First Conference on Computational Psycholinguistics August 11, 1997
revised August 15, 1997; converted to pdf June 2005

Inferring Processing Times from a Corpus of Conversations

Nigel Ward[1]
University of Tokyo

1 Motivation

One of the ultimate goals of psycholinguistics is to know what goes on in the mind of a language user, including what all the processes are, how they fit together, and how they operate in everyday language use. Psycholinguistics research today, however, mostly focuses one or another specific aspect of language processing, more or less in isolation, in controlled, unnatural tasks.

This paper proposes a new research technique that addresses these problems.

2 The Proposed Technique

Measuring processing times requires, of course, the ability to determine both the start and end point of processing. The end point is simply the point at which the output of the response begins. The difficulty is determining the start point.

In controlled experiments this start point is the point at which the stimulus is given, or at which the "go" signal is given. In free conversation things are of course less simple, but it turns out that, for one special case, there is a kind of "go" signal in conversation.

This is the cue which the speaker gives when he wants to receive back-channel feedback, also sometimes called listener responses. These are, roughly, the expressions which a hearer produces during the speaker's turn. In English, typical expressions used in back-channel feedback include *yeah*, *uh-huh*, *hm*, *right* and *okay*.

In both Japanese and English this cue, or at least one of the cues, is a region of low pitch (Ward 1996; Ward 1997). The current best characterization of the cue is given as a predictive rule in Figure 1.

For Japanese this cue is quite strong; in our corpus well over half the occurrences of back-channel feedback are preceded by a low pitch cue by the speaker. Moreover, in most cases it seems that the speaker produces the low pitch cue immediately after having expressed some noteworthy information. Therefore we can use the location of this cue as our start point for measuring processing times; believing that it marks the "go" point, that is, the point at which the listener commits to producing an utterance, and also believing that it approximately marks the point at which the listener has received sufficient information to choose his utterance.

[1]Mech-Info Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113 Japan; nigel@sanpo.t.u-tokyo.ac.jp; http://www.sanpo.t.u-tokyo.ac.jp/~nigel/

3 Some Observations

First, let us note that the production of back-channel feedback in conversation appears to be very fast: on the order of 350 milliseconds (P5 of Figure 1). This is faster than typical response times in, for example, picture naming tasks. In any case, it seems that the demands of being an active and cooperative listener provide considerable time pressure.

Table 1 indicates the distribution of processing times associated with some common back-channel responses. For each row the grunt listed is the prototypical elements of a phonetic and semantic category. (Note that these grunts are not lexical items, in that their phonological content is not fixed, but seems to vary directly with the content conveyed, we believe.) For example, the row labeled *aa* includes grunts transcribed as *a*, *aa*, *aa-a*, *aaaaaaa*, *ahhh*, *a–un*, *aa-aaaa-aaaa-a*, *hha*, *haan*, *ha-heeee* and so on, but not *aa-soo-desu-ka*. The time numbers shown are the times between the cue and the observed onset of back-channel feedback, as computed from our corpus of 80 minutes of Japanese conversations. These averages are computed using only those instances which were correctly predicted by the rule, where a correct prediction is one where the onset predicted by the rule is within 350 milliseconds of the onset of the observed back-channel feedback.

With reference to the time for *un*, the most common and the most neutral form of back-channel, and also with reference to the average time for grunts, Table 1 suggests that responses indicating incomplete understanding or agreement (the *aa* group) take relatively longer, those indicating sur-

Upon detection of	
a region of pitch less than the 28th-percentile pitch level and	(P1)
continuing for at least 110ms,	(P2)
coming after at least 700ms of speech,	(P3)
providing you have not output back-channel feedback within the preceding 1.0 seconds,	(P4)
350ms later you should produce back-channel feedback.	(P5)

Figure 1: Model of the Decision of When to Produce Back-channel Feedback

response category		time	instances	
sound	meaning	(ms.)	used	total
other grunt	various	290	90	207
un	neutral	295	82	180
mm	comtemplative	321	11	32
oo	receipt of new information	341	9	19
aa	delay, filler	364	27	64
hee	surprise	375	12	20
unn	agreement	395	13	31

Table 1: Average Processing Times for some Common Categories of Back-channel Grunts

prise (the *hee* group) take even longer, and those indicating deep or thoughtful agreement (the *unn* group) take even longer. This is also seen in Figure 2.

Table 2 suggests how processing time may depend on the pragmatic force of the back-channel feedback. (Note that everything in Table 1 fits into the "grunts" category here.) One necessary caveat is that the average time for "echo" (feedback in the form of echoing back one of the speakers words) may be misleading since many such cases appear to be cued by another pitch factor, one that precedes the low pitch regions. Another caveat is that all the data is raw, uncompensated for the probable existence of word length effects or word frequency effects.

