Temporal planning in the production of Australian English compounds

Ivan Yuen, Nan Xu Rattanasone, Gretel McDonald, Rebecca Holt, and Katherine Demuth

Macquarie University

ivan.yuen@mq.edu.au;nan.xu@mq.edu;katherine.demuth@mq.edu.au

Abstract

Listeners rely on prosodic cues to disambiguate syntactic structures. One such ambiguity arises from how nouns are grouped in a sentence. Grouping nouns together as compounds compared to non-compounds should result in temporal adjustment within the word. We investigated how speakers disambiguated the two types using temporal planning, and how these temporal cues were exploited during perception. As expected, compounds showed shorter durations than the non-compounds, with the first word of compounds being shorter than in non-compounds. Compounds were also recognized faster than non-compounds in an eye-tracking task, suggesting a close link between production and perception.

Keywords: temporal planning, compounds, duration

1. Introduction

Listeners can use prosodic cues to disambiguate sentences such as (1).

(1) Steve or Sam and Bob will come.

This sentence entertains at least two possible interpretations, depending on how the nouns are prosodically grouped. In (1a), 'Steve or Sam' functions as one constituent and 'Bob' as another. However, in (1b) 'Steve' constitutes one entity and 'Sam and Bob' another.

(1) a. Steve or Sam # and Bob will come.

(1) b. Steve # or Sam and Bob will come.

The way nouns are prosodically grouped also affects their temporal organization within the sentence. In [1] speakers were instructed to disambiguate the sentence as either (1a) or (1b). When the intervening boundary after 'Sam' in (1a) occurred within a strong-weak Abercrombian foot unit, the duration increased. On the other hand, there was no increase in duration when the unit fell within the larger syntactic grouping. Prosodic cues such as phrase/boundary tone, pitch range and pauses have also been reported to be at the speakers' disposal, disambiguating constituents [2].

While the use of pauses to mark syntactically-structured groupings does not spread to adjacent units, other cues, such as phrase-final lengthening, do [3]. Thus, a speaker's decision to employ different temporal cues to disambiguate syntactic structures provides a glimpse into the scope of the speech planning process.

One type of structural ambiguity arises from the difference between compounds and non-compounds. Two nouns, for example, 'rain' and 'coats', can be combined to form the noun compound, 'raincoats'. In a sentence such as 'I see rain, coats and maps', there are three direct objects. However, the compound 'raincoats' results in only two direct objects in the same sentence. Thus, although the number of syllables in the sentence stays the same, the syntactic structure differs.

The distinction between compounds and non-compounds can be manifested in different ways, with some speakers using durational cues such as pauses or/and lengthening, and others using pitch excursion size and/or pitch reset [4].

The classical distinction between compounds and non-compounds is stress and metrical strength [5, 6, 7]. On the basis of the compound stress rule, the leftmost word is assigned a primary stress, resulting in a 'falling stress contour' in compounds. That is, the leftmost word receives greater metrical strength than the rightmost word, with longer duration associated with stress. Recently, [8] has shown that the phonetic realization of stress (in terms of duration, intensity and pitch) in American English adjective-noun compounds varies in different prosodic contexts. For example, the vowel duration of the leftmost stressed word tends to be longer than the other word in the compound, but such durational differences disappear in sentence-final position. This is probably due to the effect of utterance-final lengthening on the rightmost word of the compound. This might be why speakers opt for other cues such as pitch or intensity to distinguish compounds from non-compounds in different prosodic contexts.

At the lexical level, polysyllabic compounds are a single morphosyntactic and phonological word, whereas non-compounds contain multiple morphosyntactic and phonological words. According to [9], the temporal organization of words is subject to polysyllabic shortening, despite their morpheme and syntactic boundaries. However, [10] showed that the phonological word was accessed as an encoding unit during speech planning, because speakers took longer to prepare a non-compound word than a compound. In addition, sentence duration was longer in non-compounds than in compounds, suggesting that the choice of the encoding unit could affect how the sentence is manifested temporally. But the reported long duration in non-compounds in [10] could also be attributed to the insertion of pauses. It is worth noting that [9] did not examine compounds, whereas [10] based their findings in Dutch compounds and non-compounds. As compounds are recursive phonological words containing internal structure, this raises the question about the granularity of encoding in speech planning. How does encoding manifest in the temporal structure of the compound? Can listeners resort to such temporal cues in distinguishing compounds from non-compounds, as soon as these cues are available?

