Variations along the Contextual Continuum in Task-Oriented Speech

Gregory S. Aist (gaist@cs.rochester.edu)
Computer Science, University of Rochester, RC 270226
Rochester, NY 14627 USA

Ellen Campana (ecampana@bcs.rochester.edu)
Brain & Cognitive Sciences and Computer Science, University of Rochester
Rochester NY 14627 USA

James Allen (james@cs.rochester.edu)
Computer Science, University of Rochester, RC 270226
Rochester, NY 14627 USA

Mike Rotondo (mrotondo@cs.rochester.edu)
Computer Science, University of Rochester, RC 270226
Rochester, NY 14627 USA

Mary Swift (swift@cs.rochester.edu)
Computer Science, University of Rochester, RC 270226
Rochester, NY 14627 USA

Michael Tanenhaus (mtan@bcs.rochester.edu)
Brain & Cognitive Sciences, University of Rochester
Rochester NY 14627 USA

Abstract

Spontaneous speech, with its interactive nature and close ties to world and task context, is often viewed as less desirable than carefully planned utterances – filtered through (errorful) human performance limitations, harder to understand, and certainly more difficult for computerized analysis. This paper presents evidence that under certain conditions, interactive and context-specific spontaneous speech is not only an efficient and effective means of human-human communication, but is actually simpler than its carefully planned counterpart in a number of dimensions that are important to both human and automated processing.

Introduction

Spontaneous utterances, and those that rely highly on context, have a number of characteristics that at first blush render them problematic vs. carefully planned sentences. For example: pauses; elided nouns, verbs, etc.; and abandoned words and sentences.

For these and perhaps other reasons, some language scientists have historically viewed spontaneous utterances as second-class citizens – in some linguistic circles, as a version of grammatically correct utterances but distorted by human performance limitations, or in some computational circles as a fragmentary framework with holes to be filled in by computational approximations in order to reconstruct the original communicative intention. Here is a good description of this sometimes-called 'idealist' view:

> What is actually said and communicated between people is said to be the product of 'language performance', which is governed by many other factors besides the linguistic faculty, and is profoundly distorted by speaker errors of various kinds. The goal of linguistics is to get at the underlying 'competence' of the speaker, and the study of performance is said to lie outside of linguistics proper. (Labov 1985)

Labov goes on to describe (his own) countervailing view: "The materialist view is that 'competence' can only be understood through the study of 'performance'..." -- that is, that natural language use (such as spontaneous speech, with its disfluencies and context-dependencies) is the lens through which science necessarily views language. We would extend this position to argue that spontaneous speech has in fact various advantages for the speaker, the hearer, and the course of the conversation. These advantages may pertain to the speaker – for example, there is evidence that speech produced when contextual redundancy is high economizes on articulatory effort (e.g. Lindblom 1990). Or, they may relate to the conversation more generally. Bard, Anderson, Sotillo, Aylett, Doherty-Sneddon, and Newlands (2000) claim that variations in speaking style are a source of information -- for example the lack of clarity in a speaker's communication can indicate the difference between given or new information. For a concrete example, Sampson (1998) reports:

> Grammatical discontinuity is normally taken to be a performance deviation from the competence rules (of the standard language, or of a nonstandard dialect). Perhaps surprisingly, one not uncommon pattern in CHRISTINE data is that discontinuity is used intentionally to achieve a particular communicative effect. In:
>
> *and he takes the mickey out of him which okay then he called him ...*
>
> the most plausible interpretation of which okay has it saying, in effect, "I am not going to complete the relative clause I have initiated with which, but, were I to do so, that clause would amount to a concession of the issue just raised". (Sampson 1998).

This attention to language as it is actually used is accompanied by a deep concern for context. As one recent article pointed out, "As the notion of language proficiency has evolved into one of communicative competence, so has the understanding that performance, especially in spoken language, is influenced by context. No study of spoken language use or assessment of oral proficiency could ignore the interaction between context and performance." (Wiseman 2004).

We set out to define a domain where we could explore this relationship between context, spontaneous speech phenomena, and human-human dialogue. We view the resulting utterances as produced by a common langauge mechanism, but influenced by varying considerations when incorporating semantic and contextual constraints into the generation process.

