A generative model of context inferred from a prosodic signal

Abstract

We examine how listeners can infer and apply information about a talker's linguistic variation across different social contexts, using Child-Directed Speech (CDS) as an example of a social situation in which talkers often tailor their prosodic, phonetic, and lexical choices because of their addressee's social features. Early this semester, we conducted two pilot studies to establish a link between prosody and lexical prediction. Given the success of these pilots, we designed a generative model of this association as explained by the top-down effect of context. We have prepared a set of experiments to test the fidelity of the model to human behavior by collecting both on-line and off-line data.

Introduction/Background

In the context of language processing, the extraction of information from an acoustic signal is often described as a challenge: listeners must pick out linguistic information from a noisy communicative channel (Bicknell, Elman, Hare, Mcrae, & Kutas, n.d.; Gibson, Bergen, & Piantadosi, 2013; Jaeger & Tily, 2010; Levy, Reali, & Griffiths, 2009). There is variation in how different phonemes, syntactic phrases, and even propositional messages are realized between different social groups, individuals, and even different instances of utterances. However, if we recognize that much of this "noise" is the result of structured variation in the world, it is possible to treat it as a useful part of the communicative channel. If human minds are capable of tracking the patterns of correlation between linguistic variation and the variation of other contextual features, the "noise" inherent in acoustic linguistic signals can be leveraged to inform expectations about the "noisy" signal variation itself (Fine, Jaeger, Farmer, & Qian, 2013; Kleinschmidt & Jaeger, 2015). For example, if a person's productions of /s/ sounds similar to /θ/, one might infer that they have a lisp, and therefore anticipate that a /ð/ sound will occur where a /z/ sound might otherwise be expected, facilitating one's ability to interpret this particular talker's speech.

Previous research has investigated how listeners infer social information about a talker from the acoustic signal and how this can affect language processing across linguistic levels (Niedzielski, 1999; Van Berkum, van den Brink, Tesink, Kos, & Hagoort, 2008). It has also investigated how listeners and talkers use information on their visual and conversational contexts to make choices about which words to use and how to interpret them (Altmann & Kamide, 1999; Brown-schmidt, Yoon, & Ryskin, 2015; Tanenhaus, Spivey-knowlton, Eberhard, & Sedivy, 1995). In this proposal, we posit that listeners are using a rich set of information from both the context of an utterance and from the “noise” in the signal itself to infer a more holistic meaning of utterances. We focus on the aspect of meaning that a talker’s relationship to an intended addressee brings to an instance of linguistic communication.

We choose to work with Child Directed Speech, as it is a good testing ground to demonstrate the relationship between interlocutor context and linguistic signal. Its intended addressees are an easily–identifiable group with universally recognized characteristics. Experience with the register (we were all children once upon a time, and have at least heard others speaking to children) is near-universal. As researchers, we can also make clear predictions about reference and meaning, in part due to the fact that children have a much smaller set of knowledge about the world at their disposal. Furthermore, relative to any other identifiable register, CDS has the largest and most diverse body of documentation to refer to. We have access to information about its basic phonological, syntactic, and prosodic features (Cameron-faulkner, Lieven, & Tomasello, 2003; Foulkes, Docherty, & Watt, 2005; Warren-leubecker, Neil, & Bohannon, 2016). We even have supported theories as to what the function of the register is (Cooper & Aslin, 1990; Pegg, Werker, & Mcleod, 1992; Rowe, 2008). Furthermore, corpora are available (e.g., CHILDES) that relate directly to the register from which we can extract attested stimuli.

It has be established that listeners can extrapolate information about a talker’s social identity from features of the linguistic signal (Purnell, Idsardi, & Baugh, 1999; Staum Casasanto, 2009). Here, we aim to establish that listeners can also infer information about a talker’s intended *addressee* from a linguistic signal; specifically, from prosodic information.

