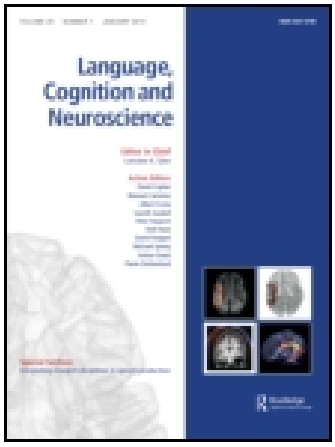

Language, Cognition and Neuroscience
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/plcp21

Do speakers articulate over-described modifiers differently from modifiers that are required by context? Implications for models of reference production

Paul E. Engelhardt[a] & Fernanda Ferreira[b]

[a] School of Psychology, University of East Anglia, Elizabeth Fry Building 1.10, Norwich Research Park, Norwich, UK

[b] Department of Psychology, University of South Carolina, Columbia, SC, USA

Published online: 07 Nov 2013.

Language, Cognition and Neuroscience, 2014
Vol. 29, No. 8, 975–985, http://dx.doi.org/10.1080/01690965.2013.853816

Routledge
Taylor & Francis Group

Do speakers articulate over-described modifiers differently from modifiers that are required by context? Implications for models of reference production

Paul E. Engelhardt[a]* and Fernanda Ferreira[b]

[a]*School of Psychology, University of East Anglia, Elizabeth Fry Building 1.10, Norwich Research Park, Norwich, UK;* [b]*Department of Psychology, University of South Carolina, Columbia, SC, USA*

(*Received 1 March 2012; accepted 3 October 2013*)

Studies have shown that speakers often include unnecessary modifiers when producing referential expressions, which is contrary to the Maxim of Quantity. In this study, we examined the production of referring expressions (e.g. *the red triangle*) that contained an over-described (or redundant) pre-nominal adjective modifier. These expressions were compared to similar expressions that were uttered in a context that made the modifier necessary for unique referent identification. Our hypothesis was that speakers articulate over-described modifiers differently from those used to distinguish contrasting objects. Results showed that over-described modifiers were significantly shorter in duration than modifiers used to distinguish two objects. Conclusions focus on how these acoustic differences can be modelled by Natural Language Generation algorithms, such as the Incremental Algorithm, in combination with probabilistic prosodic reduction.

Keywords: over-description; language production; Maxim of Quantity; Probabilistic Reduction Hypothesis; Incremental Algorithm

Reference occurs when a speaker produces a linguistic expression for a listener that uniquely identifies an object in the environment (Jackendoff, 2002). The exact form of the expression will necessarily depend on the context (i.e. the physical situation) in which the speaker and listener find themselves. For example, if the context contains multiple referents of the same type (e.g. two books that differ in colour), then the speaker must produce some type of modification (e.g. *the red book*) which will allow the listener to uniquely identify the intended referent. Therefore, the type of utterance that a speaker produces will be dependent upon the number and complexity of objects in the world. The Gricean Maxim of Quantity states that speakers should include enough information for an object to be identified, but no more (Grice, 1975, 1989). However, many studies have shown that adult speakers have a tendency to include extra modifiers, which we refer to as over-descriptions (e.g. Pechmann, 1989; Sonnenschein, 1984). Over-descriptions are referring expressions that include more modification than the context requires. Thus, if *the book* would be sufficient to identify the intended referent, then participants have a tendency to produce expressions such as *the red book* or *the book on the shelf*.

In an early study, Deutsch and Pechmann (1982) showed that speakers produced over-descriptions on approximately 25% of trials. In more recent work, Belke (2006) showed that when participants were placed under time pressure to begin speaking, they were even more likely to produce over-descriptions. In the time pressure condition, when a size modifier was required, almost all utterances had an unnecessary colour modifier, and when colour was required, approximately half had an unnecessary size modifier. In another study, Engelhardt, Bailey, and Ferreira (2006) reported that participants produced unnecessary prepositional phrase modifiers on 30% of trials in a single referent context. Thus, if the context contained only one apple, then participants were likely to produce an expression, such as *the apple on the towel*. Perhaps more interestingly, Engelhardt et al. reported data from an eye movement study which showed that listeners were almost 1 second slower to execute over-described instructions compared to instructions that were not over-described. This combination of findings begs the question of why speakers produce over-descriptions, if over-descriptions do in fact hinder comprehension performance (see also, Engelhardt, Demiral, & Ferreira, 2011; cf. Arts, Maes, Noordman, & Jansen, 2011).

On the surface, the combination of data suggests that speaker behaviour does not conform to the Maxim of Quantity (Grice, 1975), which predicts that speakers should provide enough, but no more information for an object to be identified. A related theory is the *Audience DesignHypothesis*, which assumes that speakers should construct utterances with the intention of being cooperative with interlocutors (Bell, 1984; Clark, Schreuder, & Buttrick, 1983; Clark & Wilkes-Gibbs, 1986; Levelt, 1989). Essentially, speakers should consider the needs of

*Corresponding author. Email: p.engelhardt@uea.ac.uk

the listener when formulating utterances. One way in which speakers can be cooperative with an interlocutor is to articulate clearly so as to be intelligible.