Materials and Methods

We devised a task whose basic goal is to construct objects, place them on a map, and orient them to match a (paper) card showing the target object. This task consists of five parts.

The Task: Constructing objects and placing them on a map

Part 1: Choosing an object. We devised compound shapes where each shape consisted of only a few components, each component of the shape was easy to name, but the entire shape required a complex description rather than a prenominal modifier. For example, a square with stripes could also be referred to as "the stripey square", but a square with diamonds on the corner cannot be referred to as *"the corner-diamonded square". We thus chose a set of shapes such as "a small square with a diamond on the edge", "a large triangle with a star on the corner", "a small triangle with a circle on the edge", and so forth.

Part 2: Placing the object. For the location of the objects on the map, we wanted to have directions come in as several small clauses. We first devised a map that was a Manhattan-style street grid, with named avenues and streets, and had people place objects on the intersections. However, we were surprised at how complex peoples' descriptions of intersections were - they described not only the two intersecting streets but also the corner (e.g. Southwest) and sometimes the approximate distance from the intersection and other nearby streets or landmarks as well! But we wanted to have people give directions not in terms of complex descriptions ("on the southwest corner of Fifth and Broadway, across from Pat's Deli") but rather in terms of several small clauses such as regions ("in Central Park"), landmarks ("north of the flagpole") and relative descriptions ("up a little bit"). We thus designed a map with several named regions and a couple of easily-recognizable landmarks (flags), and placed the objects within the regions (not on boundaries), on the flags, or near the flags.

Part 3: Painting the object. Objects can be painted in one of several colors. This was included to provide a uniformly easy subtask, for use as a control condition in experiments conducted in this domain.

Part 4: Rotating the object. For the rotation of objects on the map, we wanted to have both specific angle specifications available ("left forty five degrees") as well as a more interactive style ("right a bit more.. okay stop"). Thus the rotation tool we provided simply turned the objects right a bit or left a bit, with continued activation needed to turn the object more. (We did not provide a compass rose nor allow the person using the mouse to specify a particular angle to rotate.)

Part 5: Filling an object. In order to construct a complete object, the user has to specify its contents. We chose the analogy of a set of vendors' carts (e.g. Coney Island hot dog stands) that you have to fill with food and place on a map. So, the fillings of the carts were various kinds of food such as avocados, banans, cucumbers, grapefruit, and tomatos. We chose fruit since they were nameable (especially with a label printed on the screen) and quite different visually from the carts themselves (square, triangle, etc.) A cart could be filled with zero or more fruit, and on each display there could be several carts filled with various fruit. Constructing each cart was however an independent task, which meant that errors made early in the dialogue will have only limited impact on the construction of subsequent carts.

Figure 1 shows an example of the screen after placing five (5) objects on it.

Data Collection: Speech, Gaze, and Mouse

During the data collection, one person (the director) gave directions to the other person (the actor) on how to place objects on the map, move them, and so forth. In order to collect speech data, the director wore a headset microphone. In order to collect eye gaze data, the actor (sometimes) wore a head-mounted eyetracker. In order to collect mouse actions, the display software kept detailed logs, synchronized precisely to the timing of both the speech and the eye movement data. In the data described in this paper, the subject was the director and a confederate was the actor.

