On the durational reduction of repeated mentions: recency and speaker effects

Viktor Trón

International Graduate College of Language Technology and Cognitive Systems
Edinbugh & Saarbrücken
v.tron@ed.ac.uk

Abstract

There are conflicting views in the literature as to the role of listener-adaptive processes in language production in general and articulatory reduction in particular. We present two novel pieces of corpus evidence that corroborate the hypothesis that non-lexical variation of durations is related to the speed of retrieval of stored motor code chunks and durational reduction is the result of facilitatory priming.
keywords: speech production, corpus study, reduction, duration, priming

1. Introduction

Over the past decade, it has been demonstrated that discourse context systematically affects articulatory variation in speech over and above lexical, social and individual factors. It has long been noticed that frequent words are hypoarticulated relative to infrequent words of the same phonological composition (0; 0). In a series of work (0), Jurafsky and colleagues showed that predictability within the discourse context is also a key determinant of articulatory variation. In particular they found a unidirectional link between redundancy and reduction, which they destilled into the mnemonic: "inform less, less form in" and formulated the *Probabilistic Reduction Hypothesis* (PRH): More probable words are more reduced.

The PRH has the potential to subsume various static as well as dynamic sources of redundancy which are known to enhance reduction such as frequency, bigram probability, semantic association and repetition. However, it leaves open the question how one can pin down a causal connection between information-theoretic notions of probability or redundance on the one hand and reduced articulation on the other.

Listener-adaptive accounts.

The earliest known statements about the relation between redundancy and reduction framed the phenomenon in functional terms somewhat similar in vein to Jurafsky et al's mnemonic. In this view articulatory reduction is the implementation of the Principle of Least Effort and the PRH is a constraint on its application: its interaction with the Principle of Clarity. According to this, speakers can afford attenuating their pronunciation more in contexts where more information is available for the listener to identify what is said (0; 0; 0). Taking this for granted implies that articulatory reduction is a listener-adaptive process which presupposes that the speaker has to maintain and update a model of the listener.

Explicit listener-modelling has been shown to be very limited in natural spontaneous discourse and thought to be restricted to monitoring and adjustment stages of language use. Even at higher levels of linguistic processing, the way people plan referential expressions is known to be predominantly speaker-centered (0) and inferencing from common ground is subject to available resources. This leads one to question that computationally costly listener-modelling could underlie such low-level processes as articulation (0).

Priming and reduction.

Balota *et al* (0) perform a series of experiments, where they prompt speakers to produce a target word in the context of semantically related and unrelated primes. They show that semantic relatedness leads to shorter target productions, and this durational reduction is more pronounced at shorter latencies between prime and target. To our knowledge this study is the first to hypothesize that durational reduction might be directly related to memory retrieval insomuch as chunks of motor codes are sequentially accessed during production, and access facilitated by priming can lead to speeded execution, i.e., shortening.

Why duration?

Various reductive processes can be viewed as resulting from attenuated gestures due to temporal overlap and therefore regarded as side-effects of durational reduction. Also, the durational aspect of reduction is a fortunate choice for experiments since it can direcly be measured on speech output and requires only minimally tagged speech corpora.

Why repetitions?

Absolute measures of durational reduction are problematic, since norms are difficult to obtain due to the large amount of variables influencing durations (speech rate, style, prosody, segmental composition, etc.). If word durations are compared across mentions of the same word in the same discourse, most of these methodological problems are solved. Therefore we decided to explore *durational reduction of repeated mentions* in the hope that it tells us about the relationship between contextual redundancy and articulatory reduction in general described in the PRH. Fowler and Housum (0) were the first to demonstrate that second mentions of words in a discourse are shorter than first mentions and proposed a listener-adaptive functionalist account in line with the PRH.

Synopsis.

This paper presents novel results on durational reduction of repetitions that corroborate a non-functionalist mechanistic account of reduction mediated by priming. If durational

reduction is the result of priming, we expect repetitional reduction to show a recency effect similar to the one found in semantic priming by Balota *et al*. In particular, *we hypothesize that consecutive mentions of a word show more reduction for shorter latencies and asymptotically level out as the time lapse between the mentions increases*. Since repetition is the strongest possible prime–target association, we expect that the sensitive time window showing reduction is much longer than for semantic priming and also that its magnitude is larger.

Section 2. describes the corpus used in our study, overviews dataset preparation and introduces our terminology. In section 3. we test the recency effect for repetitional reduction. Section 4. contrasts reduction in self- and cross-speaker repetitions in dialogue to provide further argument against the listener-adaptive account of repetitional shortening.

