Research Report

The Old and Thee, uh, New

Disfluency and Reference Resolution

Jennifer E. Arnold, Michael K. Tanenhaus, Rebecca J. Altmann, and Maria Fagnano

University of Rochester

ABSTRACT—*Most research on the rapid mental processes of on-line language processing has been limited to the study of idealized, fluent utterances. Yet speakers are often disfluent, for example, saying "thee, uh, candle" instead of "the candle." By monitoring listeners' eye movements to objects in a display, we demonstrated that the fluency of an article ("thee uh" vs. "the") affects how listeners interpret the following noun. With a fluent article, listeners were biased toward an object that had been mentioned previously, but with a disfluent article, they were biased toward an object that had not been mentioned. These biases were apparent as early as lexical information became available, showing that disfluency affects the basic processes of decoding linguistic input.*

As utterances unfold over time, listeners rapidly extract linguistic information, making provisional commitments to meaning and structure that are remarkably time-locked to the input. However, the evidence for this conclusion comes from studies using materials that are recorded to approximate an idealized fluent delivery. In contrast, the language that speakers typically generate is rife with hesitations, false starts, repetitions, *uh*s, *um*s, and elongated forms of words, such as saying "thee" (rhyming with "tree") for *the*.

How do deviations from an ideal delivery affect real-time language comprehension? One possibility is that they introduce noise that needs to be removed so the listener can recover the linguistic message using "normal" comprehension routines, viz., routines designed to decode well-formed, fluently delivered input. A second possibility is that disfluencies create probabilistic constraints that listeners exploit, just as they use other types of probabilistic constraints, such as the relative frequency with which words and phrases co-occur.

This second view is more consistent with results from a few recent studies, which suggest that disfluency does not always interfere with comprehension, and in some cases might help listeners (Brennan & Schober, 2001; Fox Tree, 1995, 2001). For example, Fox Tree (2001) argued that *uh* signals a brief delay and serves to heighten listeners' attention to upcoming speech. Disfluency also affects parsing preferences, at least when listeners are given time to think about an utterance (Bailey & Ferreira, 2003), and may direct listeners' attention to unfamiliar referents (Barr, 2001a, 2001b). However, these studies leave open the important question that we address in this report: Is the information provided by disfluency used on-line to guide the interpretations that listeners initially consider as they process spoken language?

We focus on the distinction between reference to information previously mentioned in a discourse, labeled "old" or "given" information, and information that is new to a discourse—a distinction that is central to models of conversation and discourse (e.g., Chafe, 1976, 1994; Haviland & Clark, 1974; Prince, 1992). We demonstrate that listeners use information provided by fluent *the* (rhyming with "duh") and its commonly occurring disfluent variant, *thee uh*, to guide expectations about whether a speaker is likely to refer to something given or to something new. Moreover, we show that fluency information is integrated with the unfolding acoustic input during the earliest moments of word recognition.

Speakers are more likely to say "thee" when they are experiencing difficulty planning an upcoming segment, as evidenced by delays, repeats, or fillers like *uh* (Clark & Wasow, 1998; Clark & Fox Tree, 2002; Fox Tree & Clark, 1997). Referring to something that is new is likely to be more difficult than referring to something that is given because the conceptual, lexical, and phonological representations for new referents have not been as recently accessed. Indeed, speakers are more likely to be disfluent when referring to something new than when referring to something given (Arnold & Tanenhaus, in press). Therefore, we reasoned that listeners might expect a noun that follows "thee uh" to refer to something new rather than something given, a prediction we confirmed in an off-line study. In that study, participants viewed displays like the one illustrated in Figure 1 and heard auditory instructions such as "Put the grapes above the candle. Now put {thee uh/the}. . . ." The instructions were originally recorded in their entirety but were digitally truncated after either "Now put" (short-instruction condition) or "Now put {the/thee uh}" (long-instruction condition). Eight experimental items were combined with four fillers, which were truncated in the middle of the target word (e.g., "Now put the sa-"). Twenty-four participants were asked to identify what they thought the speaker was about to say next, by circling the label of the object on a piece of paper. After hearing the fragment "Now put," participants chose one of the new objects 63% of the time (68% in the disfluent condition, $SE = 8\%$, and 58% in the fluent condition, $SE = 8\%$). This pattern changed when listeners heard an article, however. The proportion of new objects chosen dropped to 35%

Address correspondence to Jennifer Arnold, University of Rochester, Department of Brain and Cognitive Sciences, Meliora Hall 495, RC Box 270268, Rochester, NY 14627; e-mail: jarnold@bcs.rochester.edu.

