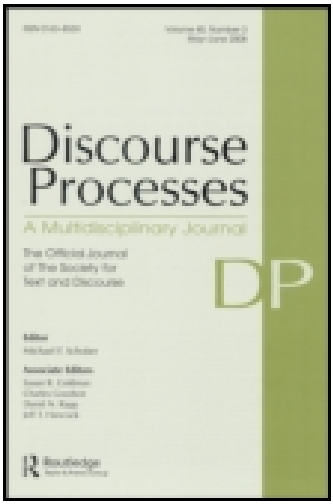

Discourse Processes
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hdsp20

Prosody and the Interpretation of Hierarchically Ambiguous Discourse

Joseph Tyler[a]
[a] Department of Linguistics, University of Michigan
Accepted author version posted online: 30 Dec 2013.Published online: 13 Oct 2014.

To cite this article: Joseph Tyler (2014) Prosody and the Interpretation of Hierarchically Ambiguous Discourse, Discourse Processes, 51:8, 656-687, DOI: 10.1080/0163853X.2013.875866

To link to this article: http://dx.doi.org/10.1080/0163853X.2013.875866

Discourse Processes, 51:656–687, 2014

ISSN: 0163-853X print/1532-6950 online
DOI: 10.1080/0163853X.2013.875866

Routledge
Taylor & Francis Group

Prosody and the Interpretation of Hierarchically Ambiguous Discourse

Joseph Tyler
Department of Linguistics
University of Michigan

Although significant attention has been devoted to prosody in discourse production, relatively little is known about prosody's effect on discourse interpretation. This article explores the ability of synthetic manipulations of prosody to bias interpretation of discourse ambiguities where a first sentence is linked to two following sentences either by coordinating (Narration) or subordinating (Elaboration) discourse relations. In Experiment 1, manipulations of pitch, pause duration, and intensity were found to influence discourse interpretation. In Experiment 2, subsets of these prosodic contrasts were compared. A bias for more coordination interpretations was found only for subsets with rising pitch at the end of the first sentence, including one where that was the only contrast, showing that rising pitch alone can disambiguate discourse. Participants also expressed more confidence when choosing a coordination interpretation after hearing a rise or a subordination interpretation after hearing a fall. Results demonstrate that the discourse disambiguation ability of prosody goes beyond ambiguities of scope and reference to hierarchical ambiguities of coordinating and subordinating discourse relations.

Joseph Tyler is now at the Department of English, Morehead State University.

Correspondence concerning this article should be addressed to Joseph Tyler, Department of English, Bert Combs Building 103, Morehead State University, Morehead, KY 40351, USA. E-mail: josephctyler@gmail.com

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hdsp.

INTRODUCTION

As competent speakers of a language, we know more than just the structure of sounds (phonology) and sentences (syntax) but also how sentences combine together to make coherent discourse. The ability to identify how sentences fit together is integral to effective comprehension of spoken or written material. Sometimes this process requires reasoning why two sentences might be uttered side by side. For example, a likely interpretation of (1) is that John's pushing Max caused Max to fall, even though this is not explicitly stated. Another possible interpretation is that first Max fell and later John pushed him. These two interpretations involve different causal and temporal relations between the two sentences. But although (1) is ambiguous, it can be disambiguated by adding lexical material like in (2) and (3).

(1) Max fell. John pushed him. (Asher & Lascarides, 2003)
(2) Max fell, because John pushed him.
(3) Max fell. Then, John pushed him.

Lexical markers of discourse relations, for example, *because* and *then*, are sometimes called discourse markers (Schiffrin, 1987), coherence markers (Kamalski, Sanders, & Lentz, 2008), cue phrases (Knott & Mellish, 1996), or connectives (Webber, Stone, Joshi, & Knott, 2003). These lexical cues help listeners identify how sentences are related and the discourse is structured.

Although listeners exploit lexical cues and general reasoning in identifying a discourse's structure, they may also take advantage of prosodic cues (e.g., variation in pitch, pause duration, intensity). We already know that listeners use prosody to disambiguate various kinds of syntactic ambiguities (Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Snedeker & Trueswell, 2003). We also know from discourse prosody production studies that prosody correlates systematically with discourse structure. For example, larger discourse boundaries tend to be produced with longer pause durations, and higher postboundary maximum pitch and intensity (den Ouden, Noordman, & Terken, 2009; Grosz & Hirschberg, 1992; Lehiste, 1982; Tyler, 2013). Theories of discourse have also proposed local hierarchical contrasts, and production studies have found prosodic correlates of them. For example, den Ouden et al. (2009) found more important discourse segments had slower articulation rates. Tyler (2013) showed related discourse segments that were hierarchically equal (i.e., coordinated) tended to be produced with relatively longer preceding pauses, higher maximum pitch, and higher maximum intensity than segments where one dominates another (i.e., subordination). Tyler also showed an interaction effect of boundary size and hierarchical structure, where the effect of hierarchical structure disappeared as boundary size increased.

But although discourse prosody production studies have shown that speakers' prosody carries information about the structure of discourse, relatively little is

known about when or how listeners use this prosodic variation in their interpretation. In discourse prosody perception research, most work has used indirect measures of linguistic perception like naturalness judgments (Smith, 2004) or judgments about a sentence's location in the discourse (Lehiste, 1982), for example, is it paragraph-final or not.

Two studies have tested whether discourse prosody can specifically affect the interpretation of linguistic expressions (Mayer, Jasinskaja, & Kölsch, 2006; Silverman, 1987). Mayer et al. tested whether prosody could bias the resolution of an ambiguous pronoun toward an antecedent in the previous sentence or an antecedent further back in the discourse. They found that when listeners had both pause duration and pitch cues, there was a significant effect of prosody on pronoun interpretation. In two follow-up experiments, they discovered that pause duration or pitch alone could not achieve the same result. It seems there needed to be cues available in both pause duration and pitch range to show an effect in listeners' interpretation. Silverman (1987) tested whether prosody could bias the interpretation of the size of the domain of a quantifier phrase (e.g., "all materials"), scoping over two sentences or many preceding sentences. He found significantly more predicted interpretations of ambiguous quantifier phrases when listeners have available both pitch and pause duration cues (84.2%) compared to when pause durations were held constant (71.7%).

Because the contrasting meanings in the discourses of Mayer et al. (2006) and Silverman (1987) can be explained as resulting from different hierarchical structures, one possible conclusion from their results could be that prosody can disambiguate hierarchical discourse structure. An alternative and equally explanatory conclusion is that prosody conveys something about the scope or distance with respect to which one should interpret a particular linguistic expression. For Mayer et al., prosody helps differentiate an antecedent in the preceding sentence from an antecedent three sentences back. For Silverman, prosody helps differentiate a quantifier phrase as scoping over two versus five preceding sentences. These studies' results cannot separate whether prosody is helping to disambiguate a hierarchical structure contrast or a distance contrast.

This article explores ways listeners use prosody to interpret hierarchically ambiguous discourse in the absence of overt lexical cues while controlling for distance, using discourses like the following:

(4) I sat in on a history class. I read about housing prices. And I watched a cool documentary.

On one interpretation, the events denoted by the second and third sentences (S2 and S3) take place during the event denoted by the first sentence (S1), that is, they are elaborating. On another interpretation, all three sentences describe separate independent events. In the first interpretation, S2 and S3 are embedded in S1, that

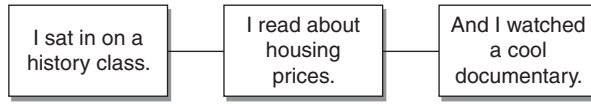

FIGURE 1 Subordinating interpretation of (1).

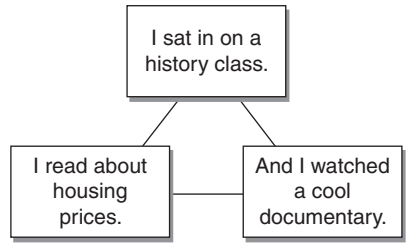

FIGURE 2 Coordinating interpretation of (1).

is, they are subordinated to S1. Using the terminology of Segmented Discourse Representation Theory (Asher & Lascarides, 2003), they are related via the subordinating relation of *Elaboration*. In the second interpretation, all sentences are at the same hierarchical level, that is, they are coordinated to each other (in Segmented Discourse Representation Theory via *Narration* relations). These contrasting meanings can be represented visually as in Figures 1 and 2, where vertical and horizontal lines indicate subordination and coordination, respectively.

Because S2 attaches to S1 in both interpretations, the ambiguity is a result not of *where* S2 attaches but *how* S2 attaches, that is, whether it is coordinated or subordinated to S1. For this reason, the experiments presented here, testing whether prosody can disambiguate discourses like (4), are testing whether prosody can disambiguate *hierarchical* discourse structure (Asher & Lascarides, 2003) while holding potential distance effects constant.