In this paper we investigated how Australian English-speaking adults distinguish compounds from non-compounds, in both production and perception, shedding light on the production-perception link in speech planning. This is part of a larger study in investigating children's ability to use prosodic cues in disambiguating compounds from non-compounds.

2. Method

The study consisted of both a production experiment and a perception/eye-tracking experiment. All participants received the perception experiment first.

The production experiment consisted of a question-answer elicitation task. To familiarize participants with the task, a

female Australian-English speaking experimenter showed participants a picture and prompted their response with a question: 'What can you see here?'. The participants responded by naming the visual objects and putting the names in a carrier sentence: 'I can see *item1, item2 and item3*' or 'I can see *item 1* and *item 2*'. The picture prompts, which contained either two or three objects, were presented one after the other in a booklet. The presentation order was counterbalanced across participants with half of the participants receiving one order of presentation, and the other half receiving the reverse order. The responses were audio-recorded. To ensure that all participant use the same object names during the elicitation task, an object naming practice session was given at the beginning of each session.

For the eye-tracking experiment, an adult female Australian English-speaker was recorded producing eight two-item and eight three-item sentences in child-directed speech style, at a sampling rate of 44.1 kHz using Audacity 2.0.5. The stimuli were embedded in the carrier sentence 'Look at the *item1, item2, and item3*' or 'Look at the *item1* and *item2*'. These sentences were then converted to stereo mode using MPEG Streamclip. Using Final Cut Pro, we combined the auditory prompts with the visual stimuli from the elicited production task to create videos for each experimental trial.

The average sentence duration was 2616 ms (range: 2280 - 3300 ms). On average, the target compounds began 474 ms from the beginning of the sentence, and the target non-compounds began 519 ms from the beginning of the sentence. The temporal characteristics that distinguished the target compounds from non-compounds included the following: (1) the average duration of the compounds was shorter than that of non-compounds (868 ms vs. 1228 ms), (2) the average duration of word 1 was shorter than word 2 in the compounds (349 ms vs. 482 ms), (3) the average duration of word 1 and word 2 in the non-compounds was the same (512 ms vs. 516 ms).

There were 8 test trials in total, with yoked pairs of compound and non-compound sentences. Each participant heard only 1 of the pair – either the compound or non-compound. This ensured that the same words were heard only once. Each trial lasted approximately 9.5 seconds and consisted of three phases: visual familiarization (4 seconds), auditory prompt (duration of the target sentence), and visual task (3 seconds). During the visual familiarization period, participants saw two sets of pictures containing three objects each. One picture exhibited three different objects, and the other two different objects plus a copy of one of the object. The location of the objects in each picture was counterbalanced to minimize any left vs. right side preference. The trials were randomized to minimize predictability. See Figure 1 for sample trial.

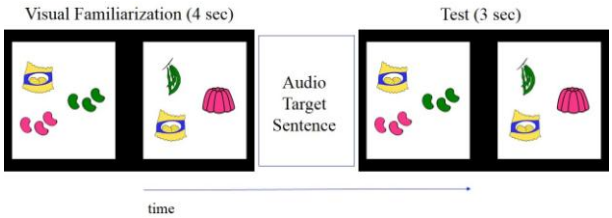

Figure 1: *Sample trial 'jelly beans and chips' vs. 'jelly, beans, and chips' in the eye-tracking study.*

2.1. Participants

Nineteen monolingual Australian-speaking undergraduates from the Sydney area took part in the experiment (5 M, 14 F). Their age ranged from 18 to 30 years (mean = 19 years). They completed the experiment for course credit. Due to bilingualism (3) or excessive creaky voice in the production experiment and/or poor sampling/inattentiveness during the eye-tracking experiment (8), data from 8 of the participants were used for analysis.

2.2. Stimuli

The target test stimuli consisted of two types: (1) a 2-3 syllable compound, and (2) two separate words. Both were embedded in a carrier sentence: 'I can see *item1* and *item2*' for the compound condition and 'I can see *item1, item2 and item3*' for the non-compound condition. The final item in each carrier sentence was a distractor item to prevent the insertion of 'and' between the target items. The sixteen stimuli used in both the elicited production and eye-tracking tasks are shown in Table 1.

