

Stimuli preparation

Source files. The source material for this study will come exclusively from subject 205 of Rehrig et al. (2015). This material consists of WAV file recordings captured during the production phase of that study. The recordings were digitally recorded with high-quality recording equipment in a acoustically dampened environment.

In all, 56 sentences (28 active and 28 passive) recorded by subject 205 will be the basis for the experiment's stimuli. The sentences were constructed using a combination of fourteen verb stems paired with animal agents and patients. Below is a sample pair of sentences:

Progressive active: *The sheep was kissing the pig.*

Passive: *The sheep was kissed by the pig.*

Artificially modifying vowel duration in source sound files. In preparation for this study, I will consider the sentences in pairs (such as above) and measure their verb stem vowel duration. The files will then be manipulated in such a way that the two waveforms will appear to have "swapped" vowel duration. For instance, if the active verb stem vowel in the aforementioned *The sheep was kissing the pig* has a duration of 46ms and its corresponding passive form has a duration of 78ms, the files will be electronically altered so that the active vowel achieves a duration of ~78ms and the passive ~46ms.

Especially critical to the success of this study will be the manipulation of these existing files so as to lengthen active verb stem vowels and shorten passive ones without disrupting the surrounding acoustic content. This will be accomplished largely by using existing features of Praat phonetic analysis software.

First, each source file will first be passed through the Penn Phonetics Lab Forced Aligner (P2FA) which automatically annotates and sets the phonetic boundaries of the sound file, rendering the results in Praat TextGrid file format. Further refining of verb stem vowel alignment (as well of alignment of all of the sentence morpheme boundaries) will be conducted manually in Praat by multiple coders. Also using Praat, the portion of the vowel where it achieves a steady state will be isolated and the period of the signal's waveform at this point will be measured and annotated.[1] This will later play a role in the shortening/lengthening algorithm.

After isolating the vowel to be shortened/lengthened, the actual manipulation will be performed using Praat's Duration Tier object (a Praat feature which operates upon the duration of a sound file). The algorithm for creating the modified Duration Tier takes into account the vowel boundaries and signal period in calculating the new duration curve which most closely models the targeted duration time. Using a short Python program to make these calculations and create a Praat Duration Tier script, the new Duration Tier is imported into Praat, applied to the original sound file in the Praat Manipulation object, and exported as a new WAV file which reflects the changes made in vowel duration.

To ensure the targeted vowel has been lengthened/shorted accurately and that no anomalous conditions have been introduced into the file in the modification process, the new WAV file will be compared with previous Praat TextGrid information and boundaries will again be refined manually. The new vowel duration will be measured to make sure it has been stretched/shortened to the desired length. Upon confirming that no unwanted alterations have occurred in any part of the modified file, the file will be admitted into the study.

Picture Pairs. The two pictures that will appear onscreen with each sentence will depict the animals described in the sentence. The only difference between the two pictures will be which animal is acting as agent and which is acting as patient. The position of the correct picture on the left or right (as well as within-picture positions of agent and patient) will be pseudorandomized.

[1] Although natural speech signals are considered to be quasi-periodic, individual segments of these signals may be treated as periodic. This is particularly the case when such waveforms are sampled from the steady-state portion of a given vowel segment (Reetz & Jongman, 2009).

PROCEDURE

Comprehension trial summary. Participants will be seated in front of a laptop screen wearing high-quality headphones in a quiet, distraction-free environment. The subject will perform a brief eye-tracking calibration test using software provided by Eye Tribe. After calibration is complete, the eye tracker will start recording and the actual experiment will commence.

A fixation cross will appear on the screen for 250ms followed by the appearance of both the target and incorrect illustrations. After 250ms pause, a recording of the sentence (unmodified for the control group and artificially altered for the experimental group) will be played. Participants will respond by selecting the correct picture on the screen. Following their response, a blank screen is presented for 500ms followed by a prompt for the next trial.