EXPERIMENT 1: PROSODIC EFFECTS ON THE INTERPRETATION OF DISCOURSE AMBIGUITIES USING A SET OF SYNTHESIZED PROSODIC MANIPULATIONS

In this section I present an experiment testing whether a set of prosodic manipulations can bias the interpretation of an ambiguous discourse. All discourses in this study were three sentences long, where sentence 2 attaches to sentence 1 via either a coordinating or a subordinating relation. For example, the discourse in (4) could be interpreted such that the narrator read about housing

prices and watched a cool documentary while sitting in on the history class (the Subord interpretation) or separate from the history class (the Coord interpretation).

Method

Participants. Forty students from the University of Michigan Psychology Subject Pool participated in this study in exchange for course credit. All reported being native speakers of American English. Ages ranged from 17 to 21 with a mean of 18.43. Of the 40 participants, 13 were men. Fourteen (35%) reported knowing a second language.

Materials. A total of 102 discourses were constructed, each discourse being ambiguous between the Coord and Subord interpretations described above. These 102 discourses were included in a norming study to test for general preferences for each discourse's interpretation as Coord or Subord. This norming study was meant to ensure the discourses were reasonably, not just possibly, interpretable as either Coord or Subord. The norming took place in an online survey through the Qualtrics survey research tool (https://www.qualtrics.com/), with participants recruited through Amazon Mechanical Turk (https://www.mturk.com/). Participants indicated whether they preferred the coordinating, subordinating, or an "other" interpretation. Eight discourses that received more than 10% "other" interpretations were excluded. The remaining discourses were ordered from most ambiguous, indicating they received a more equal number of Coord and Subord interpretations, to most biased. The 52 most ambiguous discourses in the norming study were selected as the discourses for this study, each having a second best interpretation chosen at least 25% as often as the preferred interpretation. The 48 most ambiguous discourses were used as target stimuli, with the remaining four serving as training (see Table 1 for the full list of stimuli).

All spoken materials were recorded in the sound-attenuated booth in the University of Michigan Linguistics Department's Sound Lab. Each individual sentence in each discourse was separated and entered into a list, resulting in a list of $52 \times 3 = 156$ sentences. Each sentence was then placed into a carrier context like in (5).

(5) I am going to read a sentence. I read about housing prices. I just read a sentence.
I am going to read a sentence. I sat in on a history class. I just read a sentence.
I am going to read a sentence. I went for a run. I just read a sentence.

After randomizing the order of presentation, a 30-year-old, female, native speaker of American English was recorded reading each sentence out loud one at a time in its carrier context. This reader was instructed to say the sentences as

TABLE 1
Full Set of Discourses, With Norming Results

Bias	*Discourse Text*
0	I visited my uncle in Detroit. I saw a movie. And I went for a run.
1	I spent the day at work. I played some ping pong. And I experimented with paper airplane designs.
1	I went to the gas station. I bought an apple. And I picked up some wine.
1	I sat in on a history class. I read about housing prices. And I watched a cool documentary.
1	I finished my senior project. I taught some kids how to tango. And I put on a show at school.
2	I did some work for class. I read about dogs. And I took some pictures.
2	I partied at my friend's house. I changed my status on Facebook. And I spilled juice on my shirt.
2	I went to the art fair. I bought some dinner. And I saw a performance by the Pink Flamingoes.
2	I hung out with my boyfriend. I did some homework. And I played guitar.
3	I competed in a race. I built a raft. And I gave my friend Jason a pep talk.
3	I took a trip in my convertible. I played some disc golf. And I ate a lot of beef jerky.
4	I cleaned the kitchen. I vacuumed my new rug. And I took out the trash.
4	I got my living room ready for a party. I fixed the fire alarm. And I put away my clothes.
5	I went to the market. I met up with my advisor. And I ate some good food.
5	I laid in bed for a while. I ate a bowl of chicken soup. And I played with my cat.
6	I worked on a project with my neighbor. I baked a cake. And I put up decorations.
6	I relaxed on the sand. I played some chess. And I read a novel.
7	I am getting trained for my job at the mall. I am learning to be a better public speaker. And I am figuring out how to use my new smartphone.
7	I squeezed in a workout. I walked to my parents' house. And I helped my dad move some furniture.
9	I went for a hike. I hung out with my buddies. And I scavenged for seashells.
10	I go on dates whenever I can. I go to museums. And I occasionally go out for a drink.
10	I ran some errands. I picked my dad up from the airport. And I got take-out Chinese food for dinner.
10	I ate some breakfast. I enjoyed the sunshine. And I read a few chapters of my book.
11	I worked an eight hour shift. I did some crossword puzzles. And I got yelled at by a sketchy homeless guy.
12	I walked around the art fair. I gave a friend a pep talk. And I thought about the war in Afghanistan.
12	I played fetch with my dog. I practiced my Frisbee technique. And I watched a soccer game.
12	I planned a practical joke. I bought a bucket of paint. And I wrote a letter.
12	I went to the grocery store. I got Starbucks coffee. And I picked up my photo prints.
13	I overreacted to a friend's comment. I went for a long walk. And I wrote in my diary for an hour.
13	I visited the state fair. I learned how to knit. And I saw my favorite band.
14	I stopped by my hometown. I wrote a bunch of thank-you notes. And I bought a new outfit at the mall.

(Continued)

TABLE 1
(Continued)

Bias	*Discourse Text*
14	I babysat for my neighbors. I baked a pie. And I gave my brother a call.
15	I went on a road trip. I played some disc golf. And I ate a lot of beef jerky.
15	I went to the library. I listened to a presentation about music. And I got a cup of coffee.
15	I went on a date. I saw a bunch of movies. And I almost fainted.
18	I did some work. I read a book. And I talked to my boss.
19	I went to English class. I drew some cartoons. And I gave a presentation.
21	I played on the computer. I read the newspaper. And I chatted with a friend.
21	I went to my brother's birthday party. I got a drink with Sharon. And I played some darts.
21	I worked on my computer. I listened to some music. And I looked at some photos.
22	I spent some time in Chicago. I went to the beach. And I saw an old friend from college.
22	I took care of some business. I bought some painting supplies. And I cashed a check.
22	I went home for Easter. I watched the NBA playoffs. And I ran in a race for the first time.
22	I took the dogs for a walk. I picked some wild berries. And I dropped off a letter at the mailbox.
22	I stopped by the market. I did some people watching. And I saw an accordion performance.
22	I picked up some stuff for my mom. I got some bird seed for the bird feeder. And I bought a couple rose bushes.
23	I went to my neighborhood block party. I cleaned my picnic table. And I went for a bike ride.
23	I played a game. I turned on my computer. And I relaxed on the couch.

Note. Bias indicates the difference between the number of participants who chose the Coord interpretation and the number who chose the Subord interpretation. The discourses are ordered from most ambiguous (least biased) to least ambiguous.

naturally as possible. Disfluent productions, including those that had missing words, extra words, or extraverbal interruptions like coughing and sneezing, were re-recorded afterward until all productions were fluent. In some cases, the speaker independently chose to re-record a sentence; in these cases, the final production was used.

Each target sentence was then spliced out from these readings. These sentences' prosody was manipulated in Praat (Boersma & Weenink, 2009) in the following ways. First, to guard against unintended intensity variation, all files were normed to the overall mean intensity of 57.2 dB. It is these intensity-normalized files that are used for the subsequent manipulations. Then, five different prosodic contrasts were created, motivated by both the literature on prosodic correlates of discourse structure in discourse production (den Ouden et al., 2009; Hirschberg & Grosz, 1991; Tyler, 2012; Yule, 1980) and a previous study (Tyler, Kahn, & Arnold, 2011). Tyler et al. (2011) found that speakers intending to convey one interpretation of an ambiguous discourse produced

systematic prosodic correlates, but that listeners were unable to retrieve the speakers' intended meaning. Within the overall null effect, however, there was one speaker whose productions listeners were able to correctly identify 75% of the time. This speaker's productions revealed contrasts in terminal pitch contours and pause durations, and her behavior serves as primary motivation for the manipulations.

Sentence-final pitch contours for S1 and S2, but not S3, were manipulated. For each contour manipulation, it was important to have a consistent temporal window. Because the final contour generally began at the last stressed syllable, the window for manipulation was from the last stressed syllable to the end. In a pitch manipulation object in Praat, all pitch points from the last stressed syllable to the end of the file were selected, and all but the first and last of these pitch points were deleted. Then, the last pitch point was multiplied by a factor of the pitch of the last stressed syllable's pitch point, depending on whether the sentence was a first or second sentence in the discourse and whether it was a Coord or a Subord manipulation. For Coord manipulations, S1's final pitch point was multiplied by 1.6, and S2's final pitch point was multiplied by 1.3. For Subord manipulations, S1's final pitch point was multiplied by .75, and S2's final pitch point was multiplied by 1.1. Figure 3 shows the original and manipulated pitch contours for discourse (4). Although the resulting contours have linear slopes, and thus lose the nonlinear movements from the original productions, it was a consistent way of constructing the final contours.