2. Materials and method

We present results on the Edinburgh Maptask Corpus (Maptask), a collection of spontaneous dialogue transcribed and aligned on the word level. The Maptask Corpus contains 128 dialogues totalling a 14.5 hours of speech. The corpus was cleaned of fillers, pauses, fragmentary utterances and overlapping speech in dialogue. Mentions of the same word type w (based on orthographic identity) within a dialogue define a *mention chain*, $w_1, w_2, \ldots, w_n$. Datapoints in our initial dataset are *repetitions*, i.e., pairs of consecutive mentions in mention chains, $\langle w_i, w_{i+1} \rangle$. Repetitions are indexed for their position in the mention chain, e.g., $\langle w_i, w_{i+1} \rangle$ has *position index* i. We extracted start and end times and word durations. In addition to this, for each mention pair we record durational reduction and latency. *Durational reduction* is measured directly as the duration difference between the later and earlier mention of the pair in milliseconds; this gives an intuitive scale where a value for repetitional shortening is smaller than for lengthening. The *latency* of a repetition is defined as the time lapse between the end of the earlier mention to the onset of the later mention in seconds. More precisely, latency for the pair $\langle w_i, w_{i+1} \rangle$ is $\mathrm{end}(w_i) - \mathrm{start}(w_{i+1})$ (with range $(-\infty, 0]$), which yields a mnemonic measure where repetitions with recent earlier mentions are to the right while ones with long lapsed earlier mentions are to the left. We also recorded token frequency of the word type based on occurrences within the corpus. The resulting database contains some 100,000 mention pairs with 1,000 word types. Mentions were also tagged for speaker, which allowed us to classify repetitions as *self-repetitions* if the utterers of the two consecutive mentions were identical or *cross-speaker repetitions* in case they were different. Approximately two-thirds of all mention pairs are self-repetitions.

3. Latency

Fowler and Housum's earlier finding that second mentions are shorter was confirmed. Paired t-test directly comparing durations of earlier and later mentions was highly significant (see Table 1). An alternative hypothesis would be that repeated mentions are reduced because they start later in the discourse. A control test was performed where a later mention was paired with a first mention of the same word in another dialogue with a matching onset (it starts at the same time in their respective dialogue). Paired t-test showed that repeated mentions are significantly shorter than onset-matched first mentions ruling out the possibility that durational difference between consecutive mentions is solely an artefact of different amounts of preceding discourse.

	reduction (ms)	t	df
overall	-3.01	-9.47	102603
1st pos.	-13.5	-14.7	14891
lat. $<$ 10s	-8.54	-19.9	48616
1st pos., $<$ 10s	-27.1	-17.5	4982
control	-2.69	-7.90	67754
1st. pos. control	-13.06	-12.40	11378

Table 1: Results of paired t-tests comparing the durations of consecutive mentions. All results are highly significant ($p \approx 0$).

The priming account of durational reduction predicts a recency effect of reduction: the magnitude of durational reduction is larger for short latencies between the consecutive mentions. Latency (-log scale) shows a highly significant negative correlation with reduction ($\varrho = -0.06$, $t = -19.5$, $df = 102602$, $p \approx 0$). Fig 1 demonstrates that the temporal relationship is monotonic and near linear for latency quantiles (which almost perfectly align with log latency). Reduction is attested even in the latency range of 50-60 seconds and asymptotically levels out at longer latencies.

In order to quantify the predictive power of reduction, a linear model was run with reduction as the response variable and various factors like latency (log), token frequency (log), onset and position index (log), speaker as independent variables. ANOVA shows the main effect of latency as highly significant ($F = 558.32$, $p \approx 0$). Frequent words tend to reduce less than infrequent words, and there is a significant interaction between frequency and latency ($F = 23.98$, $p \approx 0$) suggesting a floor effect: frequent words are typically short which cannot reduce as much.

The magnitude of reduction also decreases steadily with position index (see Fig 1). Note that for higher position indexes we actually see lengthening for longer latencies. The higher the position index, the shorter the expected average latency, the more likely the earlier mention will be reduced due to a short-latency prime before.

These results are compatible with the PRH since we expect predictability of a word to decrease as its previous mention is farther behind in the discourse. The temporal profile of reduction, however, is reminiscent of the laboratory findings of the Balota *et al* study and raises the possibility that we are dealing with a recency effect in repetitional priming. Since repetition serves as the most robust prime–target association possible, the sensitive range of latencies is longer than for semantic priming and the magnitude of reduction is larger.

