Commentary on

“Sentential Influences on Acoustic-Phonetic Processing: A Granger Causality Analysis of Multimodal Imaging Data”

Arthur G. Samuel

Basque Center on Cognition, Brain and Language, Donostia – San Sebastián, Spain.

IKERBASQUE, Basque Foundation for Science.

Stony Brook University, Dept. of Psychology, Stony Brook, NY, USA

In their paper, “Sentential Influences on Acoustic-Phonetic Processing: A Granger Causality Analysis of Multimodal Imaging Data”, Gow and Olson take on one of the most contentious issues in the speech perception field: Is the architecture of the speech perception system entirely bottom-up, or do higher-level representations (e.g., lexical or sentential) affect the operation of lower-level (e.g., phonetic) ones? In the literature, this argument has pitted “autonomous” models (e.g., Massaro, 1989; Norris, McQueen, & Cutler, 2000) against “interactive” ones (e.g., Grossberg, 1980; McClelland & Elman, 1986).

Both model classes ultimately grew out of a conception of spoken word recognition – the cohort model – pioneered by Marslen-Wilson and his colleagues (e.g., Marlen-Wilson & Tyler, 1980; Marlen-Wilson & Welsh, 1978). The basic premise of the cohort model was that when a listener hears a spoken word, an initial “cohort” of lexical candidates gets activated by the word’s beginning, with the cohort then winnowed down to the correct item as more information becomes available. In terms of the top-down versus bottom-up issue, the cohort theory posited a mixture: The generation of the initial cohort was argued to be determined by bottom-up factors (i.e., the set of lexical candidates that matched the acoustic-phonetic information during the first 150 msec or so of the word), but the winnowing process used both bottom-up mismatch information and top-down (semantic) information. Experiments using the gating paradigm (Tyler, 1984; Tyler & Wessels, 1983) provided early evidence for this position, as a sentence’s semantic context did not seem to constrain the nature or number of initial candidates in

the cohort, but it did affect the speed with which the candidates were eliminated from contention and thus how quickly the correct word was recognized.

As theorists progressed from the cohort model's starting point, two different directions were taken. One of these picked up on the bottom-up mechanism as the driving force, and the other expanded the role of top-down processes. The ensuing argument has been going on for decades, and has filled countless pages; it will not be resolved by a single finding, or a single paper. In part, the durability of the argument reflects the different types of evidence that each side can bring to bear. Those who argue for interactivity produce demonstrations in which manipulating a higher-level factor affects the outcome at a lower-level. Those who favor autonomous models either show null effects of a higher-level factor on outcomes at a lower level, or argue that the apparently top-down effects shown by others could instead be due to post-perceptual decision-level influences.

The prototypical domain for these two positions is the lexical influence on phonetic identification first reported by Ganong (1980). Ganong showed that the interpretation of an ambiguous phonetic segment is affected by lexical factors. For example, a segment that is midway between /d/ and /t/ will be reported as "d" if it begins "dash", but the same segment will be reported as "t" in "task". Theorists favoring interactive models take this outcome as evidence that lexical activation directly affects phonetic encoding, while theorists favoring autonomous models argue that the effect occurs at a decision level, with listeners post-perceptually combining the lexical and sublexical information.

As Gow and Olson note, one of the objections to interactive models is philosophical rather than empirical: If the system already has enough information to activate the correct word (e.g.,"task" or "dash"), what purpose is there to go back and shift the phonetic code (e.g., toward /t/ or toward /d/)? This argument also applies to increasingly popular "forward models" that seem to be modern variants of "analysis by synthesis" (Poeppel & Monahan, 2011; Stevens, 1960), as in these models the system is assumed to generate a kind of hypothesis (the synthesis) and then to match that hypothesis against the input signal. If the hypothesis is good, there is a match, but in that case the system apparently already knew enough to generate the right hypothesis; if the hypothesis is bad, there is a mismatch. So, the question is, what is gained here? Gow and Olson note that interactivity can help with error correction. Perhaps. Perhaps a better answer, at least for interactive models, is that such models actually assume that the higher-level and lower-level codes are settled on simultaneously (with a possible lag at the higher level). The selection of both the phonetic code and the lexical code will occur as the two representations resonate with each other, with compatible representations mutually supporting each other's activation. Looked at this way, there is no "going back" to the lower level when a higher level code is activated because the two codes are being determined at the same time. As we will see, this conception is quite plausible for the relationship of acoustic-phonetic and lexical encoding; whether it makes sense in the domain of Gow and Olson's paper – sentential context effects on phonetic encoding – remains to be seen. There is evidence from multiple sources (Connine, Blasko, & Hall, 1991; Swinney, 1979) that suggests that phonetic encoding

remains malleable for about one second, a duration that is comfortably within a lexical window but not necessarily within a sentential one.