After assigning a new pitch contour to all S1s and S2s, the mean pitch and intensity of S2 and S3 were multiplied by 1.1 for the Coord condition and .9 for the Subord condition. For pitch, this was achieved with a Praat script that

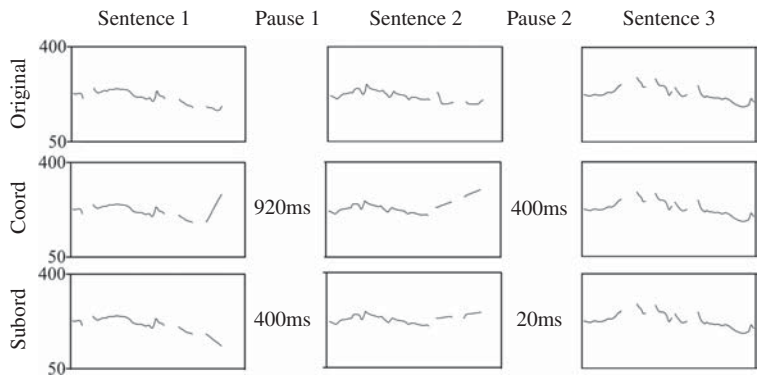

FIGURE 3 Pitch contours and pauses for original, Coord, and Subord versions of the sentences for the discourse in (4).

multiplied all pitch frequencies in the Manipulation object. For intensity, a Praat script using the *scale intensity* function reassigned mean intensity from the overall average of 57.2 to 57.2 × 1.1 = 62.92 for Coord or 57.2 × .9 = 51.48 for Subord. The files were re-examined and silences before and after the recorded sentences removed.

Then, the sentences were concatenated with intervening pauses of different durations. For the Coord condition, the first pause (P1) between S1 and S2 was 920 ms and the second pause (P2) between S2 and S3 was 400 ms. For the Subord condition, P1 was 400 ms and P2 was 20 ms. Like the pitch manipulations, these pause durations were motivated by the productions of the speaker in Tyler et al. (2011). It was these final concatenated sound files (summarized in Table 2) that were presented to participants, and it was these prosodic contrasts that correspond to the predictor *prosody* in the statistical model.

Design. Using a within-subjects design, prosody was the predictor of interest, with discourses balanced for prosody and question type. Participants heard each discourse in one of two prosodic conditions, with prosody manipulated to encourage either coordinating or subordinating interpretations. Yes/no questions directly querying the interpretation were used. To control for the affirmative response bias, question type was varied such that a yes answer sometimes indicated a coordinating and sometimes a subordinating interpretation. For example, they might hear the discourse in (4) while being presented one of the following two interpretation questions:

(6) Affirmative response = Coord: Did Sally mean that she read about housing prices and watched a cool documentary separate from history class?
(7) Affirmative response = Subord: Did Sally mean that she read about housing prices and watched a cool documentary in history class?

The predicted answer is a Coord interpretation after hearing Coord prosody and a Subord interpretation after hearing Subord prosody.

TABLE 2
Prosodic Manipulations for Coord and Subord Prosody Conditions

	Coord Prosody	*Subord Prosody*
Terminal pitch on S1	Rise with slope of 1.6	Fall with slope of .75
Terminal pitch on S2	Rise with slope of 1.3	Rise with slope of 1.1
Pause duration between S1 and S2	900 ms	400 ms
Pause duration between S2 and S3	400 ms	20 ms
Mean pitch and intensity on S2 and S3	Mean pitch and intensity multiplied by 1.1	Mean pitch and intensity multiplied by .9

Four groups of participants were created, with each group hearing 12 discourses for each combination of prosody and question bias. This way, each participant heard an equal number of discourses representing each prosodic condition and each question type. The groups were counterbalanced so each discourse was presented an equal number of times. The 48 target discourses were separated into four blocks of 12 discourses. Although the blocks were always presented in the same order, the discourses within each block were randomized, mitigating local sequencing effects. The blocking allowed for comparison between blocks of 12 to see if participants changed their behavior over the course of the experiment. From piloting, it appeared that participants may not initially use prosody in their interpretation but with repetition begin to do so. This blocking was included to check this potentiality. Within each block of 12, there were 3 discourses for each combination of prosody and question type.

Each participant group and presentation quarter was also assigned a balance of discourses with a range of ambiguity, from those where both the Coord and Subord meanings were nearly equally accessible to those where either Coord or Subord was preferred more than the other. There were no fillers. Although fillers with more blatant prosodic contrasts could have been included, this may have prevented participants from paying attention to the more subtle contrasts in the prosodic manipulations of interest. Fillers with different kinds of structural contrasts could also have been included, but this would have made it more difficult for listeners to get used to the discourses and the elicitation questions. The discourse ambiguities are difficult enough to process, giving listeners a chance to get comfortable with the relevant meaning contrast and the elicitation questions should reduce noise in the data due to processing difficulties.

Preceding the 48 target discourses were four training discourses, one in each combination of prosodic condition and question type. All participants heard the same training discourses, in the same order, with the same questions and same prosody. For participants, the first 4 discourses were indistinguishable from the remaining 48. The training discourses, which were not included in the final analysis, provided a chance for participants to get some basic familiarity with the task before their data counted.

Procedure. Participants listened to the discourses one at a time and answered associated questions in a Qualtrics survey. They were told they were going to hear a series of stories told by a woman named Sally and that those stories could be interpreted multiple ways. Their task was to answer questions about how they interpreted the stories. They were instructed to adjust the volume to comfortable levels and they could listen to each discourse as many times as desired. For each discourse, listeners answered three questions. First, they saw a page with an audio play button and an interpretation question that queried

whether they got the Coord or Subord interpretation of the discourse. The interpretation question was a yes/no question that asked *Did Sally mean that [Coord Interpretation]* or *Did Sally mean that [Subord Interpretation]*.

After answering the interpretation question, they advanced a screen and were asked how confident they were in their interpretation on a scale from 1 to 100. Finally, they answered a factual comprehension question about the discourse they had just heard to check whether they were paying attention. Participants saw only one question on the screen at a time, could not advance without answering the question, and could not go back and change previous answers.

For example, on the first screen they might hear the discourse in (4) while being presented one of the interpretation questions in (6) or (7). They were then asked how confident they were in their choice; finally, they were asked a comprehension question like "did Sally pick up some beer?" After answering all three questions, they would advance to the next discourse and continue.

Predictions. Prosody was predicted to bias interpretation, with listeners hypothesized to provide more Coord interpretations when they hear sentence-final rises on sentences 1 and 2, longer pauses and higher and louder versions of sentences 2 and 3 (i.e., Coord prosody) than when they hear a fall on sentence 1, a smaller rise on sentence 2, shorter pauses and lower and quieter versions of sentences 2 and 3 (i.e., Subord prosody).

Results

Results showed that all 40 participants got at least 91% of the comprehension questions correct, suggesting they were all attentively participating. Therefore, data from all 40 participants were included in the analysis.

Whether listeners used prosody to help disambiguate the coordinating/subordinating discourse ambiguities was tested by checking to see if listeners' interpretations matched predictions, where a predicted outcome (called "match" here) would be a Coord interpretation after hearing Coord prosody or a Subord interpretation after hearing Subord prosody. Prosody could have been tested as a predictor of interpretation, although this result is identical to testing relative likelihood of match. Using match has the benefit of making subsequent analyses simpler. For example, subsequent analyses using other predictors are getting more directly at the core question of whether those predictors affect match rate, regardless of which prosody they heard. This avoids having to do analyses for each other predictor variable (e.g., discourse ambiguity, question bias, demographic factors) with both Coord and Subord prosody independently. The use of the terms *match* and *mismatch* is not intended to say one interpretation is better than another but to indicate as simply as possible if a listener's interpretation did or did not fit predictions.

The statistical analysis used a mixed effects model with binomial outcome run within R (R Development Core Team, 2013) and fit using the lmer function in the R package lme4 (Bates, Maechler, & Bolker, 2013). The dependent variable was a binary of match versus mismatch (of prosody and interpretation). No categorical or continuous predictors were included in this initial model, testing only whether match likelihood was above chance. All models included random effects for subject and discourse (unless otherwise noted), fitting the recommendations of others (Barr, Levy, Scheepers, & Tily, 2013; Clark, 1973; Jaeger, 2008). For a discussion of the benefits of mixed-effects models with binary outcomes relative to other repeated-measures models, see Quené and van den Bergh (2008) and Jaeger (2008).

Results showed that chance of match was significant ($\beta = .355$, SE $= .089$, $z = 4.00$, $p < .001$) with a positive coefficient, meaning listeners' interpretations were significantly more likely to be as predicted by the prosody than not. Coord prosody resulted in a 59% match rate (571 of 960), whereas Subord prosody had a 57% match rate (547 of 960). Results showed no significant difference between the two conditions of prosody ($ß = .108$, SE $= .094$, $z = 1.145$, $p = .252$), indicating that neither Coord nor Subord prosody was more likely to get the predicted interpretation.