Of course, all these numbers are very preliminary. The basic problem is that there is too little data to draw any conclusions with confidence. A second problem is that the averages are computed over a corpus of many speakers, but the distribution of feedback types is not uniform, and neither is the distribution of delays.

feedback type	time	instances	
	(ms.)	used	total
Emotion/Sympathy	292	6	19
Echos	298	22	39
Laughs	306	26	58
Grunts	313	244	553
Explicit Agreement	322	55	120
Compound	391	21	41
Other	408	12	29
Inference	446	5	12
Questions	530	2	2
TOTAL	322	393	873

Table 2: Average Processing Times for Types of Back-channel Feedback

4 Problems with This Technique

The primary problem with this technique is that all the assumptions in the last paragraph of Section 2 are critical yet unproven. At least three lines of work are needed: experimental work to establish and quantify the significance the low pitch cue, corpus work to clarify the relation between the low pitch cue and the point of new information conveyed, and confirmation that this technique gives results which are compatible with those given by other methods for measuring processing times.

The unavoidable weakness of the technique is the amount of data required. Fortunately raw conversation data is now becoming readily available (in the form of the Call Home corpora), and labeling is not that onerous (with proper tools, such as "Didi",

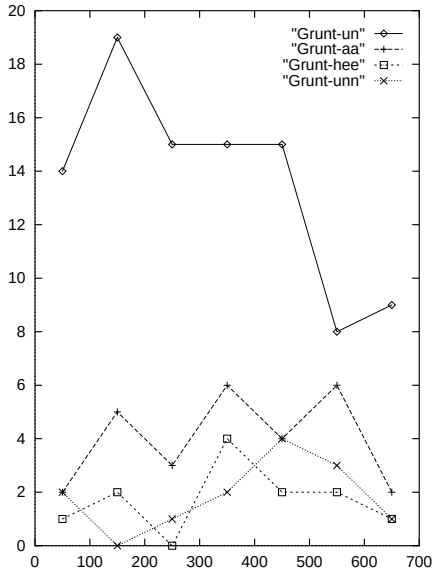

Figure 2: Distribution of Processing Times for *un*, *aa*, *hee*, and *unn*

our home-grown "dialog display tool").

5 Related Techniques

An elaboration of the above technique would be to explore the relation between timing phenomena and the speaker's and hearer's mental states. This should be possible with the use of subjective judgements, since humans are highly skilled at drawing inferences from exactly what someone says and from how (including the timing of when) he says it. One could leverage such judgements — regarding amount of attention paid, psychological distance between listener and speaker, and so on — with analysis of timing phenomena, using statistical methods. This could bear on questions of how attention affects response timing, and whether there are trade-offs between the amount of attention paid to prosodic cues vs. word recognition vs. deeper processing; or trade-offs between listening and responding at an emotional level vs. at the information level.

One could also use subjective judgements for evaluating hypotheses. For example, we could build a system that embodies rules like "produce echos of keywords 80 milliseconds early", use it to synthesize conversational responses, and have humans judge the results, perhaps on dimensions like "seems odd" or "seems normal"; or maybe even "seems depressed", "seems a little pushy", "seems thoughtful", "seems friendly", and so on, since all of these probably correlate somewhat with the details of response timing. This could be done in two ways. One possibility is with on-line experiments, with subjects conversing in real-time with a system that produces back-channel feedback. However, in pilot studies we have found that subjects in such conditions are generally not sensitive to even gross variations in response timing. A more promising possibility would be to use third-party judges, listening to the resulting conversation either live or on tape.

A related technique would be to examine back-channel responses which are suboptimal, as judged by the analyst or by the speaker himself, subsequently listening to his performance on tape with the benefit of hindsight and unlimited time. Such cases can probably be correlated with, among other things, conditions where time pressure prevents all processes from contributing. This idea is, in effect, an generalization of the techniques of speech error analysis.

6 Hopes

The potential strength of this technique is that it allows quantitative analysis of the whole language user, as he participates in normal, situated language use.

References

Ward, Nigel (1996). Using Prosodic Clues to Decide When to Produce Back-channel Utterances. In *International Conference on Spoken Language Processing*, pp. 1728–1731.

Ward, Nigel (1997). A Simple Rule for the Cooperative Timing of Utterances in Spoken Dialog. In *IJCAI-97 Workshop on Collaboration, Cooperation and Conflict in Dialogue Systems*, pp. 85–90.