Studies have shown that less predictable words are articulated more clearly (e.g. Bard et al., 2000; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Klatt, 1975; Ladd, 1996; Lieberman, 1963), which has generally been interpreted as evidence that speakers can compensate for weak context by improving (or enhancing) articulation. This increases the chances of successful communication (Ladd, 1996; Lieberman, 1963; Samuel & Troicki, 1998). The primary factor that makes a word more intelligible is acoustic prominence, and in this case, prominence is defined in terms of duration, intensity and placement of pitch accents (Peirrehumbert & Hirschberg, 1990). Much of the discussion in this research area focuses on an inverse relationship between predictability and articulatory clarity/prominence. If a word is highly predictable, then speakers can afford to articulate that word in a reduced and less intelligible manner. Another theory that makes predictions about how speakers should formulate and articulate utterances is *Uniform Information Density* (Frank & Jaeger, 2008; Genzel & Charniak, 2002; Levy & Jaeger, 2007). It assumes that speakers will optimise information transmission in order to increase the chance of successful communication. The fundamental premise of uniform information density is that speakers will keep the information content of utterances as uniform as possible. Effects consistent with these predictions have been shown at multiple levels of analysis, including the production of individual words (Aylett & Turk, 2004; Bell et al., 2003) and the inclusion of syntactically optional words (Jaeger, 2010; Levy & Jaeger, 2007). Similar to the findings from the intelligibility literature, uniform information density is directly related to how predictable a word is in a particular context. Specifically, low predictability equals high information load, and high predictability equals low information load. Studies have shown that speakers will modulate word duration so that words with high information load are spread out over a longer period of time (Aylett & Turk, 2004; Bell et al., 2003).

The fact that speakers are capable of adjusting articulation might suggest one possible mechanism in which speakers can distinguish the informational importance of particular words in an utterance. We assume that speakers tend to include unnecessary modifiers because of egocentric production processes, and that if there are articulation differences between over-described and required modifiers, then the prosodic adjustment would occur later (e.g. during phonological encoding). These predictions follow from an incremental production architecture, such as the one described by Levelt (1999). The basic assumption is that an unnecessary modifier becomes part of the conceptual representation (or pre-verbal message) in the semantic/syntactic system, and articulation differences, if they occur, would likely be implemented later in the phonological/phonetic system. Thus, our key hypothesis is that over-described (or redundant) modifiers get included, but their acoustic realisation might suggest that they are less prominent and less communicatively important compared to modifiers that distinguish two objects. In the current study, we report a production experiment designed to investigate whether speakers articulate over-described modifiers differently from modifiers that are required by the context.

Models of reference production

Over the past decade or more, there has been increasing interest in computational models and algorithms that produce referential expressions (Dale & Reiter, 1995; Mellish et al., 2006; Reiter & Dale, 2000). The goal of these models is to generate a unique referring expression given a particular set of objects. Some models in this area attempt to generate expressions consistent with the Maxim of Quantity (Appelt, 1985; Dale, 1989), and others attempt to replicate human performance as closely as possible, even in those cases where human speakers produce less-than-ideal utterances (for reviews, see Krahmer & van Deemter, 2012; van Deemter, Gatt, van der Sluis, & Power, 2012; van der Sluis & Krahmer, 2007).

One of the most prominent, and arguably, most successful models in this area is called the Incremental Algorithm (Dale & Reiter, 1995; cf. van Deemter, Gatt, van Gompel, & Krahmer, 2012). In this model, potential object attributes are ordered based on preference (e.g. colour, size and orientation). The model begins with the most preferred attribute and checks whether the target object is uniquely identifiable. If the answer is no, then the model moves to the next attribute. It will terminate when a unique referring expression is generated. This model accurately predicts some over-descriptions because if an attribute, after being selected, is made redundant by an attribute selected later, it is still not excluded (in other words, the algorithm does not backtrack). The Incremental Algorithm cannot, however, predict cases in which a speaker produces an expression, such as *the red cup*, in a context in which there is only one cup (this is what we refer to as a one-referent context). It also cannot predict a case in which a less preferred attribute is included as an over-description (e.g. saying *the big red triangle*, when *the red triangle* uniquely identifies the object). However, several studies have shown that people are much more likely to include attributes such as colour than attributes, such as size (Belke, 2006; Kaland, Krahmer, & Swerts, 2011; Sedivy, 2007). To account for over-descriptions, the Incremental Algorithm must be non-deterministic (i.e. there must be a certain probability of including *red* in one-referent contexts).

Current study

The primary goal of this study was to investigate how speakers articulate over-described modifiers, and the secondary goal was to explore how articulation differences might be implemented in computational models of reference production. With regards to the secondary goal, we limit our discussion to models that produce over-descriptions (e.g. the Incremental Algorithm), and we envisage a model in which the acoustic properties of over-described modifiers are probabilistically reduced (Jurafsky, Bell, Gregory, & Raymond, 2000). In this study, participants saw arrays of objects presented in a 2 × 2 grid (see Figure 1), and they had to produce a referring expression for the object or set of objects in one quadrant, which was indicated by an arrow. In Panel A, there is only one triangle, and so a referring expression, such as *the triangle*, is sufficient for unique identification. In contrast, Panel B contains two hearts differing in colour, and so a modifier (e.g. *blue*) is required by the context. In order to control for both individual differences in speech rate and lexical variables such as length and frequency, we analysed only pairs of adjectives in which a participant produced the same modifier (e.g. *blue*) in both types of contexts. Therefore, in all critical comparisons, we analysed utterances that contained one adjective, and the primary manipulation was whether the context made the modifier an over-description (see panel A) or whether the context made the modifier necessary for unique reference (see panel B). We hypothesised that speakers might produce these two types of modifiers differently, and specifically, we expected the over-described (or redundant) modifiers to be produced with less acoustic prominence.