It is possible this general effect of prosody on interpretation varies over the course of the experiment (e.g., a learning effect), depends on the underlying ambiguity of the discourse, or changes from participant to participant. The question about how prosody's effect on interpretation could change from the beginning to the middle and end of the study was explored by comparing the four quarters of the experiment. Every participant heard the same 12 discourses in each quarter. To test for changes in performance over time, a continuous variable for presentation order was included in the model. This variable was not a significant predictor of match ($ß = .016$, SE $= .042$, $z = .384$, $p = .701$). Therefore, participants appeared to use prosody consistently over time, neither showing a learning effect nor a fatigue effect.

It is also possible that the ability of prosody to disambiguate discourse depends on the practical ambiguity of the discourses themselves; the lexical material of one discourse could bias so much toward one interpretation that prosody would have no effect, whereas when multiple meanings are more equally accessible a factor like prosody could have an impact. Similar issues are raised in the psycholinguistic literature, where researchers have manipulated speaker awareness of an ambiguity to see if they then use prosody to disambiguate (Allbritton, McKoon, & Ratcliff, 1996; Schafer, Speer, Warren, & White, 2000; Snedeker & Trueswell, 2003). For example, Snedeker and Trueswell experimentally manipulated whether the context supports two possible meanings of a sentence or heavily biases toward one. They found that speakers use disambiguating prosody when the context supports two interpretations, but if the context heavily biases toward one interpretation

"these cues all but disappear" (2003, p. 128). This suggests prosody may be exploited only when other cues do not disambiguate.

The norming study was done to create a set of discourses where lexical content did not so heavily bias interpretation as to wash out any potential effect of the prosodic contrasts. Nevertheless, the norming study results still show variation between the selected discourses in terms of group preferences for one interpretation or the other. A covariate was included in the model that measured the absolute value of the difference between the number of people who chose Coord and Subord interpretations. The scale of this variable was from zero (equibiased) to 26 (most biased), with higher values indicating greater bias toward either Coord or Subord. The degree of a discourse's underlying ambiguity was not found to affect participants' matching discourse interpretation with prosody (ß = .001, SE = .006, $z = .186$, $p = .853$). This suggests that the degree of ambiguity of these discourses did not affect the prosodic effect on interpretation. Naturally, it is possible the degree of ambiguity would have had an impact if discourses with a wider range of ambiguity had been included.

Results so far have been discussed for the participant population as a whole, but there is substantial variation from participant to participant. One dimension of such variation is a participant's overall preference for Coord or Subord interpretations. Table 3 presents each participant's mean interpretation preference, sorted from most Coord-preferring to most Subord-preferring.

There is clear variability of interpretation preferences. To test if these preferences affect participants' use of prosody, participants' interpretation preferences were correlated with their match rate. The result was significant ($\rho = .101$, $p < .001$), with participants with greater Coord preference showing more predicted interpretations. The r^2 value for this relationship is .01, however, showing that although significant, the relationship is not particularly explanatory.

Within the 40 participants, 3 showed a high match rate (>80%), suggesting that some participants were behaving more as predicted than others. Demographic factors, including education level, gender, age, and mono- versus multilingual status, do not explain the difference, as none of these were

TABLE 3
Proportion Coord Interpretations by Participant, Sorted From Most to Least Coord-Preferring

.979	.708	.563	.500	.479	.458	.396	.333
.938	.708	.563	.500	.458	.438	.375	.333
.896	.625	.542	.500	.458	.417	.375	.292
.833	.604	.542	.500	.458	.417	.354	.250
.729	.563	.500	.479	.458	.396	.333	.250

significantly predictive of a match between prosody and interpretation. To test the impact of these high match rate participants, the three participants with the highest match rates were excluded from the analysis and prosody still had a significant effect on interpretation (ß = .355, SE = .089, $z = 3.998$, $p < .001$), indicating the effect is not only due to these participants.

In addition to concerns about the degree of a discourse's ambiguity and individual variation, the naturalness of the prosodic manipulations may affect interpretation. To test this eventuality, naturalness judgments were elicited on a scale from 1 to 5 (1 = "highly unnatural", 5 = "highly natural") for three versions of the discourses: Coord (S1- and S2-final rises, longer pauses, increase in S2 and S3 mean pitch and intensity), Subord (S1-final fall, smaller S2-final rise, shorter pauses, decrease in S2 and S3 mean pitch and intensity), and unmanipulated (original speech samples, concatenated with 400-ms intervening pauses). After excluding two participants who were not native speakers of American English, a total of 40, 42, and 42 participants rated the naturalness of the Coord, Subord and unmanipulated versions, respectively (a between-subjects design). Results show a difference in mean naturalness judgments (Coord: mean = 3.69, $SD = .271$; Subord: mean = 3.47, $SD = .418$; Unmanip: mean = 3.55, $SD = .288$).

To test whether the prosodic differences had an impact on naturalness judgments, mixed models were fit with manipulation as predictor (unmanip, Coord, Subord), each discourse's mean naturalness as a dependent variable, and discourse as a random effect. Results show Coord was significantly different from Subord (ß = –.219, SE = .046, $t = -4.74$, pMCMC[1] = .002), unmanip was different from Coord (ß = .138, SE = .032, $t = 4.34$, pMCMC = .010), and unmanip was different from Subord (ß = –.081, SE = .044, $t = -1.82$, pMCMC = .015). The effect of the manipulations then seems to be having made the Coord samples sound more natural and the Subord samples sound less natural.

The clear related question, then, is whether the varied naturalness can explain the prosodic effect on interpretation. This was tested with a mixed model with interpretation as a binary outcome, subject as a random effect, prosody (Coord or Subord) as a categorical predictor, and naturalness as a continuous predictor. Results show match was predicted by prosody (ß = .729, SE = .102, $z = 7.126$, $p < .001$) but not naturalness (ß = .084, SE = .141, $z = .598$, $p = .550$). These results indicate the effect of prosody on interpretation holds even when controlling for naturalness, and the effect of prosody on interpretation is not reducible to a naturalness difference.

[1] p values are the Markov Chain Monte Carlo p-value output from pvals.fnc in R.

Confidence. Participants were also asked to rate how confident they were in their discourse interpretation judgment. Confidence judgments were collected on a scale from 1 to 100 and were analyzed as a continuous, not binary, outcome. The data were fitted to a linear mixed model using the lmer function in the R package lme4, with random effects for both subject and discourse. Results show match was predictive of confidence ($ß = 1.56$, $SE = .65$, $z = 2.40$, $p\text{MCMC} < .013$),[2] indicating that listeners were more confident in their judgments when their interpretation was as predicted by the prosody.

Discussion

The results of this study indicate that terminal pitch on sentences 1 and 2, pause duration contrasts after sentences 1 and 2, and overall pitch and intensity contrasts on sentences 2 and 3 collectively bias the interpretation of hierarchically ambiguous discourse. Furthermore, the set of five Coord-biasing manipulations and the set of five Subord-biasing manipulations resulted in significantly more Coord and Subord interpretations respectively, meaning the overall effect is not the result of just Coord- or just Subord-biasing manipulations. Moreover, the effect for Coord prosody is not significantly different from the effect for the Subord prosody, meaning not only are both sets of manipulations contributing but they are contributing equally.

One limitation of this experiment is the use of five prosodic manipulations, obscuring the contribution of each one. Prior research shows that prosodic effects on discourse interpretation seem to be stronger when more cues are used in tandem. Silverman (1987) showed an 84% disambiguation rate with three cues (two pitch and one pause), but with the pause duration contrast neutralized the rate dropped to 71% (p. 6.27). The first experiment in Mayer et al. (2006), with both a pitch and pause manipulation, showed a significant effect of prosody on interpretation. When the experiment was conducted again with either just pause or just pitch, the effect disappeared. Silverman explains that "this is hardly surprising: the more redundantly the prosodic structure is encoded in the acoustic signal, the more likely it is that listeners will be able to recover it and use it during speech perception" (p. 6.27). He argues listeners are more able to disambiguate because they are more able to recover the prosodic structure. An alternative explanation is that there is actually intersubject variability in which cues listeners pay attention to. It is possible that the drop in the disambiguation rate when fewer prosodic contrasts are included is due to some listeners no longer having the cues

[2]The *p* value was calculated with the pvals.fnc() function. Results are presented with the commonly used MCMC (Markov Chain Monte Carlo) *p* values, as used elsewhere (Baayen, Davidson, & Bates, 2008; Rohde, Levy, & Kehler, 2011).

that were relevant for them. For example, some participants in Silverman's studies may have focused on pause duration, and neutralizing the pause duration contrast would have removed the information they were using to disambiguate. These two explanations may both be right, as everyone is likely to be able to perceive meaningful contrasts in different prosodic measures to some degree, but there is also likely to be variation between individuals as to how much they focus on any particular prosodic measure.