Pilot Studies:

We ran two pilot studies to demonstrate the connection between prosody and lexical choices with regard CDS in language processing. We hypothesize that perception of acoustic (i.e. prosodic and phonetic) markers of a particular register will generate sets of lexical and/or morphological predictions for the listener. Specifically, we hypothesize that listeners will make implicit judgments about which register they are listening to based on prosodic cues in the signal and generalize expectations about this register to make lexical predictions, which they will indicate in the testing task. Specifically, we expect a trend in which participants respond by choosing the CDS lexical item to finish the sentence more than the Adult-Directed Speech (ADS) lexical item when and only when the talker used CDS prosody in the sentence they listen to.

Pilot 1.1: Proof of concept

Method

Participants

48 Mechanical Turk participants were tested. There were no exclusion criteria, so all data sets were included in analysis.

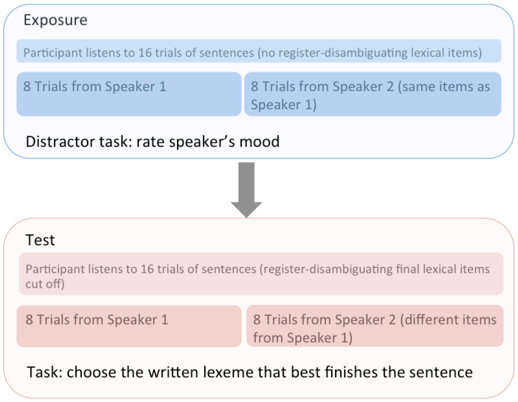
Materials

Sets of training and testing stimuli were recorded by two female talkers. Training materials consisted of eight sentence items, which were designed so that CDS and ADS interpretations were both possible in written form.. In recording, the talkers were shown an image of a child while recording CDS stimuli, and an adult while recording ADS stimuli, and asked to produce sentences in manners appropriate in these scenes.

Testing stimuli (16 items) were also sentences that could reasonably be interpreted as being directed at either a child or an adult. Each of these items had two versions: one in which the last word (a noun) was consistent with ADS-type lexical choices (e.g. "spaghetti"), and one in which the last word was a synonym for the original word, but more consistent with CDS lexical choices (e.g. "pasketti"). Lexical items for CDS were taken from the CHILDES corpus. Twenty words with forms of the diminutive suffix "-y" were identified, and stimuli sentences were constructed around them. Speakers recorded the list of testing stimuli four times, in order to include versions of each item in both CDS and ADS prosody, and that concluded with lexical items that were consistent and inconsistent with the register of the rest of the recording.

Procedure

In the training phase, participants were asked to listen to a set of 16 recordings (8 items read by each speaker). One speaker would use CDS prosody and the other would use ADS prosody. The mapping between the speaker and the register was counterbalanced across subjects. The subjects' task was to rate their perception of the speaker's mood in each recording, from "extremely happy" to "extremely unhappy". An avatar indicating the speaker's identity appeared above the audio media player in each trial.


Exposure

Participant listens to 16 trials of sentences (no register-disambiguating lexical items)

8 Trials from Speaker 1

8 Trials from Speaker 2 (same items as Speaker 1)

Distractor task: rate speaker's mood

Test

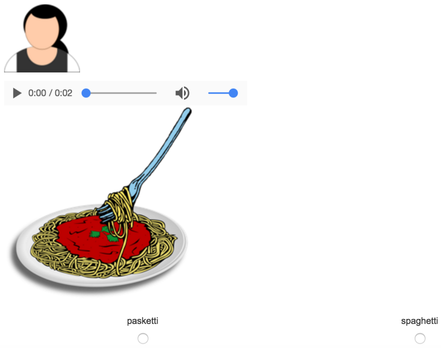
Participant listens to 16 trials of sentences (register-disambiguating final lexical items cut off)