A final issue worth considering is the relationship between over-descriptions and predictability. Many previous studies of articulatory reduction have shown that predictability is one factor that will affect how people articulate particular words in a given context. If information is highly predictable, then it is redundant and people do not need to articulate as clearly. Likewise, if information is predictable, then it has low information content and it can be shortened to maintain uniform information density (Levy & Jaeger, 2007). We view over-descriptions as providing redundant information. However, we also believe redundancy and predictability are not always going to be linked, although they have been in many of the studies discussed thus far. Our assertions in some ways hinge on the rates at which speakers over-describe. If over-descriptions are relatively uncommon, then they will be unpredictable from both a comprehension and production stand point. Moreover, at this juncture, there are no data to suggest that predictability has anything to do with why a speaker chooses to include an over-described modifier. Therefore, in this study, we examine a novel situation in which an unpredictable, yet redundant word is a candidate for hypo-articulation.

In summary, the primary purpose of this experiment was to compare the acoustic properties of modifiers that were required by the context to those that were produced as over-descriptions. If speakers articulate these two types of modifiers in different ways, then it would suggest at least one possible explanation for why speakers might include unnecessary modifiers, despite the fact that they can be detrimental to comprehension. To preview the main findings, we found that over-described modifiers were shorter in duration than the modifiers that were contextually required. Our discussion and conclusions focus on how these behavioural findings might be implemented in referring-expression production models (e.g. Dale & Reiter, 1995) via probabilistic prosodic reduction (Jurafsky et al., 2000).

Method

Participants

Twenty-four native speakers of British English (age: M = 22.82, SD = 4.25, range: 18–35; male: 35.3%) with normal or corrected-to-normal vision were recruited to participate. Participants were recruited through the University of Edinburgh employment service, and each was paid £3.00.

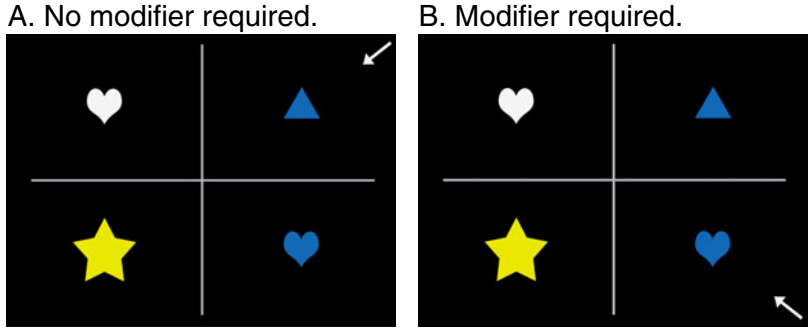

Figure 1. Example stimuli. Panel A shows a zero-modifier context and panel B shows a one-modifier context.

Materials

Stimulus materials consisted of 113 arrays of objects arranged within four quadrants. Three were for practice and 110 were the regular session items. One issue to consider when eliciting over-descriptions is overall task difficulty. In pilot work, we found that when task demands were low, participants tended not to produce over-descriptions, and when task demands were high, participants tended to over-specify a great deal. Task difficulty can be modulated in several different ways. One is to increase the number of contrasts in the display (see Figure 2). In panel A, there are three contrasting objects that vary in two attributes, and so two modifiers are required for unique identification (e.g. *the large white star*). In panel B, there are four sets of contrasting objects that vary in three attributes, and so three modifiers are required for unique identification (e.g. *the two small red squares*). We were primarily interested in one-modifier utterances when there were no contrasts in the display (e.g. Figure 1, panel A), and one-modifier utterances when the target was a member of a contrast set (e.g. Figure 1, panel B). We included two and three modifier trials as fillers to increase task difficulty, and at the same time, to provide variation with regards to the number of modifiers speakers had to produce from trial to trial. Because our analyses focused on trials in which the participant produced a single modifier, we wanted one-modifier utterances to be the modal response, and thus, across the entire experiment we also wanted to keep the mean number of required modifiers near one. To achieve this, we created 60 one-modifier trials, 20 zero-modifier trials, 22 two-modifier trials and 10 three-modifier trials. This makes the mean number of 'required' modifiers 1.18 over the entire experimental session.

The target object was rotated across the four regions of the display so that it appeared an approximately equal number of times in each region for each display type. The shapes and their attributes were also rotated for the target and the contrast objects. The irrelevant distractors objects were randomly assigned. However, the different shapes again occurred with different attributes, and the pairings were rotated across different attributes (i.e. colour, size and number). In 80 trials, there were (irrelevant) contrasting objects present in the display, and the target was not one of those. The rationale for including these contrasting (non-target) objects was to prevent participants from always focusing attention on the contrasting objects during the preview phase. For referentially 'required' modification in one-modifier trials, 20 required a colour modifier, 20 a number modifier and 20 a size modifier. Table 1 shows a list of all possible features and shapes.

Apparatus

Stimulus presentation was programmed using SR research Experiment Builder software. A 19" CRT monitor with a refresh rate of 140 Hz was interfaced with a 3-GHz Pentium 4 PC, which controlled stimulus presentation throughout the experiment. Participants responded orally into a microphone, and the software automatically recorded responses at 22.05 KHz and saved them in.wav format.

Design and procedure

As mentioned previously, to compare modifiers that were produced as over-descriptions to those that were required by the context, we examined only pairs of modifiers in which we could compare the same lexical item (e.g. *blue*) in both conditions for each participant. Thus, lexical variables and speech rate were controlled as much as possible, and the design consisted of a single variable with two levels, which we refer to as *modifier type*. One condition consisted of adjectives produced in contexts that made them *over-descriptions* and the other was adjectives produced in a context that made those adjectives *required* for object identification.