In assessing the unresolved dispute about model architecture, Gow and Olson note that the evidence that has been brought to bear essentially comes down to work using behavioral measures, or work based on fMRI. With respect to the latter, they argue that even the most sophisticated methods of processing fMRI data cannot provide the kind of temporal resolution needed to adjudicate the argument. This seems like a reasonable conclusion, given the time scale that is involved. They also suggest that behavioral methods "are at a disadvantage because they typically depend on overt judgments performed after either interaction or selection has taken place" (PAGE 4).

Although they are quite correct that most behavioral tasks do run into this problem, not all do. It is possible to design a behavioral task that gets around this limitation. For example, Samuel (2001) started with a Ganong-type effect, but never asked listeners to report how they heard these stimuli. Instead, the Ganong effect was used to drive perception of an ambiguous sound, and that lexically-determined percept in turn was used to drive selective adaptation. For example, listeners heard words like "arthritis" or "malpractice" in which the final /s/ was replaced by a sound that was midway between 's' and 'sh'. In another condition, the ambiguous sound replaced the final segment of words like "abolish" or "demolish". In a selective adaptation study, playing some sound repeatedly (the "adaptor") reduces subsequent identification of similar sounds. In this case, the same ambiguous mixture reduced subsequent report of 's' (on a continuum of test syllables that ranged from "iss" to "ish") when the ambiguous sound occurred in words like "arthritis" and "malpractice", but reduced report

of ‘sh” when the sound occurred in words like “abolish” or “demolish”. By looking for a consequence of the Ganong effect, rather than by getting reports from listeners about the words themselves, this behavioral study provides support for interactive models: The lexical activation drives perception of the ‘s’ or of the ‘sh”, but listeners are never required to decide about those sounds; instead, because of what they heard, differential adaptation shifts are found for test syllables.

The Ganong-based selective adaptation effect demonstrates that behavioral techniques can indeed be used to provide decisive results that bear on the interactive/autonomous debate, but Gow and Olson are right that most behavioral studies are open to both interpretations. Taken together with the well-founded concern about fMRI evidence, the limitations on behavioral studies leave researchers with very little evidence that truly can discriminate between the two model classes. Gow and Olson argue that their Granger causality method offers a way to overcome the paucity of decisive empirical findings, and I believe that this new method is indeed a potentially very powerful tool. In fact, Gow, Segawa, Ahlfors, & Lin (2008) demonstrated this power in a study of the lexical Ganong effect. They showed that when listeners hear a Ganong stimulus, the supramarginal gyrus (SMG) causally increases activation in the superior temporal gyrus (STG) during a window 280-480 msec after word onset. This is the time period during which lexical information became sufficient to determine which word the listener was hearing. Critically, based upon a substantial body of previous research, there is good reason to believe that the SMG is a “lexical” region, and that the STG is an “acoustic-phonetic” region. Therefore, if the (lexical) SMG modulates the activity of the (acoustic-phonetic) STG, the Granger causality analysis provides

evidence for neural activation patterns that are consistent with interactive models, and contrary to autonomous ones.

As noted, an essential piece of the argument is the knowledge that SMG is processing lexical information, and that STG is processing acoustic-phonetic information. Gow and Caplan (2012) acknowledge the importance of such independent knowledge of a region's function when using Granger causality – there must be a clear prior understanding of what an area is doing, based on fMRI studies, lesion studies, etc. An important issue for Gow and Olson's use of Granger causality to look for sentential effects is the weaker prior knowledge in this domain. They note that there is a "largely unsystematic mapping between supralexical semantic or syntactic representation and speech sounds" (PAGE 23). This concern is compounded by their finding a relatively large number of regions (over a dozen) that show some kind of Granger causality on (acoustic-phonetic) STG activation, with most of these occurring in frontal regions that potentially reflect the kind of problem-solving (and thus decision level) effect that would be expected by autonomous, rather than interactive, models.

Because of the uncertainty about higher-level areas, and how they might influence acoustic-phonetic encoding, Gow and Olson look for a lexical mediator: Perhaps sentence context predicts a particular word, and then the activated word could affect acoustic-phonetic processing. This approach leads the authors to a slightly indirect causal route: Although they do not see a direct Granger causal link between the lexical SMG region and STG, they do find that left posterior medial temporal gyrus (left pMTG) does affect SMG, and since SMG has been shown (Gow et al., 2008) to

affect STG, the causal relationship between left pMTG and STG is taken as evidence for lexical mediation of sentential context effects on phonetic encoding.

While this is a possible interpretation of the results, it raises a rather fundamental question about how sentential context could produce a top-down effect on phonetic processing. The approach requires sentential context to predict a specific word, and then to have the already-established interactive lexical effect produce the influence on phonetic processing. With this approach, the sentential effect is really just a variation on the lexical effect. There is nothing inherently wrong with this, though it seems to lose some of the mystique that an independent sentential effect might have. It also runs into a potential conflict with the available behavioral literature. Gow and Olson are aware of this conflict: Work by Connine (1987; Connine & Clifton, 1987) suggests that sentential effects on phonetic encoding differ from those produced by lexical context.

