INDIVIDUAL DIFFERENCES IN WORKING MEMORY CAPACITY AND THEIR EFFECT ON SPEECH PROCESSING

Annie C Gilbert[1,2], Victor J. Boucher[1], Boutheina Jemel[2,3]

[1]Laboratoire de Sciences Phonétique, Université de Montréal, Canada
[2]Laboratoire de Recherche en Neurosciences et Électrophysiologie Cognitive,
Hôpital Rivière-des-Prairies, Canada
[3]Centre de Recherche Fernand-Séguin, Département de Psychiatrie, Université de Montréal, Canada
annie.gilbert@umontreal.ca, www.phonetique.info

ABSTRACT

Though tests of working memory (WM) correlate with scales of language development, it is unclear how WM capacity relates to spoken-language processing. However, Gilbert et al. (2014) have shown that listeners perceptually chunk speech in temporal groups (TGs) and that the span these TGs influences memory of heard items. Assuming that WM capacity links to this processing of speech in groups, listeners with the highest WM spans would be better at recalling items from long TGs. To examine this, we presented two sets of stimuli (utterances and sequences of meaningless syllables) containing long TGs. After each stimuli, listeners had to determine if a target item was previously heard. An analysis using *GLME* models showed that correct recognition memory of items heard in utterances was significantly better for listeners with high WM spans than for listeners with smaller spans. The effect was marginally significant for sequences of nonsense syllables.

Keywords: Speech processing, Perceptual chunking, Rhythm, Working memory, Digit-span.

1. INTRODUCTION

A body of research has shown that performance on tasks that measure the span of working memory (WM) correlates with scales of language ability, especially vocabulary size (e.g. [1, 2]). Moreover, a history of findings shows that deficits on WM tasks characterize several pathologies of language and verbal development (see [3] for a review). Yet, despite this body of work, there is a surprising paucity of research on how WM actually operates in processing *spoken* language. In fact, studies on the role of WM in language processing have focused almost exclusively on written material where the “reading span” linked to constraints on WM [4]. This work has shown that performance on WM tasks such as the “reading span” correlates with the parsing of text as well as the understanding of discourse [5-7]. Unfortunately, such findings cannot extrapolate to spoken language processing. The difference with respect to spoken language, however, is that WM would apply to some extent or *span of speech*, and reflect constraints that would be measured not in terms of text units but in terms of some number of sound-based items.

Among the few studies that have examined how WM applies to speech processing, an experiment by Gilbert et al. [8] made use of Event-Related Potentials (ERPs) to investigate how the span of perceptual chunks influence the memory trace of heard items. They used amplitude variations of the N400 as an index of the quality of the memory traces with relative decreases in amplitude being associated with better memory for heard items. In their experiment, listeners were presented with utterances containing temporal groups (TGs) of 3 and 4 mono-syllabic words followed by a prompt-word which was occasionally present in the heard utterance. The ERPs showed that listeners were *perceptually chunking* the utterance in terms of TGs, and that the size of these groups influenced the amplitude of the N400 in a way that suggested an involvement of WM. In particular, words presented in 3-syllable TGs had a better memory trace compared to those in larger TGs. A related experiment [9] showed a similar effect of TGs in sequences of meaningless syllables indicating that the span of the chunks influenced the quality of memory traces regardless of the content of speech stimuli. These results suggest that constraints on WM can operate on-line in terms of the span of a perceptual chunk or TG, which affects the memory trace of heard elements.

On the other hand, the designs of [8] does not show the link between recognition of items in perceptual chunks and *individual* memory capacity as measured via such WM tasks as the digit span [10]. If it is the case that the span of perceptual chunks of speech reflects constraints on a listener’s WM, then individual scores on digit-span tasks should correspond to differences in the recognition of items in perceived chunks. With the purpose of verifying this general prediction, we elaborated a behavioural version of the experiment of [8] that we adapted using longer TGs in order to increase the memory load. Our general hypothesis is that individual WM capacity as indexed by a digit-span task would lead to subject-related differences in memory of heard speech items. Specifically, we expect that listeners with better digit-spans will have better recognition memory for items presented in relatively long perceptual chunks or TGs.