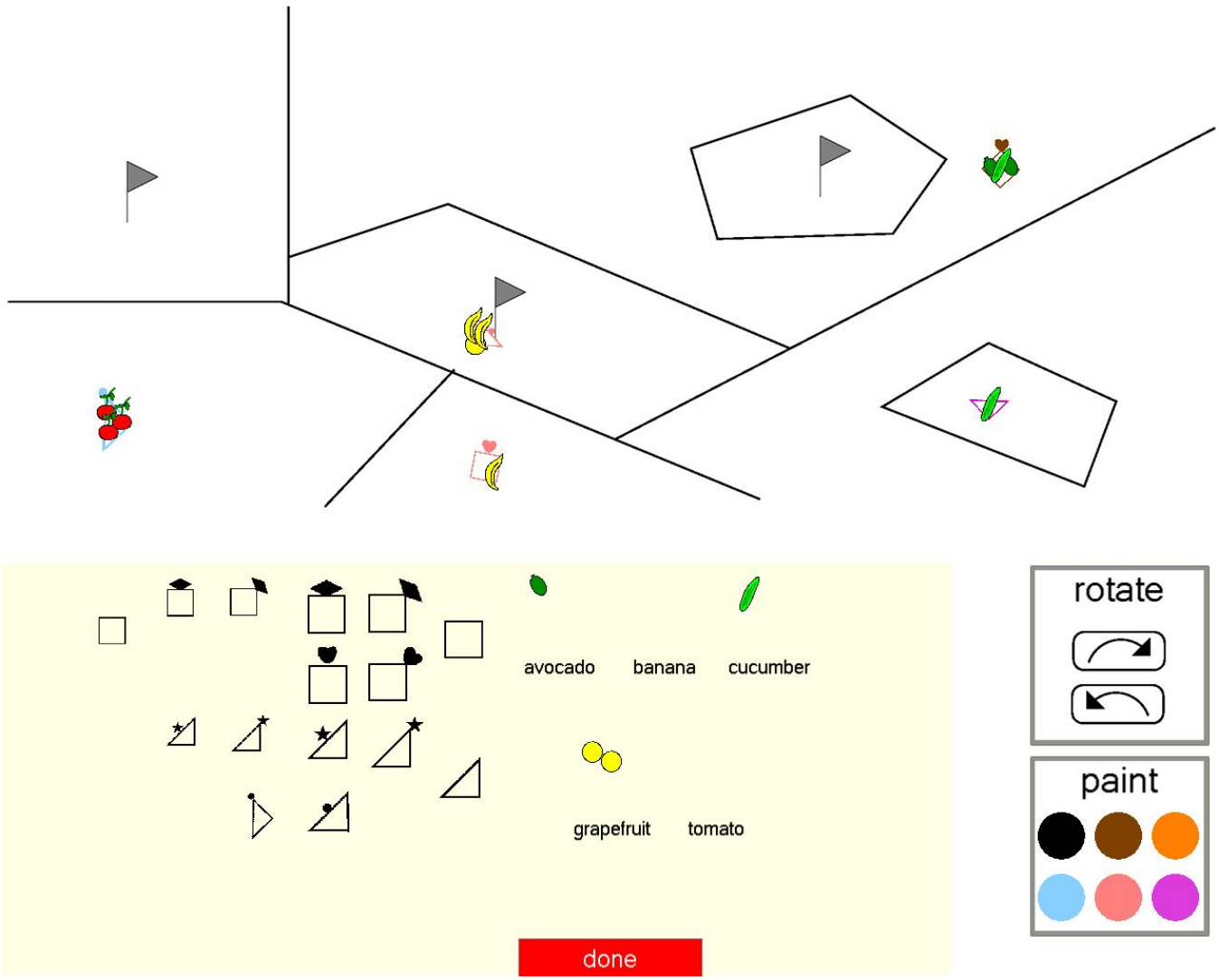

Figure 1: Fruit carts screen after placing five objects.

We collected data in a dedicated room, with the actor wearing the eyetracker and the director sitting just behind and also looking at the screen. A dozen subjects participated, each of whom specified twenty objects to place on the map; thus, a total of 240 dialogues were collected. Each object had five attributes – shape, tag, tag place, location on the map, and angle – with those attributes being either "prototypical" (e.g. Forty five degrees left; in the center of a region) or "atypical" (e.g. Twenty six degrees right; two centimeters below and one centimeter to the left of the flag in Central Park.) We intended this mix of attributes to elicit a variety of language use – everything from carefully planned utterances such as "Put a large square on the flag in Central Park" at one extreme, to interactive mayhem such as "That needs to go left of there... no, more above the flag but to the left a bi- okay that's good stop right there." The resulting recordings were then transcribed word for word.

A note on ambiguity

It is still the case that at first look, interpreting utterances continuously leads to excessive spurious ambiguity. For example, in the (incomplete) sentence "Now right by...", *right* could mean either *close* or *west* or *clockwise* – but the remainder of the sentence will typically resolve this passing amibiguity, e.g. "... a couple of degrees." However, it is exactly here that dialogue and task context can be brought to bear to eliminate the ambiguity (even in passing.)