EXPERIMENT 2: ISOLATING THE SYNTHESIZED PROSODIC MANIPULATIONS INFLUENCING THE INTERPRETATION OF DISCOURSE AMBIGUITIES

Experiment 1 explored the ability of synthetic manipulations of prosody to bias the interpretation of ambiguous discourse. The prosodic contrast had an overall effect on interpretation, with discourses in the Coord prosody condition resulting in more Coord interpretations than those in the Subord prosody condition. But because there were five total prosodic manipulations that constituted the prosodic contrast, it was unclear which one or ones contributed to the effect. This experiment presents the results of a series of follow-up studies that test various combinations of prosodic contrasts, which helps isolate which ones are driving the effect on discourse interpretation.

Method

The studies in Experiment 2 differ from Experiment 1 in two ways. First, instead of using the University of Michigan Psychology Subject Pool, participants were drawn from Amazon Mechanical Turk, with its concomitant differences in payment, setting, and other factors. Second, the prosodic contrast is composed of different sets of prosodic manipulations, except for one study that is a replication and so contains the same prosodic manipulations.

Participants. All participants in these studies participated via Amazon's Mechanical Turk service in exchange for 2 dollars. Amazon Mechanical Turk, a crowdsourcing platform where requesters post tasks that workers can complete for a set fee, has become an increasingly popular source of participants for behavioral research (Kittur, Chi, & Suh, 2008). It offers the benefits of a more diverse subject pool, fast data collection, inexpensive rates and avoids experimenter bias (Paolacci, Chandler, & Ipeirotis, 2010). Moreover, concerns about Mechanical Turk participants have been shown to be largely unfounded and manageable, resulting in data that are almost indistinguishable from data from more traditional laboratory experiments (Sprouse, 2011b). It has also begun

to be used in linguistics research specifically (Gibson, Piantadosi, & Fedorenko, 2011; Sprouse, 2011a; Sprouse & Almeida, unpublished data). All participants reported being native speakers of American English except one, whose survey was excluded. Surveys by repeat participants ($n = 9$) were excluded, using the method of checking for a repeated IP address (Berinsky, Huber, & Lenz, 2010). In all, 313 total surveys were included in the analysis.

Materials. The prosodic contrast in Experiment 1 was a combination of five different prosodic manipulations (for details, see Experiment 1); this study is referred to as PsychPool12345 because it used participants from the Psychology Subject Pool and contained all five prosodic manipulations. Experiment 2 tests subsets of those prosodic manipulations for an effect on interpretation, with each subset labeled as MTurk for its participant pool and the numbers corresponding to the prosodic contrasts it included (summarized in Table 4). So, for example, MTurk12 contains a prosodic contrast that differs only in S1 and S2 terminal pitch, whereas MTurk12345 replicates PsychPool12345 but with Mechanical Turk participants.

The studies that neutralized the contrast in sentence-final pitch (MTurk345, MTurk2) achieved this neutralization by flattening the final pitch to a constant Hz value.

Design. The design for each study in Experiment 2 was the same as in Experiment 1.

Procedure. The participants from Mechanical Turk took the same survey and received the same instructions as the Psychology Subject Pool participants in Experiment 1. Although it is unknown in what exact context they took the survey, they were instructed to be in a quiet environment and to have good headphones.

Predictions. The five Coord prosody manipulations (sentence-final rises on sentences 1 and 2, longer pauses and higher and louder versions of sentences 2 and 3) and any subsets of those prosodic manipulations were predicted to bias toward Coord interpretations. The five Subord prosody manipulations (a fall on sentence 1, a smaller rise on sentence 2, shorter pauses and lower and quieter versions of sentences 2 and 3) and any subsets of those manipulations were predicted to bias toward Subord interpretations.

Results

Results showed that 11 of the 313 participants in the data set performed poorly on the comprehension questions ($>20\%$ incorrect). After excluding these 11 participants, the final data set contained a total of 302 participants. Demographic

TABLE 4
Prosodic Manipulations in Each Study

Prosodic Manipulations in Each Study	*Coord Prosody*	*Subord Prosody*	*PsychPool12345*	*MTurk12345*	*MTurk12*	*MTurk1*	*MTurk2*	*MTurk345*
1: Terminal pitch on S1	Rise with slope of 1.6	Fall with slope of .75	X	X	X	X		
2: Terminal pitch on S2	Rise with slope of 1.3	Rise with slope of 1.1	X	X	X		X	
3: Pause duration between S1 and S2	900 ms	400 ms	X	X				X
4: Pause duration between S2 and S3	400 ms	20 ms	X	X				X
5: Mean pitch and intensity on S2 and S3	Mean pitch and intensity multiplied by 1.1	Mean pitch and intensity multiplied by .9	X	X				X

Note. X indicates the prosodic contrast was present in that study; an empty cell means Coord and Subord prosody were the same on that dimension.

data were collected from each participant at the end of the survey (Table 5). There was more variety among participants from Mechanical Turk in terms of age and education. The gender distribution was similar for both participant pools, with around twice as many women as men for most studies.

Results were modeled in the same way as in Experiment 1, using a mixed-effects model with binomial outcome fit using the lmer function in the R package lme4. The dependent variable was a binary of match versus mismatch (of prosody and interpretation). No categorical or continuous predictors were included in this initial model, testing only whether match likelihood was above chance. All models included random effects for subject and discourse (unless otherwise noted).

In Table 6, results are shown for the models for each study and for the percentage of matching interpretations for each prosodic condition and each study. These results show participants in studies MTurk12345, MTurk12, and MTurk1 are significantly more likely to choose the interpretation as predicted by the prosodic condition (i.e., match).

The other two studies (MTurk2 and MTurk345) showed no such effect. Figure 4 plots the match rate (on the *y*-axis) for each study (on the *x*-axis), with .50 indicating the responses are at chance level. The results of three studies (MTurk12345, MTurk12, and MTurk1) demonstrate the ability of prosody to affect listeners' discourse interpretation.

To test if Coord prosody and Subord prosody contributed differently to the overall effect, a predictor variable for Coord versus Subord prosody was added into the model. If prosody is a significant predictor of match, then the Coord and Subord prosody conditions were contributing different amounts to the overall effect. Results show that four studies showed a significant effect of prosody (MTurk12345: ß = –.282, SE = .075, $z = -3.748$, $p < .001$; MTurk1: ß = –.208, SE = .076, $z = -2.748$, $p = .006$; MTurk2: ß = –.324, SE = .073, $z = -4.433$, $p < .001$; MTurk345: ß = –.236, SE = .074, $z = -3.182$, $p = .001$) and one did not (MTurk12: ß = –.112, SE = .076, $z = -1.466$, $p = .143$). All four of these studies that showed an effect had a negative coefficient for prosody, such that going from Subord prosody (coded as 0) to Coord prosody (coded as 1) indicated a significant reduction in match likelihood. Therefore, for these studies match rate was higher in the Subord prosody condition. This effect is visible in Figure 5, which plots interpretation (on the *y*-axis) against Coord and Subord prosody separately (on the *x*-axis), clustered by study. Each cluster of two vertical bars corresponds to a study, with the right bar being Coord prosody and the left bar being Subord prosody. Mean match is plotted on the *y*-axis, with more matching interpretations making the bar higher.

Thus far, results have been discussed regardless of the relative ambiguity of the discourse. Although all 48 discourses are ambiguous, as shown by the

TABLE 5
Participant Demographics for Each Study

	MTurk12345	*MTurk12*	*MTurk1*	*MTurk2*	*MTurk345*	*PsychPool12345*
Total participants	60	58	60	63	61	40
Male	43%	33%	33%	32%	30%	33%
Multilingual (#yes)	10	11	3	13	7	14
Age						
Mean	35	34	33	35	36	18
Min–max	18–59	18–61	18–62	18–64	19–71	17–21
SD	11.03	11.83	10.44	13.24	11.95	.863
Education[a]	1: 1	1: 1	1: 3	1: 0	1: 1	1: 0
	2: 11	2: 9	2: 10	2: 10	2: 4	2: 20
	3: 18	3: 17	3: 21	3: 19	3: 20	3: 20
	4: 19	4: 19	4: 14	4: 20	4: 22	4: 0
	5: 4	5: 5	5: 4	5: 6	5: 4	5: 0
	6: 7	6: 7	6: 8	6: 9	6: 10	6: 0

[a]Legend for Education question: 1 = Did not complete high school; 2 = High school; 3 = Some undergraduate education; 4 = Undergraduate degree; 5 = Some graduate education; 6 = Graduate degree.