2. METHOD

2.1. Participants

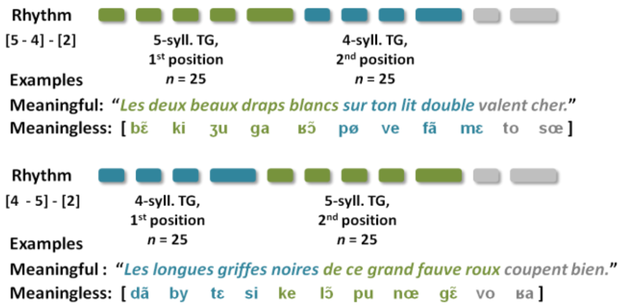
Participants were 20 native speakers of French recruited at the Université de Montréal (aged 19 to 41 years; average 25.5 years; 7 men). All presented normal hearing in terms of a standard audiometric evaluation and a normal memory span according to the digit span test of the WAIS [10] (overall, average normalized score : 10.16, std dev.: 2.4).

2.2. Stimuli

2.2.1. Stimuli design

The stimuli comprised two sets of sequences of 11 syllables: the first set included 50 meaningless sequences of syllables, and the second 50 meaningful French utterances. These stimuli were elaborated with 3 temporal groups (TGs) of 4, 5, and 2 syllables, as illustrated in Figure 1. Moreover the order of the TGs was made to vary: one subset contains TGs of 5, 4, and 2 syllables whereas the other subset contains TGs of 4, 5, and 2 syllables. This design served to compare effects of size and position of TGs while reducing suffix effects (located in the last 2-syllable group, which was not considered in the analysis).

Figure 1:. Schematic representation of TG structures used, and example of stimuli (meaningful and meaningless).


Rhythm
[5 - 4] - [2]
5-syll. TG,
1st position
n = 25
4-syll. TG,
2nd position
n = 25
Examples
Meaningful: "Les deux beaux draps blancs sur ton lit double valent cher."
Meaningless: [bɛ̃ ki ʒu ga ʁɔ̃ pø ve fɑ̃ mɛ to sœ]
Rhythm
[4 - 5] - [2]
4-syll. TG,
1st position
n = 25
5-syll. TG,
2nd position
n = 25
Examples
Meaningful : "Les longues griffes noires de ce grand fauve roux coupent bien."
Meaningless: [dɑ̃ by tɛ si ke lɔ̃ pu nœ gɛ̃ vo ʁa]

Every stimulus was followed by a target (syllable or monosyllabic lexeme) selected from either the first TG (in 50% of the cases) or the second TG (50%). Filler stimuli were also created to vary the presentation of rhythmic, syntactic and intonation patterns. Both types of stimuli were controlled with respect to the following attributes.

Meaningless series of syllables were created using consonant-vowel (CV) syllables of French. Each series was balanced with no repeated C or V within a series and no combination creating recognizable lexemes. Syllable order was controlled so that no consecutive syllables shared a common point of articulation (to prevent confounding effects on recognition recall).

Meaningful utterances were created using monosyllabic lexemes and functors with a high index of familiarity in French [11]. These were arranged in a given syntactic structure so that the first TG always contained the subject, the second TG contained a complement to the subject, and the final TG contained the verb phrase. All were literally plausible.

Note that the TGs we used (4- and 5-syll.) are longer than the average TGs found in normal speech [12, 13], but are not uncommon. They also match or exceed the optimal group size of 3-4 items for serial recall [14]. We chose these limit-size groups to minimize possible ceiling effects (by reference to the near-ceiling effects reported in [8]).