Prior to the experiment, participants were shown all of the shapes and colours that they would see during the experiment. Participants were instructed that they would see an array of objects and after a short time (i.e. 1.5 seconds) an arrow would appear in one of the quadrants. This was the participants' clue as to which object (or set of objects) they should describe. They were instructed to

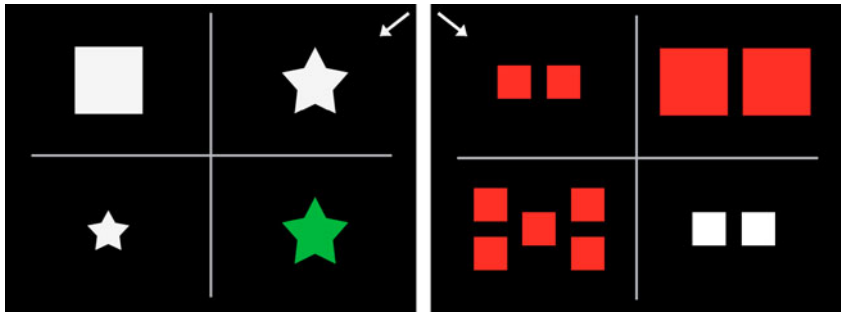

Figure 2. Example stimuli for displays requiring multiple modifiers. Panel A requires two modifiers and panel B three modifiers.

Table 1. Attributes manipulated in the experimental stimuli.

Numbers	Size	Colours	Shapes
One	Big	Blue	Circle
Two	Large	Grey	Diamond
Three	Little	Green	Heart
Four	small	Orange	Moon
Five		Pink	Square
Six		Purple	Star
		Red	Triangle
		White	
		Yellow	

Note: Only two sizes of objects were used.

formulate their utterance so that if someone else heard it they would be able to uniquely identify the object in the array. No specific instructions were given concerning over-descriptions; that is, participants were not told to avoid or to produce them. Participants were however, instructed not to use directional terms (i.e. left/right or top/bottom) in their referring expressions, and participants invariably complied. If the participant under-described during the practice, the experimenter explained how the utterance did not provide unique reference (i.e. a hypothetical listener could not select the intended object). The order of trials was randomly assigned for each participant, and participants pressed the space bar on the keyboard after they had finished speaking. Testing was conducted in a sound-treated room, and the entire testing session lasted approximately 20 minutes.

Analysis procedures

We conducted two sets of analyses. In both, we utilised linear mixed-effects models using the lme4 package in R, and p values were estimated using Markov Chain Monte Carlo sampling (Baayen, 2008; Bates, Maechler, & Dai, 2008). In all comparisons, we included both subjects and items as crossed-random factors. In the first set of analyses, we compared the modifiers and nouns from the two conditions of interest on measures of duration, pitch and intensity. In the second set, we looked at a number of additional variables, which we refer to as covariates, to rule out several alternative explanations of the findings.

Trials were excluded if a disfluency was uttered immediately before, during or after the modifier. This resulted in the exclusion of four trials. Utterances were segmented using Praat (Boersma & Weenick, 2010). Marking of word onsets and offsets in each critical utterance was done manually by two raters, who used consistent and standardised segmentation criteria (Stevens, 2002; Turk, Nakai, & Sugahara, 2006). Raters relied on a combination of auditory perception and spectral analysis in Praat. The first rater was an undergraduate research assistant who was naïve with respect to the experimental hypotheses, and the second was the first author of the study. After both raters independently analysed the data, the values for each trial were compared, and large differences (i.e. > 10 ms) were revaluated.

Results

There were 2640 trials in total. Five hundred seventy-one contained at least one unnecessary modifier, 67 were under-described (or ambiguous), and 2002 had the 'correct' amount of modification. Table 2 shows the number and percentage of under- and over-descriptions across all trials. As can be seen in Table 2, there is an increasing tendency for participants to under-describe as the number of required modifiers increased, which indicates that, as expected, these trials were more difficult. For the zero- and one-modifier trials, there were approximately 30% of trials with an unnecessary or redundant modifier, which is within the range of previous studies (e.g. Deutsch & Pechmann, 1982; Engelhardt et al., 2006). Only 19 utterances (7.2% of the total) violated the canonical adjective order in English (i.e. number, size and colour), so participants generally produced well-formed grammatical utterances.

Over-described versus required modifiers

For the modifier type analysis, we obtained 83 pairs of modifiers (62 – colour, 19 – number and 2 – size), in which participants produced the same adjective in both conditions (i.e. over-described vs. required). For these utterances, we also analysed the head nouns in each utterance. The mean number of syllables for the over-described noun phrases was $M = 2.86$, SD = 0.23 and the mean number of syllables for the required noun phrases was $M = 2.82$, SD = 0.45. The difference was not statistically significant ($p > 0.10$). Prior to the inferential analysis, we examined the data for outliers. Data points that were greater than four standard deviations from the mean in each condition were replaced with the mean for that condition. This resulted in the replacement of only three data points (i.e. <2% of the data).

Table 2. Number of over- and under-described responses for each display type.

	Over-described	Under-described
Zero modifier	132 (27.5%)	0 (0.0%)
One modifier	447 (31.0%)	4 (0.3%)
Two modifier	1 (0.2%)	16 (3.3%)
Three modifier	0 (0.0%)	47 (19.6%)

Note: Percentages represent the percentage within each display type.

Results for duration, intensity and pitch are shown in Table 3. For pitch, we analysed the maximum F0, and the mean pitch for voiced segments of the stressed syllable in both adjectives and nouns. For intensity, we also looked at the maximum (db) and the mean (db). Results showed two significant differences. First, the over-described modifiers were significantly shorter than those that were used to distinguish two objects. There was no difference in the duration of nouns. However, the pattern with the nouns reversed, such that over-described adjectives occurred with longer nouns. The analysis of mean pitch revealed significant differences for nouns, but not for adjectives. Nouns paired with over-described modifiers had lower mean pitch. The analysis of intensity was not significant for either the adjectives or nouns.