As Gow and Olson note, Connine's experiments involved a comparison of lexical disambiguation of a phonetic ambiguity (i.e., the Ganong effect) with sentential disambiguation, focusing on reaction time effects rather than category boundary shifts (because both types of disambiguation push identification of an ambiguous item in the direction one would expect). The critical patterns of reaction times come from the crossing of three stimulus types with two points along the test continua. The three stimulus types were Ganong-style word-nonword pairs (e.g., "dice"-"tice", or "dype"-"type"), nonword-nonword pairs (e.g., "dicel"-"ticel"), and word-word pairs in sentence contexts (e.g., "dent"-"tent", in either "She drives the car with the [dt]ent", or in "She saw the show in the "[dt]ent"). In all three cases, the critical words or nonwords were phonetically ambiguous; for the Ganong and sentence experiments, the context could

shift the interpretation of the ambiguous sound, and for the nonword-nonword pairs a shift was induced by paying some subjects more for voiced answers (e.g., “dicel”) and other more for unvoiced answers (e.g., “ticel”). The two points along the test continua were (1) the category boundary region, and (2) the endpoint regions. Connine found that for the Ganong case, reaction times near the category boundary region shifted in a way that matched the identification shift, but that there was no effect on reaction times for items near the continuum endpoints. In contrast, the payoff manipulation for the nonword-nonword pairs shifted the identification function near the boundary region, but only affected reaction times near the endpoints – the complement of the lexical effect. Critically, for sentential context, the results mirrored those with payoffs: There was an identification shift near the boundary, but reaction times were only affected near the endpoints. Connine concluded that the lexical shift is a true interactive top-down effect, but that the sentential effect is a decision-level effect because it matches the pattern found for an explicit decision bias manipulation – differential payoffs for different decisions.

Gow and Olson recognize that the dissociation that Connine posits for lexical versus sentential context is at odds with their suggestion that their sentential effect is essentially just a mediated (true interactive) lexical effect. They therefore raise some procedural questions about the Connine (1987; Connine & Clifton, 1987) studies, and point out that a later study (Borsky, Tuller, & Shapiro, 1998) failed to replicate the dissociation. I believe that Connine’s findings cannot easily be dismissed, in part because there is converging behavioral evidence, and in part because there is an important procedural difference in the failure to replicate that may well have important

implications for interpreting both the conflicting studies, and Gow and Olson's own findings.

The converging evidence comes from work done on phonemic restoration, an effect first reported by Warren (1970). He removed a speech segment from a word in a sentence and replaced the segment with an extraneous coughing sound. When the resulting stimulus was played to listeners, they were consistently unable to identify what speech segment was missing – they appear to have perceptually restored it. Samuel (1981) used signal detection procedures to assess whether the restoration was in fact perceptual, or was instead some kind of post-perceptual decision effect. On each trial, listeners either heard a stimulus similar to what Warren created (speech with a segment replaced by white noise), or a stimulus in which nothing was removed but noise was superimposed on a segment. The task was to indicate whether a given stimulus was the first type (something was missing) or the second type (nothing was missing). The idea was that if people perceptually restore the missing speech in the first type, they will be hearing the second type, making this discrimination quite difficult. Poor discrimination was indexed by *d'* scores near zero. If there is simply a bias toward reporting stimuli as intact, regardless of whether they actually were or were not, this will show up as a change in *Beta*, the bias parameter in signal detection. Samuel found that lexical context (measured several ways, including a comparison of restoration in real words versus pseudowords) did produce a *d'* effect, consistent with a true top-down lexical effect on phonetic encoding. In contrast, sentential predictability of a word did not affect *d'* – it shifted the *Beta* values, consistent with a decision effect. Thus, the

results from the phonemic restoration work completely converge with those from Connine's (1987; Connine & Clifton, 1987) studies of ambiguous segments.

Given this convergence that suggests a dissociation between lexical and sentential context effects on phonetic encoding, what should one make of Borsky, Tuller, and Shapiro's (1998) finding different results than Connine (1987)? In reaction time analyses, Borsky et al. found that sentence context produced effects near the continuum boundary, not near the endpoints. This pattern matches what Connine and Clifton (1987) found for lexical context, and contrasts with what Connine found for sentences. The conflicting results may well be due to the very different types of sentence context used in the two studies. Connine used sentences that rarely included very strong associates of the critical (ambiguous) items. As in the "dent"-"tent" example given above, or in the "goat"-"coat" stimulus pair that she used ("She hurried to feed the [gc]oat", or "She wanted to wear the [gc]oat"), there is no word in the sentence context that is strongly associated with the critical word (e.g., there are many things one might feed, and many things one might wear). Borsky et al. used sentences that included stronger associates in the context, often multiple strong associates. For example, for "goat"-"coat", possible sentences were "The busy zoo-keeper wanted to cage the [gc]oat" and "The cheerful tailor had to dry-clean the [gc]oat". After hearing "tailor" and "dry-clean", regardless of the sentence context per se, "coat" is very likely to attain some activation by pure word association, presumably within the lexicon. If so, then finding the reaction time pattern seen for a lexical manipulation may be quite sensible, but may not say much about higher-level sentential context, if the effect is actually taking place entirely within the lexicon.