2.2.2. Stimuli recording

The above stimuli were produced by a native speaker of French following the pacing technique described in [8]. This technique insured the production of TGs of constant durations (4-syll TG = 1,150 ms, 5-syll. TG = 1,400 ms, 2-syll. TG = 650 ms.) marked by a lengthening of the last syllable corresponding to French prosody [15, 16]. Recordings were performed in a sound-treated booth using an external sound card (M-Audio Fast-Track Pro, 44,1kHz, 16 bits, mono). Every stimulus series was saved in an individual sound file and amplitudes were normalized. Filler stimuli (meaningful and meaningless utterances) were also recorded following different prosodic patterns to vary the presentations.

2.3. Procedure

Presentation of the sets of meaningful and meaningless stimuli was counterbalanced across subjects such that half of the participants heard the meaningless stimuli first. The stimuli were presented using insert earphones (*Eartone* 3A, EAR Auditory Systems). Participants were instructed to listen to a sequence or utterance and then determine, via a key press, as fast as possible, whether the following prompt was part of the preceding stimulus. Sound files were played back via *E-prime* 1.0 (Psychology

Software Tools) in random blocks divided by rest pauses. The sounds were delivered at a constant intensity (peak levels of 68 dBA at the ears).

2.4. Statistical analyses

The overall effect of TG size (4 vs 5 syllables) and position (1st vs 2nd) was analysed using repeated measures analyses of variance (ANOVA). As for the effect of individual differences in digit span score, this was analysed using models of *Generalized Linear Mixed Effects* (GLME; as implemented in lme4 package, version 1.1-7 [17], in the *R Project for Statistical Computing environment*, version 3.1.3 [18]). The reason for choosing this method over traditional statistical analyses is that allowed us to takes into account subject-related variability in investigating effects of digit-span on the recognition of heard elements in the test context [19]. The two models (one for each type of stimuli) were fitted with *scaled digit-span scores* and *TG properties* (size and position) as fixed effects, and *subjects* and *target items* were entered as random effects (with random intercepts).

3. RESULTS

3.1. TG size and position

Table 1 summarizes the scores of item recognition for both meaningful utterances and meaningless series of syllables. As expected, the recognition of items was much higher overall for meaningful utterances compared to meaningless sequences of syllables, which was a much harder task. A 2X2 ANOVA of recognition scores for meaningful utterances showed a significant main effect of TG position [$F(1,19) = 14.389$; $p = 0.001$; $\eta 2 = 0.431$], but no effect of TG size, and no interaction between factors [$F(1,19) < 1.859$; $p > 0.189$; $\eta 2 < 0.089$].

Table 1: Overall mean accuracy and standard error of recognition scores as a function or TG length and position.

		utterance type	
		meaningful	**meaningless**
TG position	**TG size**	accuracy (%) *SE*	accuracy (%) *SE*
1st	4-syll.	89.6 *5.6*	63.1 *10.1*
	5-syll.	92.2 *5.2*	58.7 *10.6*
2nd	4-syll.	94.8 *4.0*	75.2 *8.8*
	5-syll.	94.8 *3.7*	76.3 *8.8*

Overall, then, recognition scores were significantly higher for targets heard in the second TGs compared to targets heard in the first TGs.

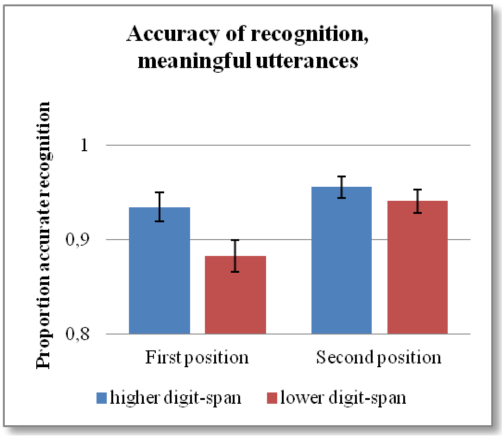
The same pattern of results was found for meaningless syllable series: there is a significant main effect of TG position [$F(1,19) = 29.186$; $p < 0.001$; $\eta 2 = 0.606$], but no effect of TG size and no interaction between factors [$F(1,19) < 1.089$; $p > 0.31$; $\eta 2 < 0.054$]. Again, recognition memory of targets is better for items heard in the second TGs compared to items heard in the first TGs.

