In: Proceedings of AMLaP 2004.

Extending Wundt's Principle for incremental processing

Markus Guhe
Cognitive Science Department, Rensselaer Polytechnic Institute

Language production is a complex task, and despite limited resources, humans can speak fluently at a constant rate. The means to achieve this performance is incremental processing; i.e., utterances and sequences of utterances are not planned completely in advance but are generated on the fly. Particularly suited for incremental processing are cascaded architectures. They consist of processes that are arranged in a fixed sequence, so that the output of one process is the input to its successor. This parallel processing of a sequential information stream is, together with the accompanying strict unidirectionality of the information flow, characteristic of incremental processing.

As a principle for cognitively adequate incremental processing Levelt (1989: 26) proposes Wundt's Principle: *Each processing component will be triggered into activity by a minimal amount of its characteristic input.* (Input/output is characteristic if a process can operate on the information, e.g. a phonological process cannot use syntactic information.) I propose the following extension: *Each processing component will be triggered into activity by a minimal amount of its characteristic input and produces characteristic output as soon as a minimal amount of output is available.* (Guhe 2003: 82) This extension is in contrast to other principles put forward in the literature. For example, it contradicts principles that reason about the 'when-to-say'. In the context of a language production system that generates self-repairs Kilger & Finkler (1995: 9), for example, state: 'In order to both fulfill the time constraints and avoid overt repair incremental output production should be used to utter succeeding parts of the sentence [i.e. after the initial output increment] as soon as necessary [...] rather than uttering them as soon as possible.' This means, for improving the quality of the output and for reducing the number of self-repairs the output is not generated immediately but only after a component-internal quality assessment. Levelt's original formulation affords both possibilities.

With respect to computational complexity this difference is important. Extended Wundt's Principle has the advantage that processing is fast and requires few resources, e.g., no resources are spent on reasoning about 'when-to-say'. There is various empirical support for the principle. For example, Bock et al. (2003) show in an eye-tracking study that the content to be verbalised is selected early and the utterance then follows the established pattern. Repp & Sommer (2003) show that in generating elliptical expressions, decisions about the generated constituents are made rapidly. INC (the *incremental conceptualiser*, Guhe 2003) is a model of the first component in Levelt's (1989) language production model that operates according to the proposed processing mechanisms. INC generates preverbal messages (semantic structures) for observed motion events. Comparing INC's output with recorded verbalisations of the same scenes supports the cognitive adequacy of Extended Wundt's Principle.

Although Extended Wundt's Principle accounts for fast output by basing the output on partial information, early-generated outputs must often be changed in light of subsequent inputs. Furthermore, some of these changes cannot be avoided principally, cf. Guhe & Schilder 2002. I propose two mechanisms to alleviate this problem:

(1) specialised *increment buffers* that temporarily store results of incremental processes (increments) so that the writing process can still access these increments and is able to change them
(2) *indirect feedback*, which is feedback that is not realised as direct transmission of information from a process to a preceding one in the cascade, but the process giving feedback alters a memory that it shares with the process receiving feedback. Such a shared memory can be one of the specialised buffers or a general memory structure like working memory. A modification of the later process affecting operations of the earlier process is indirect feedback. Indirect feedback has the advantage that the earlier process does not need to evaluate the feedback explicitly. This means, no resources are spent on processing the information that would be given by a direct feedback. Processing indirect feedback is simply part of the standard way of processing the information present in a shared memory.

Both mechanisms preserve the strict unidirectionality of the information flow, which is desirable, because mechanisms like direct feedback or revision-based processing are more costly in terms of resources.

References

Bock, Kathryn, David E. Irwin, Douglas J. Davidson, & Willem J.M. Levelt (2003) Minding the Clock. *Journal of Memory and Language* 48(4): 653--685.

Guhe, Markus (2003) *Incremental processing for language production*. Doctoral Dissertation, Department of Informatics, University of Hamburg.

Guhe, Markus & Frank Schilder (2002) Incremental generation of self-corrections using underspecification. In: M Theune, A Nijholt & H Hondorp (eds.) *Computational Linguistics in the Netherlands 2001. Selected papers from the twelfth CLIN meeting*, 118–132. Amsterdam: Rodopi.

Kilger, Anne & Wolfgang Finkler, 1995, *Incremental Generation for Real-Time Applications*. Technical Report RR–95–11, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI).

Levelt, Willem J.M. (1989) Speaking: *From intention to articulation*. Cambrigde, MA: MIT Press.

Repp, Sophie & Kathrin Sommer (2003) Verb coding in complex sentences: Evidence from eye-movements in the production of coordination with and without ellipsis. In: *Proceedings of AMLaP–2003: 9th Annual Conference on Architectures and Mechanisms for Language Processing*, August 25–27, 2003, Glasgow, Scotland, p. 137.