Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

How do both phonological and syntactic complexity influence speech planning?

Ivan Yuen[1], *Katherine Demuth*[2], *Stefanie Shattuck-Hufnagel*[3]

[1]Department of Language Science and Technology, Universität des Saarlandes
[2]Department of Linguistics, Macquarie University
[3]Speech Communication Group, Research Laboratory of Electronics, MIT
ivyuen@lst.uni-saarland.de, Katherine.demuth@mq.edu.au, sshuf@mit.edu

Abstract

Wheeldon & Lahiri [1] reported that speech initiation time (RT) was influenced by the phonological complexity of the initial prosodic word (PWd) in an immediate-production task, but by the number of PWds in a delayed-production task that enabled preplanning. Thus, different tasks are sensitive to different types of complexity. The current study explored RT for different complexity types: number of syllables in the PWd (monosyllabic vs disyllabic), location of a complex PWd (early vs late), and number of PWds (3 vs. 4 vs. 5) using a reading aloud task. The presence of a disyllabic noun increased RT only in 4-PWd sentences, where the syntactic complexity of the SUBJECT and OBJECT NPs differed. This suggests that RT may increase in the absence of parallel syntactic structures in SUBJECT and OBJECT position, highlighting the need for further research on the role of complexity at different levels in the speech planning process.

Index Terms: prosodic word, speech planning, phonological complexity, syntactic complexity, parallel structures

1. Introduction

It is widely assumed that more complex spoken utterances require more extensive processing to plan and produce. If so, this suggests that defining the types of complexity involved might be useful for guiding controlled studies of how this multi-faceted factor influences the speech planning process. Previous work by Wheeldon and Lahiri [1] has shown that two measures of complexity, i.e., the number of words in the first PWd of an utterance, and the total number of PWds in the utterance, can have an effect on the latency (response time) to begin speaking an utterance. This has been shown in different tasks, where the complexity of the first PWd, manipulated by the presence vs absence of cliticization, affected RT in an immediate speaking task, while the total number of PWds in the utterance affected RT in a delayed speaking task. The RT differences were correlated with manipulation of the prosodic organization of the word string, even when the number of words and syllables remained constant. These results identify several factors involved in estimating the complexity of an utterance which can influence latency, including the structure of the constituents (e.g., whether the first PWd contains a clitic), the number of constituents (e.g., PWds) and the nature of the elicitation task (e.g., enabling pre-planning vs not). Yet manipulation of these factors can result in utterances that also differ in syntactic complexity. Thus, longer RT may arise from the complexity of encoding at several different levels of structure in an utterance. A better understanding of this issue is needed to shed light on how complexity at different linguistic levels can interact to influence RT during speech planning.

In this study we continue the exploration of complexity effects on speech production planning, focusing on 1) a different type of PWd complexity (i.e. di-syllabic nouns (as in [2], [3] and [4]), rather than verb+clitic structures (as in Wheeldon & Lahiri, henceforth W&L) and 2) the placement of the complex PWd in the utterance (as in [5] (early, in the SUBJECT, vs late, in the DIRECT OBJECT)), while also varying the number of PWds in the utterance. Although we did not have a specific hypothesis about the effect of placement of a disyllabic word on RT, the assumption that speech can be incrementally planned suggests that, if the number of syllables in the target words has an effect, a disyllabic word that occurs late in an utterance may be less likely to influence RT than one that occurs early in an utterance.

2. Method

Fifteen (2M, 13F) Australian English-speaking adults participated in the study, with a mean age of 26;8 years (Range = 19 – 59). They were recruited from the student pool at Macquarie University, Sydney, in exchange for course credit. They also provided written informed consent, in accordance with the approval from the Human Research Ethics Committee at Macquarie University (52020173714492).

Eighteen target sentences were created as stimuli, varying along three dimensions: the number of syllables (SYLs) in the lexical word (a dimension of complexity of the PWd); the total number of PWds in the utterance; and the position of the disyllabic word in the utterance (SUBJECT vs. OBJECT position). The number of SYLs was varied by replacing a monosyllabic noun with a disyllabic noun (e.g. queen vs soldier, nurse vs teacher). The position of the disyllabic noun was manipulated to occur early or late, i.e. in SUBJECT position vs. in OBJECT position, in Subject + Verb + Object (SVO) sentences (e.g. The soldier loves meat vs The nurse loves patients). The number of PWds was manipulated by adding an adjective (e.g. in subject position, as in The queen vs The smart queen; or in object position, as in drinks tea vs drinks green tea). For comparison, the control sentences for each set of 3, 4 or 5 PW sentences contained only monosyllabic words. Table 1 lists the target sentences in our experimental conditions.