3.2. Individual differences in memory capacity.

3.2.1. Impact on recognition of items in meaningful utterances

As for the mains results bearing on the effects of individual span capacity on item recognition, the analyses of these effects used the *GLME* models. As noted, these models were fitted with recall accuracy as dependent variable, and digit span score and TG position as fixed effects (keeping subject and targets as random effects). TG position was kept in the models given that the above results (section 3.1) showed significant position effects on overall scores of item recognition. The *GLME* model fit for the meaningful utterances yielded significant main effects of both digit span score [$Z = 2.956$; $SE = 0.221$; $p < 0.01$] and TG position [$Z = 2.018$; $SE = 0.328$; $p < 0.05$]. This indicates that participants with higher memory span had better recognition scores than participants with lower spans, even though effects of TG position remained significant. These effects are illustrated in Fig.2.

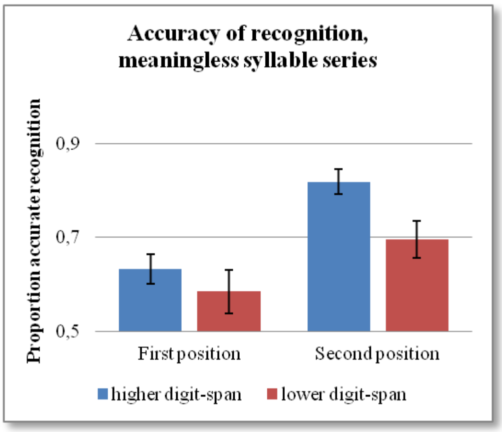
Figure 2: Proportion of target-item recognition for meaningful utterances as a function TG position and individual differences in memory span. Note that the differences in memory capacity are displayed in terms of a median split of digit-span scores.



3.2.2. Impact on recognition of items in meaningless sequences of syllables.

As for the recognition of items presented in nonsense sequences, the *GLME* model fit yielded a significant effect of TG position [$Z = 3.982$; $SE = 0.2062$; $p < 0.001$], but the effect of digit span was only marginally significant [$Z = 1.877$; $SE = 0.1706$; $p = 0.0605$]. These results suggest that participants with higher memory capacity may not be better at recognizing items presented in meaningless series of syllables. As in the case of meaningful utterances, recognition of items within TGs was significantly influenced by the position of the TGs. The effect is seen in Fig. 3. (Note that we adjusted range of the abscissa for Fig. 2 and 3 so illustrate the similar effects across the two types of stimuli).

Figure 3: Proportion accurate recognition of targets following meaningless series of syllables with regard to TG position and individual differences in memory capacity (median split on digit span scores).



4. DISCUSSION / CONCLUSION

Research has shown that individuals' WM memory capacity can correspond to a "reading span" which impacts memory of script and various aspects of text processing [4-7]. However, no study has examined how listeners' WM capacity can relate to some "span of speech" that impacts memory of heard sound-based items. Previous work by Gilbert et al. [8, 9], has shown that listeners perceptually chunk speech in temporal groups (TGs), and that the span or size of these groups affects immediate memory of heard items. The above experiment aimed to investigate whether listeners' WM capacity, as measured using a task of digit span, links to their recognition of items within relatively long chunks or TGs (of 4 and 5 items). We reasoned that, if there is a link between WM and item recognition, then individuals' with high WM capacity would manifest greater recognition memory of items presented in these long TGs. The above results generally confirm this link but with some provisions.