This analysis, of course, leads one to ask whether the stimuli in Gow and Olson's study were likely to have produced their effects via word association within the lexicon; if so, then finding a causal path between the lexical SMG and the acoustic-phonetic STG would largely be a confirmation of what Gow et al. (2008) demonstrated for pure lexical context. In looking through the sentences shown in the appendix, my impression is that there is a real mix, with some sentences clearly being more like those used by Connine (i.e., relatively weak associations from words within the sentence and the critical word) and others more like those used by Borsky et al. (i.e., strong paired associates). This mix might account for both the finding of the lexical causality (associative priming within the lexicon), and for the indirect connection that was observed (pMTG activation of SMG, followed by SMG activation of STG). If consistently strong associates had been used, it is possible that the Granger analysis might have found direct SMG to STG causality.

This is of course just speculation, based on differences in the behavioral literature. Thus, I suggest the following thought experiment (or perhaps, real experiment): What might happen if in each of the sentences that were tested here, all of the non-noun and non-adjective words were muffled by filtering/noise sufficiently to make them entirely unintelligible, leaving the nouns and adjectives intact? For most of the sentences, this would preserve the strongest associates of the target words, but would render the sentences into non-sentences – some words interspersed with mostly unintelligible sound. If the effects reported here are in fact based on simple word associations, rather than any kind of syntactic/semantic sentence processing, then these new stimuli should produce the same pMTG to SMG to STG causal chain. If

instead the current results reflect some kind of actual *sentential* predictability, then this chain should be broken.

This thought experiment should not be viewed as a criticism of the Granger causality approach that Gow and his colleagues have been pioneering in the speech domain. On the contrary, this approach offers a potentially powerful new tool in our arsenal for teasing apart the operation of the speech system, and I believe that further work of this type will prove to be extremely illuminating.

References

Borsky, S., Tuller, B., & Shapiro, L. P. (1998). ‘‘How to milk a coat:’’ The effects of semantic and acoustic information on phoneme categorization. *Journal of the Acoustical Society of America, 103*(5), 2670-2676.

Connine, C. M. (1987). Constraints on interactive processes in auditory word recognition: The role of sentence context. *Journal of Memory and Language*, *26*(2), 527-538.

Connine, C.M., Blasko, D.G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints. *Journal of Memory and Language*, *30*, 234-250.

Connine, C. M., & Clifton, C., Jr. (1987). Interactive use of lexical information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 13*(3), 291-299.

Ganong, W.F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 110-125.

Gow, D. W., & Caplan, D. N. (2012). New levels of language processing complexity and organization revealed by granger causation. *Frontiers in psychology, 3*, 506.

Gow, D. W., Segawa, J. A., Ahlfors, S. P., & Lin, F.-H. (2008). Lexical influences on speech perception: A Granger causality analysis of MEG and EEG source estimates. *NeuroImage, 43*(3), 614-623.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review, 87*, 1-51.

Marslen-Wilson, W.D., & Tyler, L.K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*, 1-71.

Marslen-Wilson, W.D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*, 29-63.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology, 21*(3), 398-421.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1-86.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *The Behavioral and brain sciences*, *23*(3), 299-325; discussion 325-270.

Poeppel, D., & Monahan, P.J. (2011). Feedforward and feedback in speech perception: Revisiting analysis by synthesis. *Language and Cognitive Processes*, *26(7)*, 935-951.

Samuel, A.G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110*, 474-494.

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science, 12*, 348-351.

Stevens, K. N. (1960). Toward a model for speech recognition. *Journal of the Acoustical Society of America*, *32(1)*, 47-55.

Swinney, D.A. (1979). Lexical access during sentence comprehension: (Re)consideration of sentence context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*, 645-659.

Tyler, L.K. (1984). The structure of the initial cohort: Evidence from gating. *Perception & Psychophysics*, *36*, 417-427.

Tyler, L.K., & Wessels, J. (1983). Quantifying contextual contributions to word recognition processes. *Perception & Psychophysics*, *34*, 409-420.

Warren, R.M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*, 392-393.

Funding

Support provided by Ministerio de Ciencia E Innovacion, Grant PSI2014-53277.