TABLE 6
Results Testing Whether Likelihood of Match Was Different From Likelihood of Mismatch

	Est.	*SE*	*z Value*	*Significance*	*Coord Interpretations After Hearing Coord Prosody (%)*	*Subord Interpretations After Hearing Subord Prosody (%)*
MTurk12345	.216	.048	$z(2880) = 4.475$	$p < .001$	51.9	58.8
MTurk12	.203	.051	$z(2784) = 3.958$	$p < .001$	53.6	56.3
MTurk1	.186	.067	$z(2880) = 2.782$	$p = .005$	51.9	56.8
MTurk2	−.021	.036	$z(3024) = -.583$	n.s.	45.4	53.5
MTurk345	−.008	.038	$z(2928) = -.216$	n.s.	46.9	52.7
PsychPool12345	.355	.089	$z(1920) = 3.998$	$p < .001$	59.5	57.0

Note. A column is included for each study as well as percentages for match for both Coord and Subord prosody.

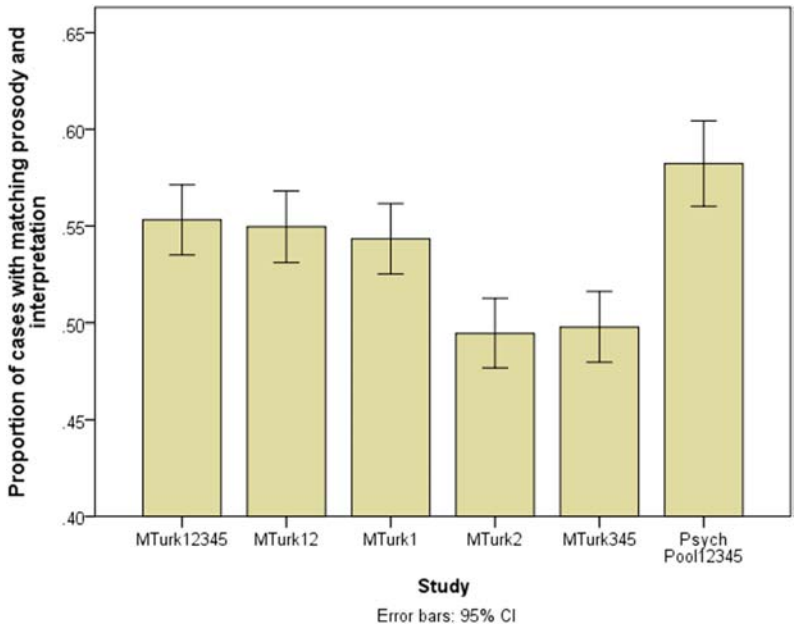

FIGURE 4 Mean match rates for each study (1 = match, 0 = mismatch), with 95% confidence intervals.

norming study, they still show some degree of underlying bias to a Coord or Subord interpretation. Each discourse was assigned an ambiguity score calculated as the absolute value of the difference between the number of people who chose Coord and Subord interpretations in the norming study. This variable captures how equibiased the ambiguity is, regardless of whether the bias is toward Coord or Subord. If this underlying ambiguity score predicts match likelihood, then the underlying ambiguity of a discourse had an impact on prosody's effect on interpretation.

A continuous variable for a discourse's underlying ambiguity was included in the mixed model with match as a dependent variable. Results show an effect for studies MTurk12345 (ß = .010, SE = .005, $z = -1.981$, $p = .048$) and MTurk12 (ß = .010, SE = .005, $z = -2.059$, $p = .040$) but no others (MTurk1: ß = .005, SE = .005, $z = .929$, $p = .353$; MTurk2: ß = –.006, SE = .005, $z = -1.183$, $p = .237$; MTurk345: ß = .000, SE = .005, $z = -.054$, $p = .957$). If all three Mechanical Turk studies that showed an effect of prosody on interpretation (MTurk12345, MTurk 12, MTurk1) are included at once, underlying ambiguity shows a trend but does not come out as significant (ß = –.005, SE = .003, $z = -1.727$, $p = .084$). The fact that the effect for

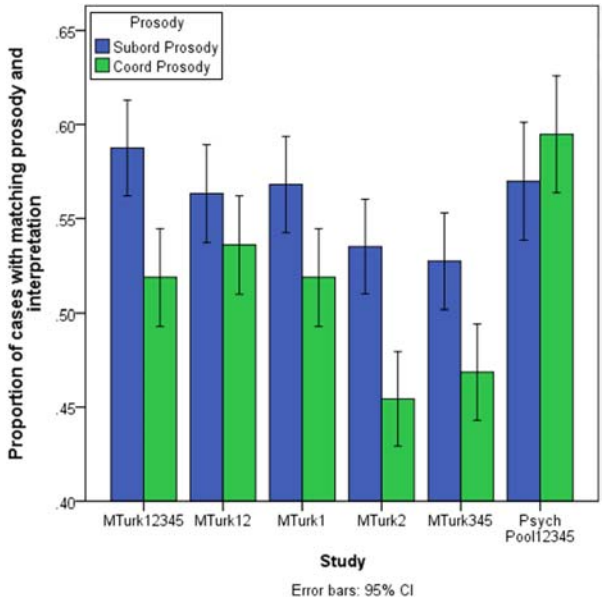

FIGURE 5 This graph plots each study on the *x*-axis and mean interpretation on the *y*-axis (1 = Coord, 0 = Subord), with 95% confidence intervals. The right column for each study indicates results for Coord prosody, whereas the left column indicates results for Subord prosody.

MTurk12345 and MTurk12 gets reduced to a trend when combined with MTurk1 suggests that those effects are weak, if present at all. Another interpretation is that the underlying ambiguity effect only shows up when multiple prosodic contrasts (MTurk12 and MTurk12345) are present but not when only one is included (MTurk1). Perhaps having more information in the prosody interacts with underlying ambiguity in a way that leads listeners to exploit the underlying bias more. The role of the discourses' underlying ambiguity could have been clearer if instead of the 48 most ambiguous discourses, a wider range of bias in the discourses was used.

Although the results above show prosody affecting interpretation, a separate concern is whether participants changed their behavior over the course of the experiment. To test this question, a continuous variable was included in the model that coded whether a judgment was made in the first, second, third, or fourth quarter of the experiment. Presentation quarter was not predictive of match in any of the studies (MTurk12345: ß = –.060, SE = .034, $z = -1.779$, $p = .075$; MTurk12: ß = –.053, SE = .034, $z = -1.553$, $p = .120$; MTurk1: ß = .051, SE = .036, $z = 1.429$, $p = .153$; MTurk2: ß = –.031, SE = .033, $z = -.944$, $p = .345$; MTurk345: ß = –.005, SE = .034, $z = -.161$, $p = .872$).

This suggests that participant behavior with respect to prosody's effect on interpretation did not change over the course of the study.

In these studies, the participant's interpretation of the discourse was elicited with two types of questions, counterbalanced such that half of the questions that participants saw would have a "yes" answer correspond to a Coord interpretation and the other half would have a "yes" answer correspond to a Subord interpretation. Overall, there was a bias toward answering "yes" (ß = –.221, SE = .028, $z = -7.795$, $p < .001$), where across all studies 55% of answers were "yes" answers. This affirmative answer bias did not differ across the prosodic conditions: When prosody is entered in the models as a predictor of a "yes" response, the result is not significant overall (ß = .049, SE = .034, $z = 1.395$, $p = .163$) or in any individual study (MTurk12345: ß = .049, SE = .075, $z = .648$, $p = 517$; MTurk12: ß = .089, SE = .076, $z = 1.162$, $p = .245$; MTurk1: ß = .107, SE = .075, $z = 1.421$, $p = .155$; MTurk2: ß = –.016, SE = .073, $z = -.218$, $p = .827$; MTurk345: ß = .011, SE = .075, $z = .150$, $p = .880$). Therefore, the type of question did not affect listeners' use of prosody in interpretation.

There is also substantial variability from participant to participant. Instead of listing participant interpretation means for each study ($n > 300$), Figure 6 plots a line for each study with participants sorted from most Coord-preferring to most Subord-preferring. The graph shows consistent variation within studies in terms of individual preferences. The relationship between these preferences and participants' use of prosody was explored by checking for a correlation between participants' interpretation preferences and match rate. Results showed a significant Pearson correlation for MTurk12345 (ρ = .053, $p = .004$), MTurk2 (ρ = .048, $p = .008$), MTurk345 (ρ = .052, $p = .005$), but not MTurk12 (ρ = .010, $p = .610$) or MTurk1 (ρ = –.010, $p = .598$). The correlation values are low even when significant. The r^2 values for each study are also quite low (MTurk12345: $r^2 = .003$; MTurk12: $r^2 < .001$; MTurk1: $r^2 < .001$; MTurk2: $r^2 = .002$; MTurk345: $r^2 = .003$), suggesting any patterns that show up are not particularly explanatory.

As discussed in Experiment 1, naturalness judgment data showed the Coord versions of the discourses with all five manipulations (PsychPool12345, MTurk12345) had higher naturalness scores than the Subord versions. To test whether interpretation in MTurk12345 was predictable by naturalness, a mixed model was fitted with interpretation as a binary outcome, subject as a random effect, prosody (Coord or Subord) as a categorical predictor, and naturalness as a continuous predictor. Results show match was predicted by prosody (ß = .426, SE = .079, $z = 5.376$, $p < .001$) but not naturalness (ß = .103, SE = .110, $z = .937$, $p = .349$). These results replicate the results for Experiment 1, showing that the effect of prosody on interpretation holds even when controlling for naturalness and is not reducible to a naturalness difference.