1a: Turn that one to the left about ninety degrees.
2a: Now right by... {Dialogue context yields *clockwise*
since we just talked about turning.}
... a couple of degrees

1b: We'll start in Central Park.
2b: Now right by... {Task context yields *close*
since there's no active object
to rotate or move.}
... the flag we need a large um square.

Data Labeling

The transcripts were then labeled with two categories. The first was prototypical "all-at-once" utterances. The second was prototypical "continuous" utterances. Utterances that were somewhere on the continuum between these two extremes were labeled as "both"; utterances that didn't seem to be categorizable along this dimension were labeled as "neither".

Utterances labeled as "all-at-once" were those whose semantics could be grounded independently of the surrounding dialogue context, up to pronominal reference. For example, "Move a large plain square to the flag in Central Park" has a fully specified verb (*move* a shape to a location), object (a *shape* unique within the domain), and adjunct (a particular *flag* in a specific region of the screen). Continuous utterances on the other hand rely on the surrounding context -- dialogue and/or task -- for their grounding. For example, "up a bit more" contains a direction (up), but might rely on the last action to identify the intended verb, and on the selected shape on the screen to identify the direct object.

Overall, the "continuous" utterances were about half as common as the "all-at-once" utterances (Figure 2), although there was – as expected – variation among subjects. We now turn to an analysis of the various contrasts between continuous and all-at-once utterances.

Results

We begin with some general observations about the nature of the data collected. Then we'll look at two areas. The first is the contrast between continuous and all-at-once utterances. The second is how usage of these strategies varied among the dialogues.

Some Striking Phenomena

The design of this corpus, and our initial analysis of these data, yield some points of interest to mention.

1. End-of-sentence boundaries tend to be fairly clear (at least to a human listener). Where a sentence begins, however, is quite difficult to say precisely, due to disfluencies, abandoned utterances, and so forth. This is in contrast to domains where speakers might tend to begin a sentence clearly, such as information retrieval ("Search for books by Kurt Vonnegut").
2. There seem to be two distinct strategies that people can employ: saying a direction all at once ("Put it one inch below the flag") or continuously ("Put it near the flag [pause] but down a bit [pause] a bit more [pause] stop.") We suspect that the harder the task is, the more likely people are to employ the continuous strategy. (The corpus is designed and balanced to make testing that hypothesis feasible – the subject of future work)
3. Besides a pure All-at-once and Continuous strategy, people sometimes switch between them, employing Both. For example, the director might tell the actor to place an object "right on the flag [pause] down a bit [pause] keep going [pause] stop." We see these as possibilities along a continuum, using the same language mechanisms yet according different emphasis to the strategies.

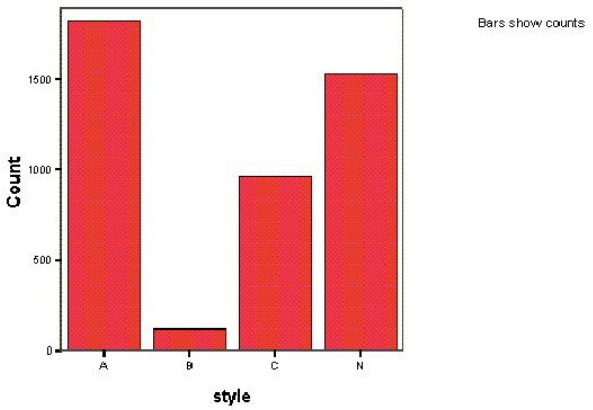

Figure 2: Frequency of style.
A=all-at-once, B=both, C=continuous, N=neither.

Continuous Utterances are Shorter and Use Shorter Words (than All-at-once Utterances)

We first characterize the differences between the prototypical "continuous" utterances and the prototypical "all at once" utterances.

What types of sentences occur? We next present an extended example of dialogue that includes all-at-once (in **bold**), continuous (in underlining), both **(bold** and underlined**),** and neither (plain) utterances.

- - - - - - - - -

take the uh large triangle with the star
and um
put that
um to the
right
um
side of the uh
flag in pine tree mountain
er the right side
and
<laughter> um down a little
um
then rotate it so that
the
the hypotenuse is
almost
horizontal but
tilted a little sli- like one more rotat- yeah
and um make that orange
um maybe a little closer to the flag
and down
yeah that should be good

- - - - - - - - -

As this example shows, people were naturally fluent in using (and blending) both styles of interactions.