Analyses of covariance

We considered three covariates in follow-up analyses. The purpose of these analyses was twofold. First, we wanted to ensure that the two significant effects that we reported above could not be explained by alternate variables, and as reported in Table 3, both effects were robust when all three covariates were controlled. Second, we wanted to investigate whether any of the covariates interacted with the modifier type variable. The first covariate was gender, which we examined because males typically speak within a lower pitch range than do females. Not surprisingly, gender produced a main (or marginal) effect on all four pitch measures (modifier-maximum F0: $t = 5.98$, $p_{MCMC} < 0.001$; modifier-mean F0: $t = 7.04$, $p_{MCMC} < 0.001$; noun-maximum F0: $t = 1.88$, $p_{MCMC} = 0.06$; noun-mean F0 $t = 3.12$, $p_{MCMC} < 0.01$). However, gender did not interact with modifier type on any dependent measure.

The second covariate was trial order. The primary concern with this variable is based on the fact that speakers tend to acoustically reduce and shorten repeated words (Bard & Anderson, 1994; Bard & Aylett, 1999; Bard et al., 2000; Clark & Wasow, 1998; Metzing & Brennan, 2003). Therefore, one possibility for the length differences that we observed is that they occurred later in the experiment, after participants had already produced several instances of a particular word. Results showed that trial order interacted with modifier type only for noun duration $t = 2.45$, $p_{MCMC} < 0.05$. Nouns that occurred with required adjectives showed the expected negative relationship between trial order and duration ($r = 0.22$, $p = 0.05$). Nouns occurring with over-described adjectives did not show a negative relationship ($r = 0.07$, $p = 0.55$). Thus, there was some evidence that noun length was affected by trial order, but only in cases in which the noun phrase contained a required modifier. In the Discussion, we present possible explanations for this noun duration by trial order interaction. Trial order did not interact with any of the other dependent measures.

The final covariate was whether the display contained a competitor object that matched the property of the adjective. For example, if participants produced *the blue triangle* with a display like Figure 1, panel A, then we looked at whether the display contained another blue object. The reason for investigating this issue was because the presence of the blue heart might lead to more acoustic prominence on the noun to contrast the two blue objects (i.e. *the blue TRIANGLE* vs. *the blue HEART*). In critical trials, 46% had an object that matched the attribute of the modifier produced. When this factor was included into the linear mixed-effects models, it did not produce a main effect and it did not interact with modifier type on any of the acoustic measures.

Table 3. Means and standard deviations for duration (msec), pitch (Hz) and intensity (db) of the over-described and required modifiers and head nouns.

	Over-described		Required		*t* value
	M	SD	*M*	SD	
Modifiers					
Maximum F0 (Hz)	199.4	48.6	212.5	53.3	$t = 1.35$, $p_{MCMC} = 0.17$
Mean F0 (Hz)	175.7	40.1	177.9	40.3	$t = 0.84$, $p_{MCMC} = 0.40$
Intensity-max (db)	57.2	10.4	57.6	9.6	$t = 0.96$, $p_{MCMC} = 0.34$
Intensity-mean (db)	53.0	10.4	53.2	9.7	$t = 0.49$, $p_{MCMC} = 0.63$
Duration (ms)	284.2	35.7	374.8	94.3	$t = 3.65$, $p_{MCMC} = 0.0004^{a,b,c}$
Nouns					
Maximum F0 (Hz)	235.2	75.8	249.0	77.4	$t = 0.51$, $p_{MCMC} = 0.89$
Mean F0 (Hz)	162.7	36.1	175.1	50.4	$t = 2.84$, $p_{MCMC} = 0.005^{a,b,c}$
Intensity-max (db)	54.9	10.3	54.9	9.9	$t = 0.09$, $p_{MCMC} = 0.93$
Intensity-mean (db)	49.7	10.0	49.9	9.5	$t = 0.53$, $p_{MCMC} = 0.60$
Duration (ms)	533.7	78.7	513.4	111.0	$t = 0.47$, $p_{MCMC} = 0.64$

Note: Superscripts indicate that the effect remains significant when including a covariate (a = gender, b = trial order, c = modifier match).

Summary

The purpose of this experiment was to determine whether speakers articulate over-described modifiers differently compared to modifiers that are required by the context. We found that over-described modifiers were significantly shorter in duration. There were also differences in the mean pitch of the head nouns. Both effects remained significant when contributions of gender, trial order and modifier match were controlled.

Discussion

The primary purpose of this study was to investigate the production of over-described modifiers, and the secondary purpose was to link the behavioural findings to computational models of reference production. Previous studies have shown that speakers tend to include extra modifiers, and other studies have shown that extra information can be detrimental to comprehension (e.g. Engelhardt et al., 2011). We hypothesised that speakers might articulate over-described modifiers differently from those that are required by the context. If they did, then it might suggest one explanation for the puzzling discrepancy between production and comprehension, and speaker's lack of adherence to the Audience Design Hypothesis (Bell, 1984; Bard et al., 2000). Our results showed three main findings. The first was that over-described adjectives were significantly shorter than adjectives that were required by the context, which is consistent with our main (behavioural) prediction. The second is that nouns that were produced with over-described modifiers had lower mean pitch. The third was an interaction between trial order and noun duration, which was only significant for nouns paired with required modifiers. In the remainder of the discussion, we first consider the significance of these findings for psychological theories of reference production, and for over-descriptions more generally. In the second section, we explore the implications of our findings for models of reference production (e.g. Dale & Reiter, 1995).