10.21437/Interspeech.2025-236

Table 1: *List of target sentences. Disyllabic target words shown in italics*

	Monosyll Control	**Disyll SUBJECT (DS)**	**Disyl OBJECT (DO)**
3 PWds	The queen drinks tea The nurse loves cheese	The *soldier* loves meat The *teacher* drinks wine	The prince keeps *horses* The nurse loves *patients*
4 PWds	The smart queen drinks tea The strong nurse loves cheese	The strong *soldier* loves meat The smart *teacher* drinks milk	The queen keeps strong *horses* The guard likes fresh *flowers*
5 PWds	The sad queen drinks green tea The short nurse loves soft cheese	The dumb *soldier* loves fresh meat The cute *teacher* drinks cows' milk	The proud duke keeps fit *horses* The tall guard loves fresh *flowers*

The control and target sentences share the same SVO word order, to minimize any potential influence of word order differences. Note that, in the 3-PWd sentences, all NPs are syntactically simple (no adjective), whereas in the 5-PWd sentences, both the SUBJECT NP and the DIRECT OBJECT NP are syntactically complex (adj+N – as in strong horses, fresh flowers). In contrast, in the 4-PWd sentences, either the SUBJECT NP or the DIRECT OBJECT NP is syntactically complex, but not both. The implications of this aspect of the stimulus design will be discussed below.

The lexical frequency of different lexical words was controlled to be as comparable as possible across the sentence stimuli, using Subtlex [6]. Table 2 summarises the 'average word frequency per sentence' in our experimental conditions, as derived according to Total frequency of all lexical words in an utterance / Number of lexical words in an utterance. An additional 46 filler sentences were mixed with the 18 target sentences to yield a list of 64 sentence items. Each participant repeated the list three times in separate blocks. Each blocked presentation was randomized in DmDX [7] for individual participants.

Table 2: *Average word frequency grouped by the number of prosodic words (PWds) and disyllabic conditions*

	Monosyllabic Control	**Disyllabic SUBJECT (DS)**	**Disyllabic OBJECT (DO)**
3 PWds	2584	2595	2548
4 PWds	3106	3000	3007
5 PWds	2843	2561	2678

Participants sat at a table in a sound-treated studio facing a display monitor, with two external speakers on either side. A condenser microphone on a boom stand with a horizontal arm was placed about 30 cm away from the lip corner of participants' mouth, to record the participant's spoken responses. Each experimental trial began with a visual display of a black fixation cross against a white background in the centre of the display monitor for 500 milliseconds (ms). The fixation cross then disappeared and the display monitor stayed blank for variable intervals from 250 to 2250 ms. This delay interval was varied to prevent the speaker from anticipating an inter-trial interval and being entrained to a regular spoken rhythm, which might mask stimulus-related differences in RT. This was followed by a visual display of a green circle against a white background in the centre of the display screen for a period of 500 ms. After a subsequent interval of 550ms with a blank screen, a sentence was visually displayed in one line in the centre of the screen, concurrent with a 1KHz-beep being played through the external speakers for 250 ms. Participants were instructed to read aloud the visually displayed sentence as soon as possible after these auditory/visual 'Go' signals. Prior to testing, five practice sentences were used to familiarize participants with the procedure. All trials were presented using DmDX [7]. Participant responses were recorded onto a separate PC using Audacity, at a sampling rate of 44.1KHz. Response time was measured from the onset of the beep to the start of a sentence response, as determined by visual inspection of the onset of the article in a waveform and spectrogram display using Praat [8].

3. Results

A total of 784 recorded utterances were used in the analysis, after excluding incorrect productions (n = 10), disfluencies (n = 9), restarts (n = 2), skipped productions (n = 4) (total n = 25), and response time outliers, defined as either values less than 200ms, or those smaller or greater than 2 SD from the by-participant mean (n = 1). Generalized linear mixed effects models were constructed to fit RT, using the lme4 package [9] in R [10]. The random structure included by-participant and by-item intercepts and slopes for predictors, under the constraints of model convergence and overfitting. The maximal random structure was identified on the basis of AIC (Alkaike Information Criterion). Predictors included Condition (3 levels: monosyllabic control vs. disyllabic SUBJECT vs. disyllabic OBJECT), and number of PWds (3 levels: 3 vs. 4 vs. 5). Significance testing of the final model was then evaluated using the function Anova() in the car package [11]. In the omnibus analysis we used the monosyllabic sentences in the 3-PWd set as the referent control.