Listener-adaptive views that anchor the notion of redundancy in PRH to information content in the message have a hard time to accommodate that redundancy of a word can decrease with later mentions to account for lengthening of

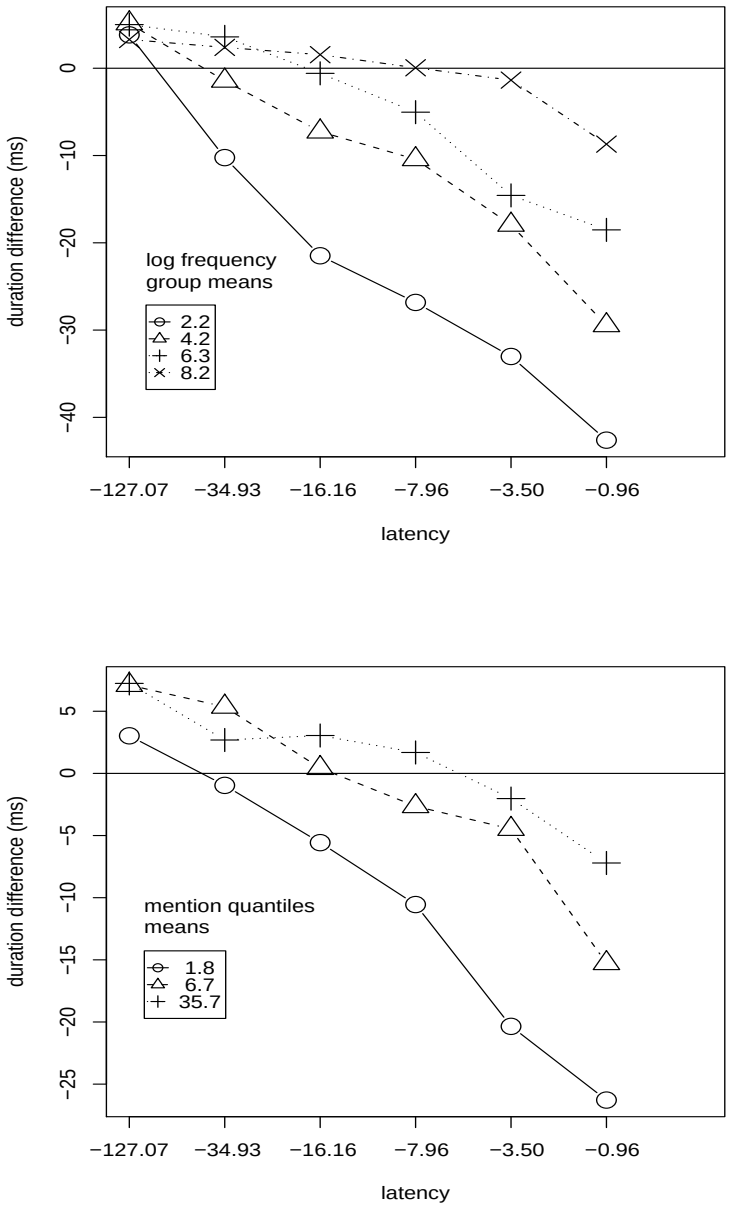

Figure 1: Reduction (more reduced is down) as a function of latency (more recent on the right). Duration difference betweeen consecutive mentions is averaged over 6 latency quantiles and plotted for 4 different frequency groups (top) and 3 mention quantiles (buttom) in the Maptask Corpus.

repetitions with high position index and long latency. A priming account, however, allows us to link such low-level temporal dependency of reduction to known recency effects on facilitation.

We found similar effects with word type: content words reduce more than function words for the same latencies and frequency, however word length is still a confound. This possibility is eliminated by using a normalized measures of reduction for the two corpora specifying alignment at the phoneme level which show exactly the same interactions.

4. Self-repetitions

In a second experiment we compared durational reduction of self-repetitions and cross-speaker repetitions. According to listener-adaptive accounts, durational reduction depends on the redundancy of a word given the speaker's model of the listener. Speakers cannot be sure that the listener processed a word, whereas listeners are expected to update their speaker's model upon successful comprehension of the message. This implies that the redundancy of a word is less influenced by an earlier mention when the speaker is also the utterer of the earlier mention than when she is the listener interpreting it. According to this self-repetitions

	reduction (ms)		paired t-test		
	self	cross	t	df	p
lat.<10s	-34.5	-19.4	3.27	1077	$< .001$
lat.>10s	-6.40	-5.44	0.29	4546	$> .5$
overall	-15.43	-9.94	2.06	3351	$< .05$

Table 2: Durational reduction in self- and cross-speaker repetitions.

are to be equally or less reduced than cross-speaker ones. More mechanistic views of dialogue assume a strict parity of representations which predicts identical activations on all levels in production and comprehension (0). This approach predicts no difference between self- and cross-speaker repetitions.