Quantitatively speaking, there were differences in sentence length among the four categories of utterances (Figure 3). Analysis of variance with sentence length as the dependent variable and style & subject as independent variables yielded a mean of 8.72 +/- 0.12 words for all-at-once vs. 6.85 +/- 0.18 words for continuous utterances (F=313.679,df=3,p<0.001),a difference of 1.87, or 21%.

What types of words occur? Certainly some words occur more often in one type vs. another type of utterance. For example, *angle* occurred 19 times in the "all-at-once" style, and only once in the "continuous" style. On the other hand, *back* occurred only twice in the "all-at-once" style, but 33 times in the "continuous" utterances. One curious case is *that* vs. *that's:*

	All-at-once	Continuous
that	366	123
that's	48	116

-- of particular interest since occasionally *that's* and *that is* are conflated in studies of language.

Quantitatively, we examined word length as it related to the style of the utterance. As Figure 4 shows, there was a significant difference between the wordlength for the all-at-once conditions vs. that in the continuous conditions. Analysis of variance with word length as the dependent variable and style & subject as independent variables yielded a mean of 3.95 +/- 0.02 letters for all-at-once vs. 3.74 +/- 0.03 letters for continuous utterances (F=37.35,df=1,p<0.001) - a difference of 0.2 letters, or 5%.

The Use of Continuous Style Increases over Time

We wanted to see if there were variations in usage over time. In particular, would the relative frequencies of all-at-once vs. continuous change as subjects gathered more practice in the task? Figure 5 shows the relative percentages. A clear trend is evident – as the dialogues progressed, the use of the "continuous" style become more and more common. Logistic regression with style={all-at-once|continuous} as the outcome variable, subject as a categorical variable, and trial as a numeric variable showed that trial was indeed a significant predictor of style (B=0.104 +/- 0.037, exp(B) = ~1.11, p < 0.01). This is a particularly interesting result since it shows that not only does the frequency of the various styles change over time, but in fact the continuous strategy increases in use over time as the experiment progressed.

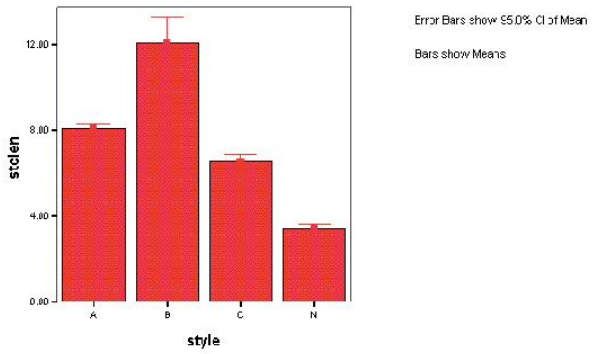

Figure 3: Sentence length by style.

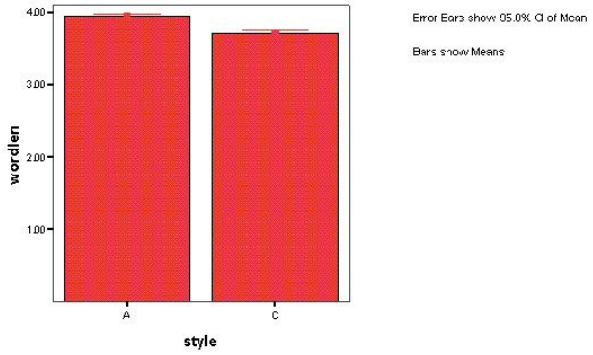

Figure 4: Word length by style.

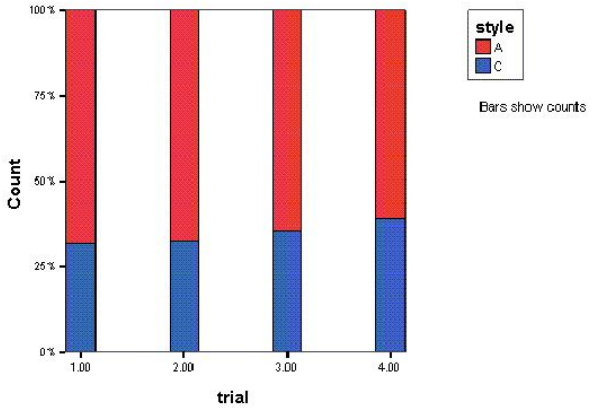

Figure 5: Relative percentage of styles all-at-once and continuous over time.