The main results using *GLME* statistical models showed that individuals with a greater memory capacity were significantly better at recognizing items in presented TGs of meaningful utterances. A similar effect was seen with presented sequences of meaningless syllables except that the effect was marginally significant (which suggest a potential problem of statistical power and sample size). It is important to note that these results complement the findings of Gilbert et al. [8, 9, 20] who found significant effects of both TG size (3- and 4-syllables) *and* position on memory of heard items presented in utterances and meaningless sequences of syllables. The absence of effects of TG size in the present study was expected: since we intended a difficult recognition task to evaluate effects of a wide range of digit-span scores, we used large chunks of 4 and 5 items, which exceed the easier, optimal chunks of 3-4 items [14]. In this context, the present findings show that listeners' WM capacity as measured via such tests as the digit span can relate to a "span of speech" or perceptual chunk. However, further confirmation of this link likely requires a larger sample of participants than that of the present study.

5. REFERENCES

[1] Baddeley, A., Gathercole, S.E., Papagno, C. 1998. The phonological loop as a language learning device. *Psyc. Rev*. 105(1): p. 158-173.

[2] Gathercole, S.E., Baddeley, A.D. 1993. *Working memory and language*. Hillsdale, USA: L. Erlbaum Associates.

[3] Baddeley, A. 2003. Working memory and language: an overview. *J. of Com. Dis.* 36: p. 189-208.

[4] Daneman, M., Carpenter, P.A. 1980. Individual differences in working memory and reading. *J. of Verb. Learn. and Verb. Behav*. 19(4): p. 450-466.

[5] Just, M.A., Carpenter, P.A. 1992. A capacity theory of comprehension: Individual differences in working memory. *Psyc. Rev*. 99(1): p. 122-149.

[6] MacDonald, M.C., Just, M.A., Carpenter, P.A. 1992. Working memory constraints on the processing of syntactic ambiguity. *Cog. Psyc*. 24: p. 56-98.

[7] Fedorenko, E., Gibson, E., Rhode, D. 2006. The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *J. of Mem. and Lang.* 54: p. 541-553.

[8] Gilbert, A.C., Boucher, V.J., Jemel, B. 2014. Perceptual chunking and its effect on memory in speech processing: ERP and behavioral evidence. *Frontiers in Psychology*. 5(220): p. 1-9.

[9] Gilbert, A.C., Boucher, V.J., Jemel, B. In Press. The perceptual chunking of speech: a demonstration using ERPs. *Brain Research*.

[10] Wechsler, D. 2008. *Wechsler Adult Intelligence Scale - Fourth Edition*. San Antonio, TX: Pearson.

[11] Desrochers, A. 2006. OMNILEX : Une base de données sur le lexique du français contemporain. *Cahiers Linguistiques d'Ottawa*. 34: p. 25-34.

[12] Dauer, R.M. 1983. Stress-timing and syllabe-timing reanalyzed. *J. Phonetics*. 11: p. 51-62.

[13] Martin, P. 1999. Intonation of spontaneous speech in French. *Proc.of the 14th ICPHS*. p. 17-20.

[14] Cowan, N. 2000. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behav. and Brain Sci.* 24: p. 87-185.

[15] Fant, G., Kruckensberg, A., Nord, L. 1991. Durational correlates of stress in Swedish, French and English. *J. Phonetics*. 19: p. 351-365.

[16] Delattre, P. 1966. A comparison of syllable length conditioning among languages. *Intl. Rev. of Applied Ling.* 4: p. 183-198.

[17] Bates, D., et al. 2014. *lme4: Linear mixed-effects models using Eigen and S4.* R package version 1.1-7. http://CRAN.R-project.org/package=lme4

[18] Team, R.C. 2014. *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* http://www.R-project.org/.

[19] Baayen, R.H., Davidson, D.J., Bates, D.M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *J. of Mem. and Lang.* 59(4): p. 390-412.

[20] Gilbert, A.C., Boucher, V.J., Jemel, B. 2012. Effects of temporal chunking on speech recall, in *Proc. of the 6th International Conference on Speech Prosody*, Q. Ma, H. Ding, and D. Hirst, Editors. Tongji University Press: Shanghai, China. p. 524-527.