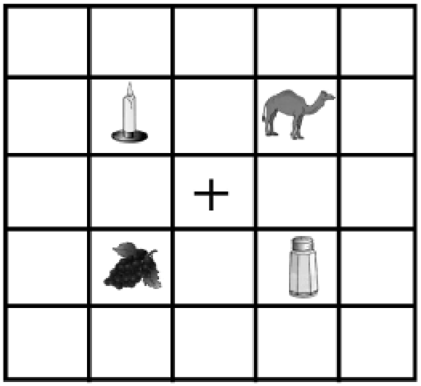

Fig. 1. Example of the visual displays used in the experiment.

($SE = 7\%$) following "the" and increased to 83% ($SE = 8\%$) following "thee uh," resulting in an interaction between disfluency and instruction length, $F_1(1, 23) = 12.0$, $p < .005$, $\eta_p{}^2 = .34$; $F_2(1, 3) = 15.7$, $p < .05$, $\eta_p{}^2 = .84$).[1]

Thus, the fluency of the article affects listeners' expectations about given and new referents. Do these expectations influence how listeners interpret linguistic input during real-time language comprehension? We investigated this question by exploiting the brief ambiguity created when names of potential referents begin with the same sequence of sounds (e.g., in the case of Fig. 1, the initial consonant and much of the first vowel in "candle" and "camel" are acoustically indistinguishable). As listeners hear a word, multiple lexical candidates become partially activated and compete for recognition (e.g., Marslen-Wilson, 1987). Competition is strongest for words that overlap at their onset—so-called cohort competitors.

Striking evidence for cohort competition comes from studies monitoring eye movements as people follow spoken instructions (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Approximately 200 ms after the onset of a spoken word, or after about the time it takes to program and launch a saccadic eye movement, people begin to look more at pictures with names that match the input than at pictures with unrelated names (Allopenna, Magnuson, & Tanenhaus, 1998). The time course of fixations closely reflects the degree of activation for lexical candidates and thus can be used to trace the time course of lexical access in continuous speech.

If the form of an article influences the earliest moments of processing, listeners should be biased toward cohort competitors that are given when a noun is preceded by a fluent article and toward cohort competitors that are new when a noun is preceded by a disfluent article. For example, consider what might happen if listeners view the display shown in Figure 1 and hear the following instructions: "Put the grapes above the candle. Now put the camel. . . ." About 200 ms after the onset of the word *camel*, listeners should initially look to the candle more than to the camel if the instruction is spoken fluently. But if the second instruction is spoken disfluently ("Now put thee uh camel. . . ."), they should initially look to the camel.

METHOD

Participants

Twenty-four native English speakers from the University of Rochester community participated in exchange for $7.50.

Materials and Procedure

One of the challenges of studying disfluency is to maintain experimental control and at the same time have the disfluent instructions sound natural and be plausibly attributable to difficulty in generating a referring expression. The first author recorded the stimuli, modeling the disfluencies on naturally occurring tokens. However, we told participants that the recordings came from a participant who had given these instructions to a partner in a previous experiment. Participants were shown examples of the graphic displays that had supposedly been shown to the speaker to indicate what she should say; the explanation emphasized that the speaker was told what to say, but not how to say it. A postexperiment questionnaire verified that all the participants believed the story. (In the off-line experiment described in the introduction, the same story was used. In this case, 2 of 24 subjects did not believe the story and were replaced.)

Using an Applied Scientific Laboratories head-mounted eyetracker, we monitored participants' eye movements while they viewed computer displays like the one shown in Figure 1 and followed auditory instructions to move objects around with a mouse. Two objects in each display were cohort competitors (e.g., candle-camel), and two were distractors with no phonetic overlap with the cohort competitors. The first instruction sentence referred to one of the cohort competitors and therefore established it as given and the other cohort competitor as new. The given cohort competitor was always the second noun phrase in this sentence, making it given but not highly focused (e.g., "Put the grapes below the candle."). Highly focused entities are preferentially referred to with pronouns (Brennan, 1995; Gordon, Grosz, & Gilliom, 1993) or, if a noun phrase is used, reduced stress (Dahan, Tanenhaus, & Chambers, 2002). In contrast, accented noun phrases are preferred for reference to both new entities and given but unfocused entities.