8 Trials from Speaker 1

8 Trials from Speaker 2 (different items from Speaker 1)

Task: choose the written lexeme that best finishes the sentence

Figure 1: Structure of pilot study


0:00 / 0:02

pasketti ○

spaghetti ○

Figure 2: Example of display during test trial

In the testing phase, participants listened to 16 sentences (16 different items, 8 items spoken by each talker) in which the final noun was cut off (e.g., "Do you want any more [~~pasketti/spaghetti~~]"). The mapping between the speaker and the prosodic register was counterbalanced across subjects, as was the register actual word spoken at the end of each sentence (although subjects never heard this word) to control for possible timing and co-articulation effects. Images depicting the final nouns in testing stimuli were displayed underneath testing sound clips in order to suggest a visualization of the noun. Subjects were presented with the two register-biased versions of the final noun in the sentence and asked to choose which word they think the speaker would use to finish the sentence.

Results and Interpretation

Contrast-coded data from the testing phase was submitted to a two-way mixed-effects logistic regression model in R (lme4). The model included register, speaker identity, and their interaction as fixed effects and by-subject and by- item random intercepts. There was a significant main effect for both independent variables (speaker: β = -1.2168, *p* < .001; register: β = 1.4658, *p* < .001). The interaction between these two factors was marginally significant at the p < .05 level (β = 0.3449, *p* = 0.09612). The results suggest that, for the most part, if the audio input carried CDS acoustic cues, then listeners were more likely to complete the sentence with a CDS lexical choice than if the input did not carry CDS acoustic cues. The baseline rate at which this trend appeared, however, in part depended on which speaker they were hearing (see Figure 3: Proportion of CDS lexical responses by speaker (Pilot 1)).

The register effect is consistent with our hypotheses to a level that is statistically significant. As demonstrated in the figure, the register effect was more pronounced with Rachel than Holly. Each item in the test phase was only heard once overall, and for each participant, each talker only produced an utterance of a particular test item once. As a result, it was impossible to include test item as a main effect in the model that would converge under reasonable conditions.

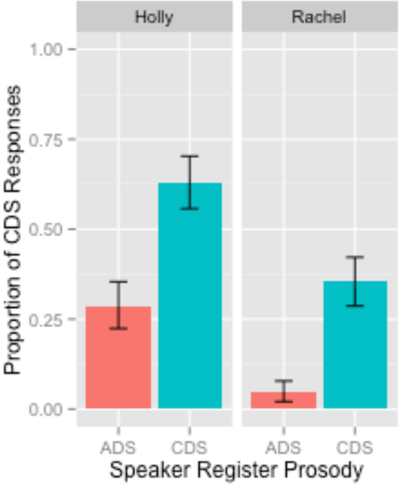
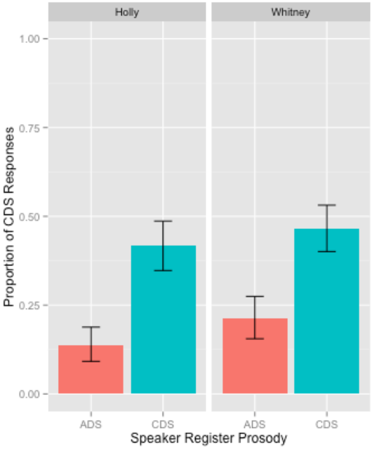


Figure 3: Proportion of CDS lexical responses by speaker (Pilot 1)

Pilot 2: Sanity checks and further details

In the first pilot, we ran twice the number of subjects we had originally intended because of concerns about subjects not clicking "play" on the audio and simply clicking through without listening. Auto play was not implemented because of current software issues at Qualtrics. In this follow-up pilot, we added an HTML autoplay work-around. For the test stimuli, we used infant-directed speech rather than toddler-directed speech, added four new items to the test phase to gauge them for future use, and replaced Rachel as a speaker with Whitney, who is a mother. We also counterbalanced whether or not the talker ended stimuli sentences with lexical items consistent with the prosody for that trial to avoid confounding co-articulation issues.