The results of the glmer model yielded no effect of number of PWds on RT, but a significant effect of Condition (Table 3). There was a trend towards significance for the interaction. The model structure was RT ~ Condition * PWd + (PWd | participant) + (1 |item), family = Gamma (link = "log")).

Table 3: *Summary of estimated p-values for the factors in the glmer model fitted on RT*

	Chi-square	**df**	**p-value**
Condition	8.4	2	.02*
PWd	3.6	2	.16
Condition * PWd	8.9	4	.06

To better understand what drives the effect of Condition, separate analyses were carried out for 3-, 4- and 5-PW sentences. For each by-number-of-PWd analysis, the comparable monosyllabic condition served as a reference. RTs

for 3 pairs of conditions were evaluated: (1) for monosyllabic control vs. disyllabic SUBJECT nouns, (2) for monosyllabic control vs. disyllabic OBJECT nouns, and (3) for disyllabic SUBJECT vs. OBJECT nouns, with adjustments for multiple comparisons.

In the by-number-of-PWd analyses, neither the results for the 3-PWd sentences nor those for the 5-PWd sentences revealed any systematic differences in RT for the 3 pairs of conditions. However, the results for 4-PWd targets showed a different pattern; here, RTs for utterances with either a disyllabic SUBJECT noun or a disyllabic OBJECT noun (a stimulus never contained both) were significantly longer than for utterances with monosyllabic controls (see Figure 1). In sum, we did not find an across-the-board effect of the presence of disyllabic words or of the number of PWds, nor of their position within the utterance. Instead, the RT difference for this immediate reading task seems to be specific to the disyllabic conditions in 4 PWd-sentences, as shown by the significant values for the 4-PWd Disyllabic SUBJECT (DS) and Disyllabic OBJECT (DO) conditions shown in Figure 1.

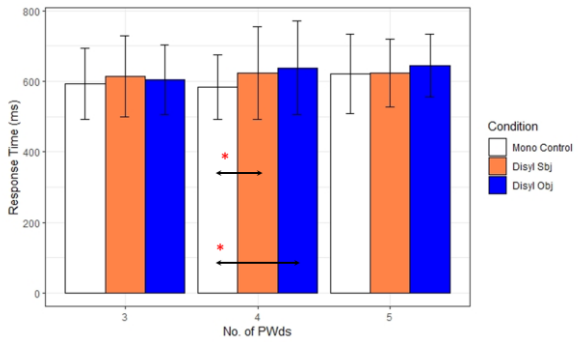

Figure 1: *Mean response time (RT) grouped by the number of prosodic words (PWds) and disyllabic conditions with +/- 1SD.*

4. Discussion & Conclusion

This study set out to explore the possible effects of different phonological complexity on speech production planning, as operationalised in terms of the number of syllables in a PWd (monosyllable vs disyllable), the placement of a larger PWd within the utterance (early, in the SUBJECT NP, vs late, in the DIRECT OBJECT NP), and any possible interactions among these factors, in 3-, 4- and 5-PWd sentences. We anticipated that complexity arising from encoding an additional syllable (comparing stimuli with monosyllabic vs. disyllabic nouns) or PWd (comparing stimuli with 3, 4 and 5 target PWds) would delay speech initiation. In addition, if our reading-aloud task is similar to Wheeldon & Lahiri's immediate production task, i.e. if it taps into the same aspect of the speech planning process, then complexity in a noun phrase early in the utterance would be expected to delay initiation, resulting in longer RTs.

However, these predictions were not borne out in our results. When comparing the general effect of additional syllables or PWds, we did not find any difference in RT (similar to the mixed-block results in Exp. 1 from Meyer et al., [2]. Nor did we find any overall differences in RT as a function of the early vs late location of a complex/larger PWd in an NP within an utterance. This suggests that the reading aloud task in the current study and the immediate production task in Wheeldon & Lahiri might elicit different production strategies. In Wheeldon & Lahiri's immediate production task, a participant silently reads a target phrase and then incorporates it in a sentence response immediately after hearing an eliciting question. Our reading aloud task also has a perceptual aspect, but one of a different type: participants could potentially scan the entire sentence and/or re-focus on a particular constituent that they are planning to produce. Such a strategy is not readily available in the immediate production task used in Wheeldon & Lahiri. This methodological difference may interact in some way with sentence planning; further research is needed to clarify any effect(s) of task-specific strategy on speech production planning, specifically in relation to additional perceptual or comprehension processes.