Examining the results in Table 3 reveals that the pitch, intensity and duration measures were consistently lower for the over-described modifiers. The pattern was not quite as clear for the nouns, and in fact, the duration of nouns paired with over-described adjectives was actually longer than nouns paired with required adjectives. Therefore, we believe that our results are not consistent with prosodic reduction of the entire noun phrase, although others may debate this conclusion. In previous work, lower pitch, lower intensity and shorter duration are all typically linked with one another, and so, our mixed results with respect to noun duration and noun intensity leads us to believe that the adjective is more affected by acoustic reduction rather than the entire noun phrase (Aylett & Turk, 2004, 2006; Bell et al., 2003, 2009; Jurafsky et al., 2000). Relatedly, there was also evidence to suggest that the nouns paired with required modifiers also underwent reduction over the course of the experiment. Recall that we observed a significant interaction between trial order and noun duration, but only with nouns paired with required modifiers. This interaction cannot simply be due to repetitions (i.e. repeated words tend to be shortened and articulated less clearly) because the reduction only affected the required modifier condition. We believe that this interaction is likely due to Uniform Information Density, which assumes that more informative words will be spread out so that the informational profile of an utterance is as uniform as possible (Levy & Jaeger, 2007). If a noun phrase contains a required modifier, then the modifier is highly informative with respect to unique reference assignment. In this case, the informational content of the noun is comparatively less than the informational content of the modifier and the noun can be reduced, although again, this speculation is debatable.

The novel and important finding from the current study concerns the acoustic reduction of a redundant, but unpredictable word. Thus, our results suggest that acoustic reduction is associated with more than just word predictability in context. There are three theories that make related predictions concerning predictability/redundancy and the duration of words in speech. The Smooth Signal Redundancy Hypothesis (Aylett & Turk, 2006) predicts an inverse relationship between redundancy and duration so that information can be evenly spread out across the speech signal. The Probabilistic Reduction Hypothesis (Jurafsky et al., 2000) predicts that word forms will be reduced when they are more probable. Reduction refers to shorter overall duration, reduced vowels and final segment deletion. Uniform Information Density (Levy & Jaeger, 2007) assumes that words with high information load will be 'spread out' (i.e. will have a longer duration) in order to maintain a uniform information profile over the entire utterance. All three theories make articulatory predictions based on how probable or how predicable a word is in a given context.

In the current study, approximately 30% of trials in the zero- and one-modifier conditions contained an unnecessary or redundant adjective modifier. Thus, within our experimental context, over-descriptions were infrequent compared to modifiers produced to distinguish contrasting objects, and yet, we observed effects that were consistent with prosodic reduction of redundant information (i.e. the over-described modifier is low in information content). One counter argument to this explanation might be that across the entire experiment, most trials did require an adjectival modifier, which makes adjectives more predictable in the experiment than in everyday situations. However, this explanation does not account for the effects of context that we observed. In particular, on this view, it is not clear why over-described (or redundant) modifiers

were shorter than contextually required modifiers. In summary, our results are largely consistent with most studies of prosodic/articulatory reduction, and the key novelty of our reduction effect is that it was found for words that are not highly probable or predictable based on the overall rates of over-description.

One issue that our data do not address is whether the acoustic differences between over-described and required modifiers are due to speakers' deliberate attempts to formulate helpful utterances or whether the differences are an automatic outcome of language production and outside speakers' direct control. Our production experiment was conducted without a listener co-present, and so the finding of significant durational differences suggests that the effect is not fully attributable to intentional audience design (Bell, 1984; Clark et al., 1983; Clark & Wilkes-Gibbs, 1986; Horton & Keysar, 1996). Future work will need to investigate whether the effect is exaggerated with an interlocutor present, which would suggest that audience design affects the magnitude of the articulation differences that we observed. We return to this issue below.

Implications for models of reference production

The secondary goal of this study was to explore how the observed articulation differences might be implemented in algorithms designed to capture the generation of referring expressions (Krahmer & van Deemter, 2012). As a starting point, models need to produce over-descriptions, and coming from a psycholinguistic background, we are most interested in models designed to replicate human performance (Ferreira, 2007). Two prominent models are the Incremental Algorithm (Dale & Reiter, 1995) and the Greedy Algorithm (Dale, 1989). The former chooses attribute properties based on a fixed preference order, and does not exclude attributes that turn out to be non-distinguishing. The latter chooses attributes based on the number of distractor objects that it excludes. Because there can be an equal number of distractors that are ruled out by different attributes, the Greedy Algorithm has the additional benefit of being non-deterministic: If size and colour both rule out the same number of distractors, then the algorithm will probabilistically choose one or the other. Previous behavioural work suggests that colour should be chosen more frequently compared to size (Sedivy, 2006, 2007).

In the context of the current results, we propose that referring expression generation algorithms should be implemented in such a way that the over-described or redundant modifiers are probabilistically reduced, with reduction primarily affecting duration. The notion of automatic speech generation systems incorporating naturalistic prosody is not entirely novel (Hirschberg, 2002). However, the current study allows for quantitative predictions regarding a particular type of redundant information (i.e. when speakers include unnecessary modifiers in referring expressions). Specifically, over-described modifiers should on average be three-quarters of the length of required modifiers. However, we should mention that there are two ways to view these duration comparisons. Do people lengthen the modifier because it is required or do they shorten a modifier because it is redundant? In our view, the inclusion of a required modifier should be considered the baseline, which follows the assumptions of the Maxim of Quantity (Grice, 1975). Therefore, we have chosen to discuss our findings in terms of the over-described/redundant modifiers being shortened. At the same time, we must acknowledge that the durational distributions do overlap, and therefore, the reduction mechanism, if it intends to model human behaviour must be probabilistic (Jurafsky et al., 2000).