Table 1: *Test stimuli*

Compound	2-word Non-compounds
Icecream	Ice, cream
Icecubes	Ice, cubes
Goldfish	Gold, fish
Raincoats	Rain, coats
Jellybeans	Jelly, beans
Jellyfish	Jelly, fish
Waterguns	Water, guns
Waterslides	Water, slides

2.3. Predictions

On the basis of the findings in [10], we expected longer duration in the target non-compound words than in the compounds. We also expected pauses and an increase in pause duration to separate the individual words in the non-compound condition.

If the encoding unit is the phonological word (PW) in speech planning and production, we predicted temporal readjustment of the embedded phonological words in the compounds only. This would result in a difference between the internal temporal organization of compounds and non-compounds. It is also likely that there would be a phrase boundary intervening between word 1 and word 2 in the non-compounds, resulting in boundary-related lengthening. This would lead to word 1 being longer in the non-compounds than the same word unit in the compounds.

We also predicted that these temporal cues would be accessed as soon as available in the perception/eye-tracking task, facilitating disambiguation.

2.3.1. Sentence elicitation: acoustic coding

Using Praat [11], we first identified the target stimuli as either compounds or non-compounds. These items were then further annotated into their component words using waveforms and spectrograms.

The onset consonants of the component words in both compounds and non-compounds were classified as (1) null, (2) stop, (3) approximant, and (4) fricative. The following criteria were employed to identify the beginning of the words. First, we used the onset of clear F2 and voicing for items in Group (1). For items in group (2), we marked the beginning at the onset of the burst release. For items in group (3), we used the intensity minimum and the lowest formant transition in F2 to

code for the beginning of /w/ and the lowest formant transition in F3 to mark the beginning of /ɹ/. Items in group (4) were coded for the beginning of high energy noise.

The same principles were applied to identifying word final consonants in both the target compounds and non-compounds. For nasal consonants, nasal formants were used as the cue to determine word offset. We also annotated the duration of the orientation phrase 'I can see…' in the carrier sentence.

The durations of the compounds/non-compounds, precursor phrases and component words were extracted from a total of 128 sentences for statistical analysis.

3. Results

3.1. Sentence elicitation

First, we examined whether compound durations were different from non-compound durations, using a paired t-test. With an alpha level of 0.05, as predicted durations were significantly longer in non-compounds than in compounds (t (7) = -8.659, p < .001). The mean duration was 691 ms (standard deviation = 86 ms) for compounds versus 1289 ms (standard deviation = 213 ms) for non-compounds (see Figure 2). This is consistent with findings in [10] and the duration patterns of the stimuli modeled by the Australian English speaker.

However, this durational difference might be due to slower speech rate in the non-compound condition. We therefore conducted a paired t-test using the orientation phrase 'I can see …' duration as the dependent variable to check for speaking rate variation. There was no significant difference between the two conditions (t (7) = -1.279, df = 7, p = .242). The mean orientation duration was comparable for compounds (mean = 496 ms) and non-compounds (521 ms), indicating that the durational difference between the compounds and non-compounds could not be attributed to speaking rate variation. In other words, the durational difference did not arise from phonetic implementation.

However, the long duration of the non-compounds could be due to the inclusion of a pause between the two words. There was an average of 269 ms pause duration in the non-compounds, but an average of only 32 ms in compounds. Thus, as expected, speakers employed pauses to separate compounds from non-compounds, indicating a boundary between the two words in non-compounds.

As frequency, which is closely linked to morphological productivity, could also affect word duration [12], it is necessary to factor out this possibility. We therefore analyzed correlation between word frequency and word duration in both compounds and non-compounds. No significant correlations were observed in compounds (r (16) = -.357, p =.17 and non-compounds (r (16) = .42, p =.11).

We then addressed the question whether compounds are encoded as a phonological word (PW) during speech planning. The prediction was that the maximal projection of a PW in compounds would lead to temporal adjustment within the compound word. Durations of the two words within the compound and the non-compound were used as the dependent variable. There were two factors: word type (i.e. compound vs. non-compound) and word position (i.e. word 1 or word 2). A repeated measures ANOVA revealed a significant main effect of word type (F(1,7) = 172.723, p < .001) and word position (F(1,7) = 43.645, p < .001). There was also a significant interaction (F(1,7) = 58.921, p < .001). As predicted, durations of the two words in the compounds differed from those in non-compounds, even though the words in both conditions were PWs. This suggests that speakers gain access to and encode the hierarchical structure of the PW during speech planning.