We therefore used accented noun phrases to refer to the target object, which could be either the given or the new cohort object. This reference occurred in the second sentence, which was recorded to achieve a natural-sounding fluent or disfluent instruction: "Now put {the/thee uh} candle. . . ." Discussions of disfluency tend to focus on the form of the disfluent words, for example, whether "the" is pronounced "thuh" or "thee" and whether or not it is followed by an "uh" or "um." However, disfluent forms are often accompanied by prosodic features in surrounding regions, in particular, by word lengthening (e.g., Bell et al., 2003) and pauses (e.g., Fox Tree & Clark, 1997). Our disfluent stimuli, like naturally disfluent utterances, were characterized by both the article "thee uh" and a prosodic contour on the words "Now put" that gave the impression that the speaker was thinking: longer duration on both words (average duration = 399 ms for fluent utterances and 639 for disfluent utterances) and a higher pitch excursion on "Now" (average pitch range = 42 Hz for fluent utterances and 115 Hz for disfluent utterances).

The first and second sentences were cross-spliced to produce four conditions: fluent-given target, fluent-new target, disfluent-given target, and disfluent-new target; the 16 experimental items were rotated through these conditions. Which item in each cohort was the target

[1] In all the item analyses reported, one factor was item group, the group of items that rotated together through each condition across the different presentation lists. The effects of this factor reflect differences between the subjects contributing to the different conditions for a particular item.

(e.g., camel vs. candle) was also manipulated as a control variable. The experimental items were combined with 32 fillers, and eight randomly ordered lists were created. As a control for order effects, half the subjects saw the items in reverse order. The fillers all contained cohort competitors but were designed to remove any contingencies that would allow participants to develop experiment-specific strategies to anticipate the target word. The visual stimuli were versions of the Snodgrass and Vanderwart (1980) pictures, colored and normed for frequency, visual complexity, and familiarity (Rossion & Pourtois, 2001). These dimensions were counterbalanced across items, so that the average of each property was the same for targets and their competitors (cf. Dahan, Magnuson, & Tanenhaus, 2001). The location of the cohort objects was also counterbalanced across items.

RESULTS AND DISCUSSION

The form of the article clearly affected how listeners processed the target noun as it unfolded over time. Figure 2 presents the percentage of looks to each object (target, cohort competitor, and unrelated pictures), in 33-ms time slices. A "look" began when the participant launched an eye movement to an object and continued during the time spent fixating that object (i.e., until a new eye movement was launched).

Consider the fluent conditions first. When the target was the new entity and its cohort competitor was the given entity, looks to the competitor initially rose more quickly than looks to the target or to the unrelated pictures, beginning about 200 ms after the onset of the target word. In contrast, there were few looks to the target's cohort competitor when the target was given and the competitor was new. A similar pattern was observed by Dahan et al. (2002) for fluent utterances with accented nouns. In the disfluent conditions, however, there were more looks to the target's competitor when it was new (i.e., when the target was given) than when it was given (i.e., when the target was new). This effect emerged 200 ms after the onset of the noun, establishing that disfluency influences the processes of word recognition and reference comprehension as early as the acoustic input occurs.

To evaluate the reliability of these results, we computed the proportions of looks to the competitor in each condition for a time slice from 200 to 600 ms after the onset of the target noun (see Table 1). These proportions were submitted to participant and item analyses of variance that included disfluency, referent (given vs. new), target (camel vs. candle), and item group (item analysis only) as independent variables. The results revealed an interaction between disfluency and referent, $F_1(1, 23) = 11.398, p < .005, \eta_p^2 = .33; F_2(1, 12) = 13.091, p < .005, \eta_p^2 = .52$.