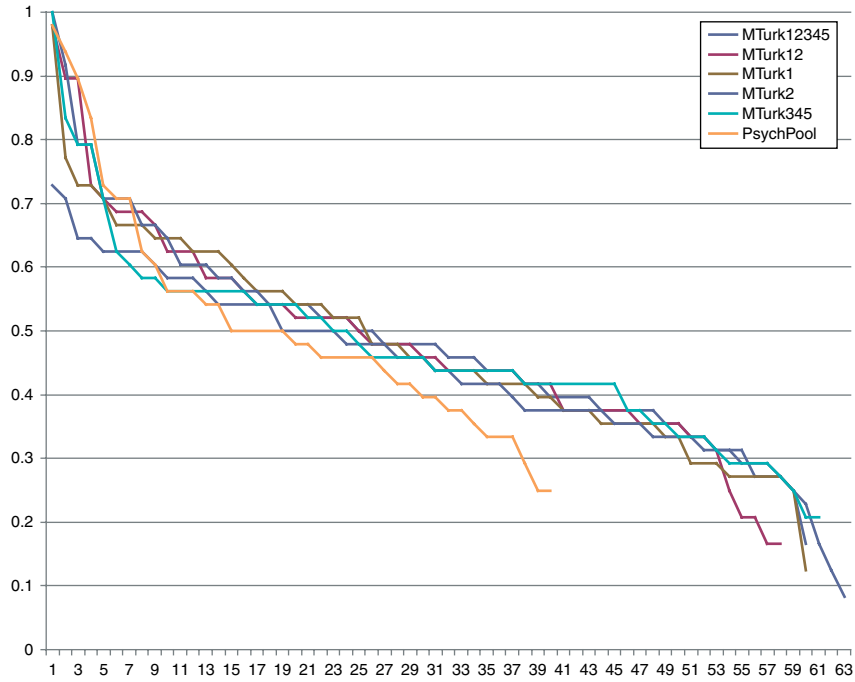

FIGURE 6 A graph of participant interpretation means (1 = Coord, 0 = Subord), sorted from high to low. Each line refers to a different study.

For each of the four studies that showed an effect of prosody on interpretation, between three and five participants had a match rate above 70%. None of the participant demographic data predicted match rate, so what led these individuals to behave more as predicted is unknown. To see if the effect of prosody on interpretation depended on just a few participants, three participants with the highest match rates were excluded from each study, and the statistical model was fitted again. All four studies continue to show an effect of prosody on interpretation (MTurk12345: ß = .219, SE = .049, $z = 4.45$, $p < .001$; MTurk12: ß = .203, SE = .051, $z = 3.958$, $p < .001$; MTurk1: ß = .186, SE = .067, $z = 2.782$, $p = .005$; PsychPool12345: ß = .355, SE = .089, $z = 3.998$, $p < .001$), showing that the effect is not dependent solely on the behavior of three participants.

Confidence. Participants' confidence ratings were modeled with a linear mixed model with random effects for both subject and discourse and confidence as a continuous dependent variable (same as in Experiment 1). These models

test whether participants were more confident in their judgments when their interpretation matched what was predicted from the prosody. One study showed an effect of match (prosody/interpretation) on confidence (MTurk1: ß = 2.037, SE = .597, $t = 3.41$, pMCMC = .001), whereas one other showed a trend (MTurk12345: ß = 1.050, SE = .595, $t = 1.77$, pMCMC = .078). The other studies showed no such effect of match on confidence (MTurk12: ß = .520, SE = .576, $t = .90$, pMCMC = .376; MTurk2: ß = .304, SE = .592, $t = .51$, pMCMC = .605; MTurk345: ß = .331, SE = .601, $t = .55$, pMCMC = .561). It is not surprising there was no effect of match on confidence for MTurk2 and MTurk345 because those studies had match likelihood at chance. If there is no indication participants are using prosodic information in their interpretation, it is less likely their confidence would be affected by whether their interpretations matched the prosodic condition. By contrast, MTurk12345 and MTurk12 showed a higher than chance match rate but no effect of match on confidence. A combined data set including all studies that had greater than chance match rates (MTurk12345, MTurk12, MTurk1, PsychPool12345) showed a significant effect of match on confidence (ß = 1.295, SE = .302, $t = 4.29$, pMCMC $< .001$). This suggests match does predict confidence, but the effect may require enough data to

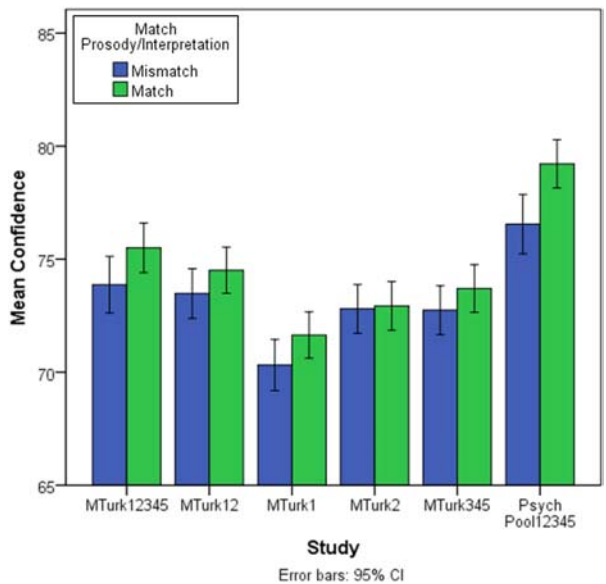

FIGURE 7 This graph plots confidence on the y-axis, with each study on the x-axis. Each study is split into a bar on the right for match results and a bar on the left for mismatch results. Error bars indicate 95% confidence intervals.

have the statistical power to detect it. Figure 7 shows participant confidence for matches and mismatches between prosody and interpretation across all studies. Confidence is higher in the match condition for all studies.

Discussion

In all, three studies found an effect of prosody on interpretation (MTurk12345, MTurk12, MTurk1) and two did not (MTurk2, MTurk345). The result for MTurk12345 replicates the results from Experiment 1 but with Mechanical Turk participants. All studies that found an effect of prosody on interpretation had manipulation 1 (a rising vs. falling terminal pitch contour contrast on sentence 1) and those that found no effect did not have this manipulation. Moreover, MTurk1 had only this contrast, indicating a S1-final rise/fall contrast alone could bias interpretation. This distribution suggests that the prosodic manipulation driving the overall effect was the rising versus falling pitch contour at the end of the first sentence.

In addition to the general effect of prosody on interpretation, it appears that Subord prosody tends to have a higher match rate than Coord prosody (see Figure 5 comparing match rate for all studies). To some degree, this may be an artifact of the set of normed discourses. Although the 48 most ambiguous discourses were used from the 102 discourses in the norming study, those 48 were slightly skewed toward Subord interpretations: 18 discourses had preferred Coord interpretations, 29 had preferred Subord interpretations, and 1 was equibiased. The result showing match rate being higher in the Subord prosody condition, suggesting an overall preference for Subord interpretations, may simply be a result of this set of discourses showing a small bias toward Subord interpretations. On the other hand, Experiment 1 did not show a higher match rate for Subord prosody. The results for MTurk2 and MTurk345, which showed no effect of prosody on interpretation, may help explain why. In Figure 5, these two studies have match rates close to the Subord match rates in the other studies. Perhaps the underlying bias for the discourses skews somewhat towards Subord interpretations, but Coord prosody (rising pitch on sentence 1) can push listeners toward Coord interpretations. If this is the case, then the effect of Coord prosody in PsychPool12345 was stronger than in the other studies.

Although it appears manipulations 2, 3, 4, and 5 had no independent effect on interpretation, it is important to exercise caution in the interpretation of these null effects. The two pause duration contrasts did not affect interpretation, but this does not mean that no pause duration contrasts would. One possibility is that listeners can hear the pause duration contrast and are not assigning any meaning to it. Another possibility is that listeners cannot even hear the contrast. In this case, they may assign meanings to some pause duration contrasts in discourse interpretation, but not to contrasts they cannot hear. Another possibility is that listeners make a decision at the end of sentence 1 about whether sentence 1

contains disambiguating material or not. The only prosodic contrast on sentence 1 is the rise/fall contrast of manipulation 1, with the other contrasts occurring between sentences, or on sentences 2 or 3. The null effects of manipulations 2, 3, 4, and 5 may then be a result of listeners already having made their interpretation and these manipulations being unable to override an earlier decision.

Another factor is the naturalness of both the discourse texts and the synthetically manipulated speech. Although naturalness of the written texts alone was not examined, naturalness scores were collected for the speech samples. While the manipulations increased naturalness for Coord prosody and decreased it for Subord prosody, these differences did not impact prosody's effect on interpretation.