In overall design, nothing was changed other than the fact that we ran 40 participants in this pilot in order to fill counterbalancing cells rather than to add power to the statistical analysis.

The results from this pilot were analyzed the same way as the previous one. There was a significant main effect for each independent variable (speaker: $\beta = 0.25492$, $p < .01$; register: $\beta = 0.83487$, $p < .001$) and no significant interaction ($\beta = -0.08739$, $p = 0.62680$).

Figure 4: Proportion of CDS lexical responses (Pilot 2)

Experiment 1

The pilot studies suggest that there is a relationship between CDS prosody and lexical choices in language processing. Another interesting question, however, is whether this relationship is the result of a rote, “dumb” connection due to co-occurrence in the natural world, or the result of a higher-level inference about why a talker might produce CDS prosody.

We can represent the theoretical model of contextual extrapolation with the following equation:

$$P(\text{interlocutors} \mid \text{signal}) \propto P(\text{signal} \mid \text{interlocutors})\, P(\text{interlocutors})$$

With such a model, it is possible to examine the generative foundations of the results we observe in the pilot studies. The model above captures the generation of such an inference. We first test its predictions of human behavior in an off-line measure where we ask participants for direct estimations of the probabilities on either side of the equation and check that they are, in fact, proportional. If the offline measure proves successful, we can then follow up with further on-line measures to provide more solid evidence for the fit of the model to human behavior.

To test the generative model of contextual inference experimentally, we aim to establish the proportionality of the probabilities of the phrases on each side of the model: context given signal, and of signal given context.

For the purpose of this study, we boil the definition the “context” down to the relationship between the talker and addressee of a particular utterance. Given that we are working with Child Directed Speech, the context is that an adult is either talking to a child or to another adult. We define the “signal” as the realization of the message sent between talker and addressee. This includes both the propositional content of an utterance and its acoustic features as well, if the utterance is spoken rather than written.

Method

Participants

We aim to run 2400 participants from Mechanical Turk on one trial each. There are no exclusion criteria, or are there inclusion criteria at this time.

Materials

Using the raw recordings of Holly (only) from which we sampled stimuli for the pilot studies, the “test stimuli” sentences were extracted in full, with the final lexical item remaining. Both versions (where the final lexical item was consistent with the register of the performed prosody of the sentence and where it was not) of each sentence were used in the experiment.

Design & Procedure

Each participant is presented with one utterance or written sentence from the pool of stimuli and a scenario designating whether the speaker of the sentence is addressing a child or another adult. They are then asked to make two ratings on a scale of 0-10. Each set of participants breaks down into 12 counterbalancing conditions in a 2 x 2 x 3 design: whether the addressee in the given scenario is a child or an adult x whether the register of the lexical item at the end of the sentence is consistent with the register of the given scenario x whether the sentence they are presented with is performed with CDS prosody, ADS prosody, or written, respectively. We include a written condition in order to give a measure of clarity as to how much the prosodic signal itself is contributing to people’s perceptions of the child-directed register, and to provide a baseline of how much the propositional content of the stimuli items affects people’s judgments for future projects. The first half of participants are asked to judge how likely they think they would be to hear the sentence in the trial in the scenario they are given, then are asked how surprised they would be to hear the sentence in the given scenario. The second half of participants are asked how likely they think the scenario is given the sentence, then how surprised they would be to discover the scenario given was the one under which the sentence occurred. Participants are asked for both a likelihood and surprise judgment because both are important to estimating the ultimate probability values we are seeking in the model, but each judgment has some drawbacks on its own. Likelihood judgments may encompass both acceptability and expectation to a different degree across participants, whereas surprise judgments may have a high floor in responses.