The current results also provide insights concerning different types of modifiers (i.e. different attribute properties). Sedivy, Tanenhaus, Chambers, and Carlson (1999) and Sedivy (2003, 2006, 2007) found that size and colour modifiers show distinct patterns in both comprehension and production. Speakers are more likely to produce unnecessary colour modifiers compared to size modifiers (see also Belke, 2006; Kaland et al., 2011). The most common account of this asymmetry is that colour is often an inherent property of an object, whereas size is relative or context dependent (Bache, 1978; Ferris, 1993; Kamp, 1975; Siegel, 1980). Our data are consistent with previous production findings: we observed very few over-described modifiers involving size. To our knowledge, this study is the first to vary sets of objects in a referential production task (i.e. the attribute of number). Speakers produced approximately one-third the number of over-descriptions involving the number attribute compared to the colour attribute. The absence of size over-descriptions is consistent with the preferred attribute order assumption of the Incremental Algorithm (Dale & Reiter, 1995). Therefore, in terms of colour and size, human performance supports assumptions of the Incremental Algorithm. Number is an inherent property of a set of objects, which makes it more similar to colour (Engelhardt, Xiang, & Ferreira, 2008). Likewise, Sedivy (2006) tested material modifiers such as *plastic* and *wooden*. To our knowledge, the attributes of number and material have not been discussed in relation to attribute preferences or where they might fall in the preferred attribute list that remains a topic for future research.

There are two issues or caveats that should be considered with respect to these findings. The first is that laboratory-based production tasks are not the same as naturalistic conversation. In our study, there was no interlocutor, speakers did not have to alternate between comprehension and production, and speakers produced similar utterances on a trial-by-trial basis. Therefore, the

results may not map perfectly onto what might be observed in more naturalistic situations. We feel that this study lays the ground work for future research that can investigate whether the articulatory differences are greater in interactive and more naturalistic dialogue situations. We also think that a great deal of work is needed to fully understand the underlying causes for why speakers include extra information. Most work to date has focused on visual or discourse saliency (Fukumura, van Gompel, & Pickering, 2010; Koolen, Goudbeek, & Krahmer, 2013). However, there may be other reasons why speakers produce over-descriptions. These factors will be important for computational models of reference production, which assume non-deterministic production architectures.

Conclusions

The primary goal of this study was to investigate whether speakers produce over-described modifiers differently from those that are required by the context. We know from previous work that speakers often produce unnecessary (or redundant) modifiers, at a rate estimated to range from 25% to 60% of trials. Most commonly, colour is included as an over-description. On the comprehension side, there is evidence that over-description impairs comprehension performance (e.g. Engelhardt et al., 2011; cf. Arts et al., 2011). The emerging picture from the comprehension side suggests that over-descriptions hinder performance (1) when visual contexts are relatively simple and (2) when the (auditory) referring expression is co-present with the visual context. In the current study, we found that speakers articulate contextually required modifiers differently from over-descriptions, and specifically, over-described modifiers are shorter than required modifiers. Thus, articulatory differences might provide one avenue for resolving the puzzling asymmetry between comprehension and production (i.e. participants tend to produce over-descriptions even though over-descriptions tend to mislead listeners). With respect to the secondary goal of incorporating the articulation findings into models of reference production, our main insights focus on a probabilistic reduction of over-described adjectives, and the data suggest that the acoustic realisation of an over-described modifier should be on average only three-quarters as long as required modifiers.

Acknowledgements

The authors would like to thank Yelda Semizer and Laura Speed for help creating the stimuli and collecting the data. We would also like to thank Sasha Calhoun for helpful comments on the data analysis procedures. This work was supported by ESRC grant RES-062-23-0475 awarded to Fernanda Ferreira.

References

Appelt, D. E. (1985). Planning English referring expressions. *Artificial Intelligence*, *26*(1), 1–33. doi:10.1016/0004-3702(85)90011-6

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*, 361–374. doi:10.1016/j.pragma.2010.07.013

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language andSpeech*, *47*, 31–56. doi:10.1177/00238309040470010201

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, *119*, 3048–3058. doi:10.1121/1.2188331

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics* Cambridge, UK: Cambridge University Press.

Bache, C. (1978). *The order of premodifying adjectives in present-day English*. Odense: Odense University Press.

Bard, E. G., & Anderson, A. H. (1994). The unintelligibility of speech to children: Effects of referent availability. *Journal of Child Language*, *21*, 623–648. doi:10.1017/S030500090000948X

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, *42*, 1–22. doi:10.1006/jmla.1999.2667

Bard, E. G., & Aylett, M. P. (1999). The dissociation of deaccenting, givenness, and syntactic role in spontaneous speech. In Proceedings of ICPhS-99, San Francisco, CA, USA.

Bates, D., Maechler, M., & Dai, B. (2008). Lme4: Linear mixed-effects models using S4 classes [Computer software manual]. Retrieved from http://lme4.r-forge.r-project.org/

Belke, E. (2006). Visual determinants of preferred adjective order. *Visual Cognition*, *14*, 261–294. doi:10.1080/13506280500260484

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*, 145–204. doi:10.1017/S004740450001037X

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*, 92–111. doi:10.1017/S004740450001037X

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. *The Journal of the Acoustical Society of America*, *113*, 1001–1024. doi:10.1121/1.1534836

Boersma, P., & Weenink, D. (2010). Praat: doing phonetics by computer [Computer program]. Retrieved 2010 from http://www.praat.org/.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, *22*, 245–258. doi:10.1016/S0022-5371(83)90189-5

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, *37*, 201–242. doi:10.1006/cogp.1998.0693

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1–39. doi:10.1016/0010-0277(86)90010-7

Dale, R. (1989). Cooking up referring expressions. In J. Hirschberg (Ed.), *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics* (pp. 68–75). Stroudsburg, PA: Association for Computational Linguistics.