One question that remains open is the role of individual variability, which was not illuminated by any of the demographic data. Another open question is how prosodic contrasts other than those examined here would affect interpretation. These questions are relegated to future research.

GENERAL DISCUSSION

The results of the experiments presented here make clear that prosody, in addition to lexical cues, can influence listeners' interpretation of hierarchical discourse. Although previous discourse prosody perception research (Mayer et al., 2006; Silverman, 1987) used ambiguous discourses whose meanings could be described as either hierarchy or distance contrasts, the discourses used in Experiments 1 and 2 contain a hierarchical coordination/subordination ambiguity while controlling for distance effects. This difference is relevant also because the effects are achieved with different prosody. Silverman and Mayer et al. achieve their effects by cuing larger boundaries (e.g., with whole sentences compressed or expanded or with preboundary lowering and postboundary raising), whereas in Experiments 1 and 2 a rising versus falling pitch contour achieved the disambiguation effect. Prosodic boundary size and terminal pitch contours may have different impacts on interpretation.

Another difference between this study and both Mayer et al. (2006) and Silverman (1987) is in the kind of interpretation required of listeners. Mayer et al. use an ambiguous pronoun, whereas the example discourse in Silverman has a universally quantified phrase with ambiguous domain restriction ("all materials"). Using pronouns and quantifier phrases as a proxy for discourse interpretation has the benefit of providing an easily interpretable task for participants, but it raises the question of whether the results are strictly about prosody's effect on the disambiguation of discourse structure or are also influenced by other factors that are known to affect the interpretation of pronouns (Kehler, Kertz, Rohde, & Elman, 2008) and quantifier phrases (von Fintel, 1994). In Experiments 1 and 2, the participants were asked questions that reveal their interpretation of how the

sentences fit together (i.e., did the events in sentences 2 and 3 happen during the event in sentence 1 or not). This approach elicits interpretations that depend on whole discourse segments and relations between them, relying less on the meaning of an individual pronoun or lexical item. This may be a more direct means of identifying prosody's effect on discourse interpretation.

Beyond methodological differences, the results of both Silverman (1987) and Mayer et al. (2006) suggest a cumulative effect of prosodic cues to discourse structure that did not appear in Experiments 1 and 2. Although raw scores suggest a slightly stronger effect when more manipulations were present, the only significant prosodic manipulation appeared to be the terminal pitch contour on the first sentence.

The comparison of these studies should also consider that Mayer et al. conducted their experiments in German with native German speakers and Silverman's participants were all native speakers of British English. Given that we know so little about discourse prosody perception at all, much less across languages or language varieties, it seems important to keep in mind that speakers and listeners of German, British English, and American English may behave differently.

A limitation of the set of experiments presented here is that it is difficult to infer relative contributions of individual prosodic contrasts or to interpret null results. On the one hand, a factorial design would better reveal how each prosodic contrast contributed to judgments and how they interacted. On the other hand, a factorial design with five prosodic contrasts is a higher-order design requiring a large amount of data that could be difficult to interpret. The research pursued here was considered more manageable, but leaves for future work questions about null effects and the relative contributions of different prosodic contrasts.

Although rising pitch at the end of sentence 1 biased interpretation toward discourse coordination in three-sentence Coord/Subord discourse ambiguities, an open question is how much this connection between discourse and prosody can be generalized to other structures. There are many possible ambiguous structures, with different amounts of preceding material, following material, or ambiguous material, that could be tested for prosody's ability to disambiguate. Moreover, because the discourses used in Experiments 1 and 2 have been constructed to make the relations between sentences consistently either Narration (the coordinating relation) or Elaboration (the subordinating relation), it raises the question of whether rising pitch could similarly bias interpretation towards other kinds of coordinating relations (e.g., Result, Contrast, Parallel or Continuation[3]).

In addition, the generalizability of these results to more natural communication behavior may depend on the relationship between ambiguity and perceived naturalness. Speakers may try to avoid unnecessary ambiguity in

[3]These relation names are drawn from Segmented Discourse Representation Theory. For a fuller discussion and definitions of these and other discourse relations, see Asher and Lascarides (2003).

their speech, meaning the ambiguous discourses could be perceived to be less natural. Additionally, these experiments used synthetically manipulated read speech in a forced choice task, distinguishing them from more natural communication.

In conclusion, the experiments presented here demonstrated that listeners can use a contrast between sentence-final rising and falling pitch to help disambiguate Coord/Subord discourse ambiguities. These results show prosody can bias the interpretation of discourse not just in terms of reference resolution or quantifier scope ambiguities but also in terms of the hierarchical structure of discourse, suggesting prosody may have a larger role to play in the comprehension of discourse than has thus far been understood.

ACKNOWLEDGMENTS

This article is based on data also in the doctoral dissertation of Joseph Tyler.

FUNDING

This research was undertaken with the support of a Humanities Candidacy Research Fellowship and a Rackham One-Term Dissertation Fellowship from the University of Michigan.

REFERENCES

Allbritton, D. W., McKoon, G., & Ratcliff, R. (1996). Reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 714–735.

Asher, N., & Lascarides, A. (2003). *Logics of conversation*. Cambridge, UK: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Maechler, M., & Bolker, B. (2013). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-2.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2010). *Using mechanical turk as a subject recruitment tool for experimental research*. Retrieved from http://huber.research.yale.edu/materials/26_paper.pdf

Boersma, P., & Weenink, D. (2009). *Praat: Doing phonetics by computer*. Retrieved from http://www.praat.org

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

den Ouden, H., Noordman, L., & Terken, J. (2009). Prosodic realizations of global and local structure and rhetorical relations in read aloud news reports. *Speech Communication*, *51*, 116–129.
Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass*, *5*, 509–524.
Grosz, B., & Hirschberg, J. (1992, October). *Some intonational characteristics of discourse structure*. Paper presented at the 2nd International Conference on Spoken Language Processing, Banff.
Hirschberg, J., & Grosz, B. (1991, February). *Intonational features of local and global discourse structure*. Paper presented at the Speech and Natural Language Workshop.
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.
Kamalski, J., Sanders, T., & Lentz, L. (2008). Coherence marking, prior knowledge, and comprehension of informative and persuasive texts: Sorting things out. *Discourse Processes*, *45*, 323–345.
Kehler, A., Kertz, L., Rohde, H., & Elman, A. J. (2008). Coherence and coreference revisited. *Journal of Semantics (Special Issue on Processing Meaning)*, *25*, 1–44.
Kittur, A., Chi, E. H., & Suh, B. (2008, April). *Crowdsourcing user studies with Mechanical Turk*. Paper presented at the 26th Annual SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy.
Knott, A., & Mellish, C. (1996). A feature-based account of the relations signalled by sentence and clause connectives. *Language and Speech*, *39*, 143–183.
Lehiste, I. (1982). Some phonetic characteristics of discourse. *Studia Linguistica*, *36*, 117–130.
Mayer, J., Jasinskaja, E., & Kölsch, U. (2006, September). *Pitch range and pause duration as markers of discourse hierarchy: Perception experiments*. Paper presented at INTERSPEECH 2006-ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA.
Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.
Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, *90*, 2956–2970.
Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413–425.
R Development Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org
Rohde, H., Levy, R., & Kehler, A. (2011). Anticipating explanations in relative clause processing. *Cognition*, *118*, 339–358.
Schafer, A. J., Speer, S. R., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research*, *29*, 169–182.
Schiffrin, D. (1987). *Discourse markers*. Cambridge, UK/New York, NY: Cambridge University Press.
Silverman, K. E. A. (1987). *The structure and processing of fundamental frequency contours*. Unpublished doctoral dissertation, University of Cambridge, Cambridge, UK.
Smith, C. L. (2004). Topic transitions and durational prosody in reading aloud: Production and modeling. *Speech Communication*, *42*, 247–270.
Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, *48*, 103–130.
Sprouse, J. (2011a). A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language*, *87*, 274–288.
Sprouse, J. (2011b). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*, 155–167.
Tyler, J. (2012). *Discourse prosody in production and perception*. Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Tyler, J. (2013). Prosodic correlates of discourse boundaries and hierarchy in discourse production. *Lingua*, *133*, 101–126.

Tyler, J., Kahn, J., & Arnold, J. (2011, September). *Speakers use prosody to communicate discourse structure, and listeners use that prosody in comprehending discourse structure*. Paper presented at Experimental and Theoretical Advances in Prosody (ETAP) 2, Montreal, Canada.

von Fintel, K. (1994). *Restrictions on quantifier domains* (Doctoral dissertation). University of Massachusetts, Amherst. Retrieved from http://www.semanticsarchive.net

Webber, B., Stone, M., Joshi, A. K., & Knott, A. (2003). Anaphora and discourse structure. *Computational Linguistics*, *29*, 545–587.

Yule, G. (1980). Speakers' topics and major paratones. *Lingua*, *52*, 33–47.