Discussion

A number of factors in this domain – and, we propose, in many domains – contributed to eliciting highly interactive speech. The first was continuous variation among parameters. For example, the angle of rotation was allowed to vary in non-45 degree increments, thus giving rise to opportunities for fine adjustments to the angle of an object. By contrast, Clark and Krych (2004) describe a construction task in which the elements are Legos – which have discrete attachment points. The second was that the actors were allowed to begin moving while the director was still speaking. This allowed the "control loop" to be closed – the director could monitor the actor's movements in real time and adjust his or her production of commands accordingly. For future work, the fact that the corpus includes mouse movement should enable us to quantify when people have enough information to begin an action. Furthermore, the eye gaze data should allow us to look at when people have enough information to disambiguate among possible actions - even before they begin to undertake an action.

What are the implications of the shift in strategy use over time? Other changes in speaker behavior over time have previously been reported; a prime example is timing changes. Bull and Aylett (1998) report that inter-speaker interval was longer in subjects' first 2 dialogues than in their next six. Mostow and Aist (1997) describe how the time between words in a reading-aloud task decreases with repeated encounters with words. One interesting possibility is that shifts in strategy – alone or with other observations such as time between sentences or words – might provide evidence of adaptation to (or learning of) the task. That would be a useful capability in both experimental explorations – to measure learning as affected by various manipulations of other variables – as well as applied domains such as intelligent tutoring systems.

What does continuous understanding buy you? In short, continuous understanding allows rapid contextualization. This is clearly the case for human-human dialogue and (we hope) can be extended to human-machine dialogue as well.

In conclusion, we have described a multimodal corpus in a novel domain: object construction and placement using continuous variables. This corpus has not only speech data but mouse movements and eye movements as well. Its contents should prove useful to those interested in building recognition systems for disfluent or continuous speech, as well as those who wish to build generation systems that can give continuous (i.e. Incremental) directions. Futhermore, we have presented evidence that under certain conditions, interactive and context-specific spontaneous speech is not only an efficient and effective means of human-human communication, but is actually simpler than its carefully planned counterpart in a number of dimensions that are important to both human and automated processing – particularly sentence length and word length. Finally, people's use of a "continuous" strategy increases over time as they become familiar with the task.

Acknowledgments

This work was carried out at the University of Rochester under a grant from the National Science Foundation (James Allen, PI) together with Elizabeth Shriberg and Andreas Stolcke at SRI.

References

Bard, E.G., Anderson, A.H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language,* 42, 1-22.

Bull, M., & Aylett, M. (1998). An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. *Proceedings of the International Conference on Spoken Language Processing* (ICSLP), Sydney, Australia. www.cogsci.ed.ac.uk/~matthewa/publications/icslp98d.pdf

Clark, H.H. & Krych, M.A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language, 50*(1), 62-81.

Labov, W. (1987). Some observations on the foundation of linguistics. Unpublished manuscript. Philadelphia: University of Pennsylvania. www.ling.upenn.edu/~wlabov/Papers/Foundations.html

Lindblom, B. (1990). Explaining variation: a sketch of the H and H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403-439). Dordrecht: Kluwer.

Mostow, J., & Aist, G. (1997). The Sounds of Silence: Towards Automated Evaluation of Student Learning in a Reading Tutor that Listens. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)* (pp. 355-361). Providence, RI: American Association for Artificial Intelligence.

Sampson, G. (1998). Consistent annotation of speech-repair structures. In A. Rubio et al. (Eds.), *Proceedings of the First International Conference on Language Resources and Evaluation*, pp. 1279–82. Granada. http://www.grsampson.net/Acao.html

Wiseman. (2004). Review of D. Boxer & A.D. Cohen (Eds.), *Studying Speaking to Inform Second Language Learning*. Clevedon, England: Multilingual Matters. Teachers College at Columbia University, *Working Papers in TESOL & Applied Linguistics*, *4*(2). www.tc.columbia.edu/academic/tesol/Webjournal/WisemanReview.pdf