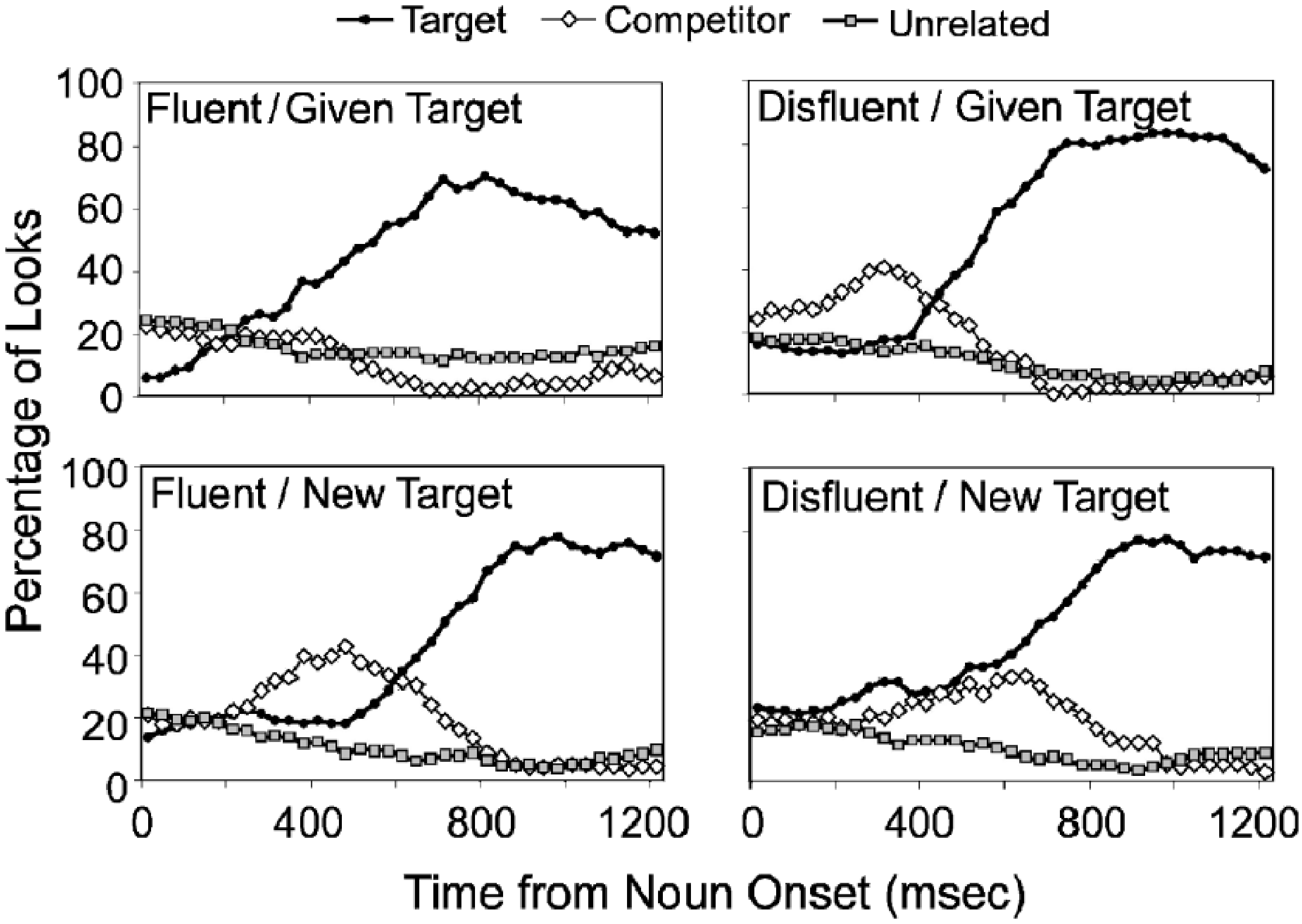

Fig. 2. Percentage of looks to the target, the target's competitor, and unrelated objects as a function of time, beginning at the onset of the target noun. Each panel shows results for a different combination of fluency (fluent, disfluent) and target (given, new).

TABLE 1

Proportion of Looks Toward the Competitor 200 to 600 Ms After the Onset of the Target Noun

Fluency of instruction	Given target (new competitor)	New target (given competitor)
Disfluent	.30 (.04)	.25 (.04)
Fluent	.15 (.03)	.34 (.03)

Note. Standard errors are in parentheses.

In sum, the form of the article influenced the earliest moments of lexical processing and reference resolution, confirming that these effects emerge during the initial on-line processing of the input. Fluent articles facilitated comprehension of expressions referring to a previously mentioned entity, whereas disfluent articles facilitated comprehension of expressions referring to new entities.

GENERAL DISCUSSION

Speakers are more likely to be fluent when mentioning something given and disfluent when mentioning something new. The results presented here show that listeners are sensitive to this contingency, combining information carried by the form of the article with the subsequent acoustic input during the earliest moments of lexical access and reference resolution. When the article was fluent, listeners were biased toward considering a previously mentioned, given entity that was consistent with the unfolding acoustic input. When the article was disfluent, this bias was reversed. In an extension of this result, we have recently shown that following a disfluent article, listeners also expect reference to novel objects that have no conventional name. For example, on hearing the disfluent "thee uh green. . . ," at the onset of "green" listeners look more at a green squiggly shape than at a known green object, like a green balloon. In contrast, if the article is fluent, they wait until the name is disambiguated. Together, these results highlight the importance of considering disfluency as an informative aspect of the speech signal, rather than as noise (also see Ferreira & Bailey, 2004). This raises important questions about how language-processing researchers should view the standard distinction between linguistic and paralinguistic aspects of the message.