Equipment/Software. The laptop will be a 13" Lenovo Yoga (or like computer). The eye-tracking device is the Eye Tribe Tracker which is calibrated using software provided by the Eye Tribe company. The software to record gaze results is a custom program written in the Java programming language and packaged as a Java application. The experiment itself will be created and run on E-Prime software.

Results analysis. Data recorded by the eye-tracking software and E-Prime will be reconciled using time-stamp information and the resultant data set will be analyzed using existing statistical techniques and software. The data should reveal how participants negotiate the unexpected "twist" provided by the artificially lengthened/shortened verb stem vowel and the subsequent resolution of the sentence meaning.

Work Responsibilities

I will perform the duration manipulation on the source files as described in the "Methods" section and implement them in E-Prime. I may also make adjustments the software I wrote to record Eye Tribe gaze data, if needed. After I have recruited participants and all data have been gathered, I will seek the assistance of my advisor(s) to perform statistical analysis of the data.

Start/End Dates

Work on the project begins on approximately May 16 and will end around September 10, 2016. Additional analysis of the data acquired in this study may continue throughout the 2016/2017 academic year.

Student Advisor Meeting Schedule

I will consult with my advisor(s) at least once a week for the duration of the project.

Previous Research Experience

In December of 2015, I was invited by Dr. Stromswold to work in her lab during the Spring 2016 semester. During that time I developed software that aids in the accurate manipulation of speech sound files. I also wrote software that records gaze data from an inexpensive eye-tracking device.

Personal Interest

Having never gone to college in the twenty or so years since I graduated high school, I started attending Rutgers University in the Fall of 2011 as a part-time, nontraditional student in pursuit of the wildly broad and alarmingly unrealistic goal of "exploring" cognitive science.

By the Spring of 2013, I turned my attention to the cognitive science subfield of linguistics, taking introductory courses in both linguistics and syntax. My syntax professor, Jane Grimshaw, strongly encouraged me to pursue my interest in linguistics and cognitive science as far as my circumstances would allow. So at the end of 2013 I separated from my full-time job and became all full-time student at Rutgers majoring in linguistics.

In the Spring of 2015, I declared a second major in another core subfield of cognitive science: computer science. This avenue of study has proved to be particularly relevant to my interest in cognitive science, providing a computational perspective on human language processing as well as practical tools with which to do research in the cognitive science of language.

After taking Dr. Stromswold's Psychology of Language course last fall, I was invited to participate in her lab. This opportunity allowed me to apply the knowledge and skills I had gained from linguistics, computer science and other relevant areas of study to solving problems and performing research in sentence processing. It is my hope that in further pursuing research in this interdisciplinary yet highly specialized area of cognitive science I will be able to make the "unrealistic goal" that I first set my sights on a reality.

References

Klatt, D,H., & Cooper, W. E. (1975). Perception of segment duration in sentence contexts. In A. Cohen, & S.G. Nooteboom (Eds.), *Structure and Process in Speech Perception* (pp. 69-89). Springer.

Mayro, A., de Lacy, P., Rehrig, G., & Stromswold, K. (2016). Locus and cause of passive verb stem lengthening. (unpublished honors thesis). Rutgers University, New Brunswick, NJ.

Reetz, H., & Jongman, A. (2009) *Phonetics: transcription, production, acoustics and perception.* Wiley-Blackwell, West Sussex, United Kingdom.

Rehrig, G., Beier, E., Chalmers, E., Schrum, N., & Stromswold, K. (2015). Robust acoustic cues indicate upcoming structure in active and passive sentences. Poster presented March 19, 2015. *The 28th Annual CUNY Conference on Human Sentence Processing.* Midtown Los Angeles Radisson Hotel, Los Angeles, CA.

Stromswold, K., Kharkwal, G., Sorkin, J.E., & Zola, S. (2013, under review). Tracking the elusive passive: The processing of spoken passives.