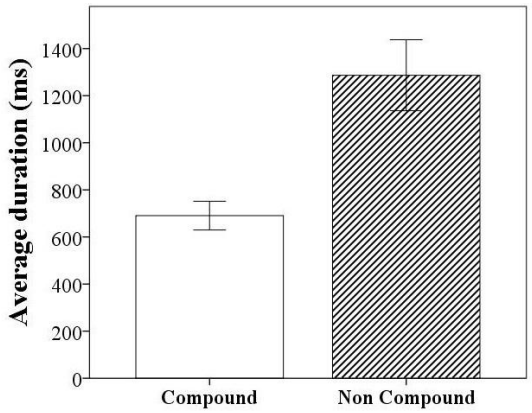

Figure 2. *Average duration (ms) of compounds and non-compounds including pause (+/-2 standard error).*

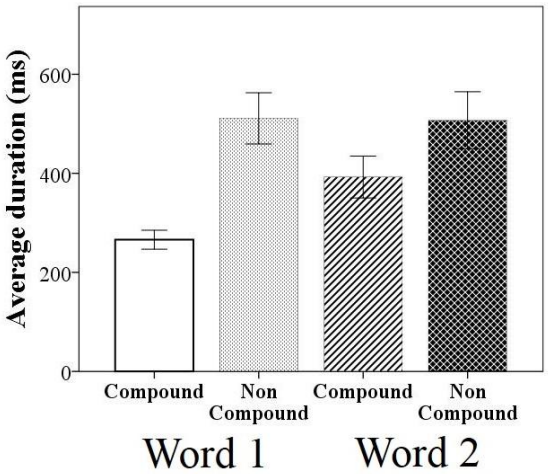

Figure 3. *Average duration (ms) of the 2 words in compounds and non-compounds (+/- 2 standard error).*

The mean duration of word 1 in the compounds was 266 ms and that of word 2 393 ms. Thus, in spite of primary stress on word 1 in the compound, it had shorter duration than word 2. This appears to be counter to the expectation derived from the compound stress rule. Yet it is also possible that the durational difference might be related to prosodic-boundary lengthening at the end of word 2 in the compounds. As regard the non-compounds, the average word 1 duration was 511 ms and the average word 2 duration was 507 ms, (see Figure 3). These duration patterns are similar to those used in the stimuli for the perception study. Thus, the durations of word 1 and word 2 in compounds were shorter than in non-compounds, suggesting temporal re-organization (polysyllabic shortening) in the former.

3.2. Eye-tracking

To examine when participants began to resolve the picture that is consistent with the spoken sentence, we conducted planned comparisons on the proportion of fixations to the target versus distractor pictures. Areas of interest were defined for each picture (65cm x 35cm) and eye-tracking data extracted from these areas. Figure 4 presents the 'time by proportion of fixation' plots for each condition.

For *Compounds*, comparisons were made from 4.8s (at the onset of component word 1) and every 200ms thereafter. With alpha set at 0.05, a significantly greater proportion of fixations to the target than the distractor was found at 5.4s ($t(7) = -3.316$, $p < .001$; mean fixations to target = 0.768, distractor = 0.301), 5.6s ($t(7) = -4.008$, $p = .003$; mean fixations to target = 0.737, distractor = 0.188), and 5.8s ($t(7) = -4.008$, $p = .003$; mean fixations to target = 0.737, distractor = 0.188). No further comparisons were made after these 3 consecutive significant results. No adjustments to alpha were made as the comparisons were a very small subset of all data collected over a 9s period. The results suggest that participants are able to decide on the correct picture at 600ms after the onset of the compounds.

For *Non-compounds*, comparisons were made from 5.2s (at the onset of component word 1) and every 200ms thereafter. A significantly greater proportion of fixations to the target than the distractor was found at 7.6s ($t(7) = -27.000$, $p < .001$; mean fixations to target = 0.982, distractor = 0.018), 7.8s ($t(7) = -6.502$, $p < .001$; mean fixations to target = 0.882, distractor = 0.118), and 8.0s ($t(7) = -2.226$, $p = .031$; mean fixations to target = 0.760, distractor = 0.239). These results suggest that participants are able to settle on the correct picture at 2.4s after the onset of the non-compounds.