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*, 233–263. doi:10.1207/s15516709cog1902_3

Deutsch, W., & Pechmann, T. (1982). Social interaction and the development of definite descriptions. *Cognition*, *11*, 159–184. doi:10.1016/0010-0277(82)90024-5

Engelhardt, P. E., Bailey, K. G. D., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean maxim of quantity. *Journal of Memory and Language*, *54*, 554–573. doi:10.1016/j.jml.2005.12.009

Engelhardt, P. E., Demiral, S. B., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, *77*, 304–314. doi:10.1016/j.bandc.2011.07.004

Engelhardt, P. E., Xiang, M., & Ferreira, F. (2008). Anticipatory eye movements mediated by word order constraints. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual conference of the Cognitive Science Society* (pp. 951–957). Austin, TX: Cognitive Science Society.

Ferreira, F. (2007). Prosody and performance in language production. *Language and Cognitive Processes*, *22*, 1151–1177. doi:10.1080/01690960701461293

Ferris, C. (1993). *The meaning of syntax: A study of adjectives in English*. London & New York, NY: Longman.

Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 939–944). Austin, TX: Cognitive Science Society.

Fukumura, K., van Gompel, R. P. G., & Pickering, M. J. (2010). The use of visual context during the production of referring expressions. *Quarterly Journal of Experimental Psychology*, *63*, 1700–1715.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of ACL-2002* (pp. 199–206), Philadelphia, PA: Association for Computational Linguistics.

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics: Speech acts* (Vol. III, pp. 41–58). New York: Academic Press.

Grice, P. (1989). *Studies in the ways of words*. Cambridge, MA: Harvard University Press.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*, 91–117. doi:10.1016/0010-0277(96)81418-1

Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication*, *36*, 31–43. doi:10.1016/S0167-6393(01)00024-3

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford: Oxford University Press.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*, 23–62. doi:10.1016/j.cogpsych.2010.02.002

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2000). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 229–254). Amsterdam: John Benjamins.

Kaland, C., Krahmer, E., & Swerts, M. (2011). Salient in the mind, salient in prosody. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 261–266). Austin, TX: Cognitive Science Society.

Kamp, H. (1975). Two theories about adjectives. In E. L. Keenan (Ed.), *Formal semantics for natural languages* (pp. 123–155). Cambridge: Cambridge University Press.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, *3*, 129–140.

Koolen, R., Goudbeek, M., & Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, *37*, 395–411. doi:10.1111/cogs.12019

Krahmer, E., & van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*, 173–218. doi:10.1162/COLI_a_00088

Ladd, D. R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT.

Levelt, W. J. M. (1999). Producing spoken language: A blueprint for the speaker. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 83–122). Oxford: Oxford University Press.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt and T. Hoffman (Eds), *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems* (pp. 1–8).

Lieberman, P. (1963). Some effects of the semantic and grammatical context on the production and perception of speech. *Language and Speech*, *6*, 172–175.

Mellish, C., Scott, D., Cahill, L., Paiva, D., Evans, R., & Reape, M. (2006). A reference architecture for natural language generation systems. *Natural Language Engineering*, *12*, 1–34. doi:10.1017/S1351324906004104

Metzing, C. A., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*, 201–213. doi:10.1016/S0749-596X(03)00028-7

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*, 89–110. doi:10.1515/ling.1989.27.1.89

Peirrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, & M. Pollack (Eds.), *Intentions in communication* (pp. 271–311). Cambridge, MA: MIT Press.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511519857

Samuel, A. G., & Troicki, M. (1998). Articulation quality is inversely related to redundancy when children or adults have verbal control. *Journal of Memory and Language*, *39*, 175–194. doi:10.1006/jmla.1998.2580

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. Journal of Psycholinguistic Research, *32*, 3–23. doi:10.1023/A:1021928914454

Sedivy, J. C. (2006). Evaluating explanations for referential context effects: Evidence for Gricean mechanisms in online language interpretation. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to world-situated language use: Bridging the language-as-product and language-as-action traditions (learning, development, and conceptual change)* (pp. 153–171). Cambridge, MA: MIT Press.

Sedivy, J. (2007). Implicature during real time conversation: A view from language processing research. *Philosophy Compass*, *2*, 475–496. doi:10.1111/j.1747-9991.2007.00082.x

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147. doi:10.1016/S0010-0277(99)00025-6

Siegel, M. (1980). *Capturing the adjective*. New York, NY: Garland.

Sonnenschein, S. (1984). The effects of redundant communications on listeners: Why different types may have different effects. *Journal of Psycholinguistic Research*, *13*, 147–166. doi:10.1007/BF01067697

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, *111*, 1872–1891. doi:10.1121/1.1458026

Turk, A., Nakai, S., & Sugahara, M. (2006). Acoustic segment durations in prosodic research: A practical guide. In S. Sudhoff, D. Lenertov, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, & J. Schlieer (Eds.), *Methods in empirical prosody research*. Berlin and New York, NY: De Gruyter.

Van der Sluis, I., & Krahmer, E. (2007). Generating multimodal referring expressions. *Discourse processes*, *44*, 145–174. doi:10.1080/01638530701600755

van Deemter K., Gatt A., Sluis I. V. D., & Power R. (2012). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science 36*, 799–836.

van Deemter, K., Gatt, A., van Gompel, R. P. G., & Krahmer, E. (2012). Towards a computational psycholinguistics of reference production. *Topics in Cognitive Science*, *4*, 166–183. doi:10.1111/j.1756-8765.2012.01187.x