Future research is necessary to address the mechanisms underlying the effects we have reported. One possibility is that listeners implicitly encode the statistical patterns created by different pronunciations of *the* in the same way that they encode and use other types of probabilistic constraints in real-time comprehension (MacDonald, Pearlmutter, & Seidenberg, 1994; Tanenhaus & Trueswell, 1995). An alternative (but not mutually exclusive) possibility is that listeners identify possible causes of a speaker's disfluency (e.g., referring to new information) and use these inferences during real-time comprehension. In this case, a disfluent article might not create an expectation for a new entity if it followed something that might plausibly distract the speaker, such as a loud noise or a flash of light. Research on these issues should shed light on how multiple sources of information are used in language comprehension, as researchers move beyond scripted, fluent utterances to study how more natural utterances are understood.

Acknowledgments—We are grateful to Dana Subik for his help in collecting and coding the eye movement data and to Katherine Crosswhite for her help with the prosodic analyses of the stimuli. Sarah Brown-Schmidt, Mikhail Masharov, and Duane Watson each made helpful suggestions. This research was partially supported by National Institutes of Health Grants HD-27206 to M. Tanenhaus and HD-41522 to J. Arnold.

REFERENCES

Allopenna, P.D., Magnuson, J.S., & Tanenhaus, M.K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419–439.

Arnold, J.E., & Tanenhaus, M.K. (in press). Disfluency effects in comprehension: How new information can become accessible. In E. Gibson & N. Perlmutter (Eds.), *The processing and acquisition of reference*. Cambridge, MA: MIT Press.

Bailey, K.G.D., & Ferreira, F. (2003). Disfluencies influence syntactic parsing. *Journal of Memory and Language*, *49*, 183–200.

Barr, D.J. (2001a, November). *Paralinguistic correlates of discourse structure*. Poster presented at the annual meeting of the Psychonomic Society, Orlando, FL.

Barr, D.J. (2001b). Trouble in mind: Paralinguistic indices of effort and uncertainty in communication. In C. Cavé, I. Guaïtella, & S. Santi (Eds.), *Oralite´ et gestualite´: Interactions et comportements multimodaux dans la communication* (pp. 597–600). Paris: L'Harmattan.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America*, *113*, 1001–1024.

Brennan, S.E. (1995). Centering attention in discourse. *Language and Cognitive Processes*, *102*, 137–167.

Brennan, S.E., & Schober, M.E. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, *44*, 274–296.

Chafe, W.L. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C.N. Li (Ed.), *Subject and topic* (pp. 25–56). New York: Academic Press.

Chafe, W.L. (1994). *Discourse, consciousness, and time*. Chicago: Chicago University Press.

Clark, H.H., & Fox Tree, J.E. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition*, *84*, 73–111.

Clark, H.H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, *37*, 201–242.

Dahan, D., Magnuson, J.S., & Tanenhaus, M.K. (2001). Time course of frequency effects in spoken word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*, 317–367.

Dahan, D., Tanenhaus, M.K., & Chambers, C.G. (2002). Accent and reference resolution in spoken language comprehension. *Journal of Memory and Language*, *47*, 292–314.

Ferreira, F., & Bailey, K.G.D. (2004). Disfluencies and human language comprehension. *Trends in Cognitive Sciences*, *8*, 231–237.

Fox Tree, J.E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, *34*, 709–738.

Fox Tree, J.E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory & Cognition*, *29*, 320–326.

Fox Tree, J.E., & Clark, H.H. (1997). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, *62*, 151–167.

Gordon, P.C., Grosz, B.J., & Gilliom, L.A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, *17*, 311–347.

Haviland, S.E., & Clark, H.H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, *13*, 512–521.

MacDonald, M.C., Pearlmutter, N.J., & Seidenberg, M.S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *1014*, 676–703.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition*, *25*, 71–102.

Prince, E.F. (1992). The ZPG letter: Subjects, definiteness, and information-status. In. W.C. Mann & S.A. Thompson (Eds.), *Discourse description: Diverse linguistic analyses of a fund-raising text* (pp. 295–325). Amsterdam: John Benjamins.

Rossion, B., & Pourtois, G. (2001). Revisiting Snodgrass and Vanderwart's object database: Color and texture improve object recognition. *Journal of Vision*, *1*, Abstract 413. Retrieved April 19, 2003, from http://journalofvision.org/1/3/413, DOI 10.1167/1.3.413

Snodgrass, J.G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 174–215.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., & Sedivy, J.E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 632–634.

Tanenhaus, M.K., & Trueswell, J.C. (1995). Sentence comprehension. In. J. Miller & P. Eimas (Eds.), *The handbook of perception and cognition: Vol. 11* (pp. 217–262). San Diego, CA: Academic Press.

(Received 7/5/03; Revision accepted 1/30/04)