Analysis, Predictions, an Interpretations

The results of the first set of participants are correlated to the results of the second half of participants. The mean response for each question type (likelihood & surprise) is taken for each counterbalancing condition in both participant sets. The values of the 12 condition cells in each participant set are submitted to a simple correlation analysis to determine the coefficient of correlation (r) and the coefficient of determination (r^2). If these values are high, it indicates a strong correlation between the two expressions on either side of the equation in our model, providing support for the validity of our model. If they are low, however, it indicates either that the model has failed to capture how humans can make linguistic inferences based on contextual information, or that our test questions fail to capture the concepts we want to ask about. The root mean squared error of this comparison is also calculated to account for different baseline probabilities between conditions.

Additionally, the results of the written conditions will be compared to the results of the conditions where participants heard the stimulus item. It is expected that results (probability estimates) in the written conditions will trend more centrally than results from conditions in which acoustic information was available.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs : restricting the domain of subsequent reference, *73*, 247–264.

Bicknell, K., Elman, J. L., Hare, M., Mcrae, K., & Kutas, M. (n.d.). Online Expectations for Verbal Arguments Conditional on Event Knowledge, *1*, 2220–2225.

Brown-schmidt, S., Yoon, S. O., & Ryskin, R. A. (2015). People as Contexts in Conversation. *Psychology of Learning and Motivation*, *62*, 59–99.

Cameron-faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech, *27*, 843–873. http://doi.org/10.1016/j.cogsci.2003.06.001

Cooper, R. P., & Aslin, R. N. (1990). Preference for Infant - directed Speech in the First Month after Birth. *Child Development*, *61*(5). http://doi.org/10.1111/j.1467-8624.1990.tb02885.x

Fine, A. B., Jaeger, T. F., Farmer, T. a, & Qian, T. (2013). Rapid Expectation Adaptation during Syntactic Comprehension. *PloS One*, *8*(10), e77661. http://doi.org/10.1371/journal.pone.0077661

Foulkes, P., Docherty, G., & Watt, D. (2005). Phonological variation in child-directed speech. *Language*, *81*(1).

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(20), 8051–6. http://doi.org/10.1073/pnas.1216438110

Jaeger, T. F., & Tily, H. (2010). On language “ utility ”: processing complexity and communicative efficiency. http://doi.org/10.1002/wcs.126

Kleinschmidt, D. F., & Jaeger, T. F. (2016). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203. http://doi.org/10.1037/a0038695.Robust

Levy, R., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. *Advances in Neural Information Processing Systems*.

Niedzielski, N. (1999). The Effect of Social Information on the Perception of Sociolinguistic Variables. *Journal of Language and Social Psychology*, *18*(1), 62–85.

Pegg, J. E., Werker, J. F., & Mcleod, P. J. (1992). Preference for Infant-Directed Over Speech : Evidence From 7-Week-Old Infants, 325–345.

Purnell, T., Idsardi, Wi., & Baugh, J. (1999). Perceptual and phonetic experiments on

American English dialect identification. *Journal of Language and Social Psychology*, *18*(1), 10–30.

Rowe, M. L. (2008). Child-directed speech : relation to socioeconomic status , knowledge of child development and child vocabulary skill *. *Journal of Child Language*, *35*, 185–205. http://doi.org/10.1017/S0305000907008343

Staum Casasanto, L. (2009). How Do Listeners Represent Sociolinguistic Knowledge? *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, 2341–2346.

Tanenhaus, M. K., Spivey-knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Comprehension Integration of Visual and Linguistic Information in Spoken Language Comprehension, *268*(5217), 1632–1634.

Van Berkum, J. J. a, van den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, *20*(4), 580–91. http://doi.org/10.1162/jocn.2008.20054

Warren-leubecker, A., Neil, J., & Bohannon, J. N. (2016). Intonation Patterns in Child-Directed Speech : Mother-Father Differences Author (s): Amye Warren-Leubecker and John Neil Bohannon III Source : Child Development , Vol . 55 , No . 4 (Aug ., 1984), pp . 1379-1385 Published by : Wiley on behalf of the So, *55*(4), 1379–1385.