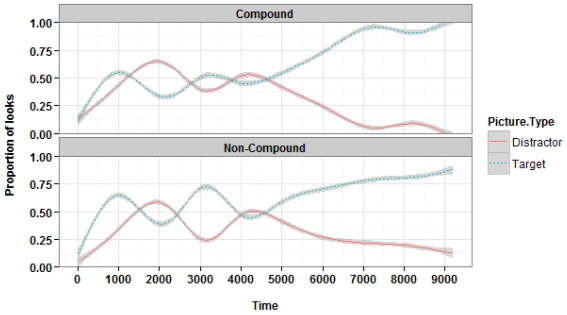

Figure 4. *Average Proportion of fixations to the Target vs. Distractor pictures across time (ms) for Compounds and Non-compounds over each trial*

4. Discussion and conclusion

The production data provided evidence for temporal planning in disambiguating compounds from non-compounds. Non-compounds were longer than compounds. This is not surprising as speakers tend to insert an intervening boundary between the two words in the non-compound condition. This was supported by the larger average pause duration between component words in non-compounds.

As the duration of the preceding phrase did not differ between the compound and non-compound condition, the temporal organization of the sentence cannot be attributed to phonetic implementation of speech rate variation. The production data also indicated that speakers accessed the hierarchical structure of the compound and adjusted the word duration within the compound. Compounds exhibited an unequal word duration distribution, whereas non-compounds showed a more distributed and equal word duration.

Interestingly, the durational cues produced by these participants were also employed to disambiguate the compounds from non-compounds in the perception task. Recognition of the compounds was rapid, within 1 second of hearing the compound words. However, it took four times longer to identify the non-compounds. This might be because the non-compound auditory stimuli are longer than the compounds (1228 ms versus 868 ms). Alternatively listeners might need to accumulate sufficient auditory cues to identify non-compounds, as the words have fairly equal durations, slowing the recognition process. On the other hand, the unequal word durations for compounds is a very salient cue, speeding up the recognition process. These results suggest that listeners can use fine-grained durational cues within a maximally projected phonological word quickly to anticipate the appropriate visual scene, facilitating identification of compounds.

5. Acknowledgements

We thank Stefanie Shattuck-Hufnagel and the Child Language Lab at Macquarie University for helpful comments and suggestions. Partial funding for this research was provided by the ARC Centre of Excellence for Cognition and its Disorders grant #CE110001021x and NIH grant #R01 HD057606 to Demuth and Shattuck-Hufnagel

6. References

[1] Lehiste, I. (1973). Rhythmic units and syntactic units in production and perception, *Journal of the Acoustical Society of America*, 54, 1228-1234.

[2] Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C. 1991. The use of prosody in syntactic disambiguation, Journal of the Acoustical Society of America, 90, 2956-2970.

[3] Turk, A., and Shattuck-Hufnagel, S. (2000). Word-boundary related duration patterns in English. *Journal of Phonetics*, 28, 397-440.

[4] Peppe, S., Maxim, J., and Wells, B. (2000). Prosodic variation in Southern British English, Language and speech, 43, 309-334.

[5] Chomsky, N., and Halle, M. (1968). The sound pattern of English. New York: Harper & Row, Inc.

[6] Halle, M., and Vergnaud, J-R. (1987). An essay on stress. Cambridge, MA: MIT Press.

[7] Hayes, B. (1995). Metrical stress theory: principles and case studies. Chicago, IL: University of Chicago Press.

[8] Morrill, T. (2011). Acoustic correlates of stress in English adjective-noun compounds, Language and Speech, 552, 167-201.

[9] Lehiste, I. (1972). The timing of utterances and linguistic boundaries, Journal of the Acoustical Society of America, 51, 2018-2024.

[10] Wheeldon, L. R., Lahiri, A. 2002. The minimal unit of phonological encoding: prosodic or lexical word, Cognition, 85, B31-41.

[11] Boersma, P., and Weenink, D. (2011) Praat: doing phonetics by computer. Version 5.2.21.

[12] Hay, J., and Baayen, R. H. (2001). Parsing and productivity. In Booji, G. E. And van Marle, J. (ed), Yearbook of Morphology, 203-235.