Although there is no overall effect of phonological complexity on speech initiation, results for the 4-PWd sentences did show significant RT differences for the disyllabic SUBJECT and disyllabic OBJECT sentences relative to monosyllabic control sentences. What could have triggered the effect of an additional syllable in 4-PWd sentences only, while having no effect on the 3-PWd and 5-PWd sentences?

One possible account, invoking the concept of parallelism in syntactic complexity and its possible effect on phonological encoding, is based on the following reasoning. Note that the 4-PWd stimuli are consistently different from the 3- and 5-PWd stimuli in one particular way: there is an incidental difference between the syntactic complexity of the SUBJECT and the syntactic complexity of the OBJECT NPs in the 4-PWd sentences. For example, the SUBJECT NP (adjective + N) is syntactically more complex than the OBJECT NP both in the monosyllabic control sentences ([The strong nurse] loves cheese) and in the disyllabic SUBJECT sentences ([The smart teacher] drinks milk). Similarly, in the disyllabic OBJECT condition the OBJECT NP is syntactically more complex than the SUBJECT NP (The guard likes [fresh flowers]). Such an incidental difference is not present in the 3- or 5-PWd sentences (Table 1), and it is specifically in those 4-PWd sentences that we observed longer RTs for target sentences with a disyllabic word in a syntactically complex NP (whether in the SUBJECT or OBJECT position). This suggests that the lack of parallel structure between the SUBJECT and OBJECT NPs may have an effect on RT, by slowing the planning process in some way.

We consider this fact in the light of observations from structural priming (e.g., [12]; see [13] and [14] for reviews). Results of this work suggest the possibility that a complex SUBJECT NP might prime the planning for a similar structure for an OBJECT NP, in the sense that parallel structure might be pre-activated and readily available to recycle in planning. This might speed up the corresponding phonological encoding process, resulting in shorter RTs. However, when the structure primed by the SUBJECT NP differs from the structure of the OBJECT NP, it may slow the planning process. Since utterance initiation time depends at least in part on the time required for phonological encoding, this lack of parallel syntactic structure might slow the initiation of speech. Perhaps this is why the effects of phonological complexity (i.e., an additional syllable) are restricted to the 4-PW stimuli, where the syntactic complexity of SUBJECT and OBJECT NPs are not parallel. This raises many questions for further research regarding what is 'complex' in the speech planning process, and suggests the possibility of multiple types of complexity, both at syntactic and phonological levels. In addition, this result highlights the importance of precisely defining the nature of complexity that is being tested in any given experimental context.

5. Acknowledgements

We thank our participants for taking part in this research.

6. References

[1] Wheeldon, L., & Lahiri, A. (1997). Prosodic units in speech production. Journal of Memory and Language, 37, 356-381.
[2] Meyer, A. S., Roelofs, A., & Levelt, W. J. M. (2003). Word length effects in object naming: The role of a response criterion. Journal of Memory and Language, 48, 131-147.
[3] Santiago, J., Mackay, D. G., Palma, A., & Rho, C. (2000). Sequential activation processes in producing words and syllables: Evidence from picture naming. Language and Cognitive Processes, 15, 1-44.
[4] Santiago, J., MacKay, D. G., & Palma, A. (2002). Length effects turn out to be syllable structure effects: Response to Roelofs (2002). Language and Cognitive Processes, 17, 15-29.
[5] Meyer, A. S., Belke, E., Häcker, C., & Mortensen, L. (2007). Use of word length information in utterance planning. Journal of Memory and Language, 57, 210-231.
[6] van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. The Quarterly Journal of Experimental Psychology, 67 (6), 1176-1190.
[7] Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. Behavior Research Methods, Instruments & Computers, 35, 116-124.
[8] Boersma, P., & Weenink, D. (2011). Praat: Doing phonetics by computer. http://www.praat.org/
[9] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. https://doi.org/10.18637/jss.v067.101
[10] R Core Team. (2019). R: A language and environment for statistical computing.
[11] Fox, J., & Weisberg, S. (2019). An R Companion to Applied Regression. Third edition, Sage: Thousand Oaks, CA.
[12] Bock, J. K. (1986). Syntactic persistence in language production. Cognitive Psychology, 18 (3), 355-387.
[13] Pickering, M. J., & Ferreira, V. S. (2008). Structural Priming: A Critical Review. Psychological Bulletin, 134 (3), 427-459.
[14] Mahowald, K., James, A., Futrell, R., & Gibson, E. (2016). A meta-analysis of syntactic priming in language production. Journal of Memory and Language, 91, 5-27.