A priming account allows that comprehension and production drive different processes so representational parity does not imply quantitatively identical degrees of activation. We hypothesize that the actual execution of motor codes in production channels more activation to articulatory representations than comprehension processes do. More activation in production would predict larger facilitation of retrieval with self-repetitions and thus more durational reduction.

In order to compare these competing hypotheses, we compiled a dataset out of first–second mentions where self-repetitions and cross-speaker repetitions were paired up and were matched for word type and log latency. A paired t-test shows that self-repetitions are significantly more reduced than cross-speaker repetitions (see Table 2). Contrast in reduction between the speaker identity condition is more robust for short latencies, levelling out at around 10-15 seconds.

Overall, our results are different from a similar experiment by Bard *et al* (0) who found no difference in reduction between self- and cross-speaker repetitions. We conjecture that this may be because repetition latency was not controlled for in their study; the lack of difference may be an artefact of shorter cross-speaker latencies or predominantly long latencies where the contrast is neutralized.

More reduction of self-repetitions contradicts a purely listener-adaptive account. Moreover, it does not support mechanistic views of discourse which assume a strict parity in the use of representations in production and comprehension. In particular, it suggests that the motor theory of speech perception (0) needs to be refined at least allowing for longer lasting activation of articulatory representations during production than in perception.

5. Conclusion

This paper presented two novel results about the durational reduction of repeated mentions: (i) Latency between consecutive mentions is inversely proportional to the magnitude of repetitional shortening. (ii) Self-repetitions are more reduced than cross-speaker repetitions at short latencies. Both results falsify a purely listener-adaptive account of reduction but are compatible with a priming account: (i) is explained by a recency effect in repetitional priming while (ii) is explained by assuming higher degree (or longer decay) of activation of articulatory representations in production relative to perception.

In sum, the results corroborate the hypothesis that non-lexical aspects of durational variation are modulated by the speed of retrieval of motor code chunks in speech production. This proposal explains durational reduction as a result of facilitatory priming and provides the causal link between redundancy and reduction stated descriptively in the PRH.

Since the temporal aspects of articulation can even be taken under conscious control, we do not dispell the possibility that inferencing listener needs from common ground can prompt listener-adaptive choice in phonetic realization. However, we second the literature suggesting a limited, resource-dependent role for this computationally costly process of explicit other-modelling. As long as more mechanistic explanations of articulatory reduction are feasible, there is no need to invoke listener modelling at the phonetic end of speech production.

6. References

Balota, D. A., Boland, J. E., Shields, L. W. 1989. Priming in pronunciation: Beyond pattern recognition and onset latency. *Journal of Memory and Language* 28, 14–36.

Bard, E. G., Anderson, A., 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.

Bybee, J. L. 2001. *Phonology and Language Use* volume 94 of *Cambridge Studies in Linguistics*. Cambridge: Cambridge University Press.

Fowler, C. 1988. Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech* 28, 47–56.

Fowler, C., Housum, J. 1987. Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language* 26, 489–504.

Horton, W. S., Keysar, B. 1996. When do speakers take into account common ground. *Cognition* 59, 91–117.

Jurafsky, D., Bell, A., Gregory, M., Raymond, W. D. 2001. Probabilistic relations between words: evidence from reduction in lexical production. In: Bybee, J., Hopper, P., (eds), *Frequency and the Emergence of Linguistic Structure* number 45 in Typological Studies in Language. John Benjamins 229–253.

Liberman, A. M., Mattingly, I. G. 1985. The motor theory of speech perception revised. *Cognition* 21, 1–36.

Lindblom, B. 1990. Explaining variation: a sketch of the h and h theory. In: Hardcastle, W. J., Marchal, A., (eds), *Speech Production and Speech Modelling*. Dordrecht, Netherlands: Kluwer 403–439.

Pickering, M., Garrod, S. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27(02), 169–190.

Wright, C. 1979. Duration differences between rare and common words and their implications for the interpretation of word frequency effects. *Memory and Cognition* 7(6), 411–9.