# HOW MANY WORDS IS A PICTURE REALLY WORTH?

*Laurel Fais, Kyung-ho Loken-Kim, and Tsuyoshi Morimoto*

ATR Interpreting Telecommunications Research Laboratories

## ABSTRACT

Subjects communicating in telephone and multimedia setting do not replace speech with visual images in the multimedia setting. Instead, they use more words in this environment. We discuss the trade-off between words and images, addressing this surprising result. Several factors are involved: use of redundant visual information; "meta-media" conversation; and a slightly greater amount of information conveyed in the multimedia setting. Suggestions concerning integration of a multimedia interface with automatic translation are made.

## 1. INTRODUCTION

The old adage about a picture being worth a thousand words captures two important concepts in the field of multimodal communication technology. The first is that certain modalities are more appropriate than others for conveying certain types of information. For example, it is generally thought that a visual image such as a map will convey locational information better than a verbal description.

This first concept plays a major role in a number of systems which automatically construct multimedia presentations [1, 2, 6]. These systems contain rules which match features of the information to be conveyed with the corresponding capabilities of the media available and select the most appropriate medium for conveying that particular information. These systems recognize that a number of media may be appropriate, but in most cases, they contain heuristics for choosing just one.

The second concept embodied in the adage concerns this choice. "One picture *is worth* a thousand words" implies that the picture should *replace* the use of words. This is the assumption behind choosing to represent, say, the location of a restaurant on a map, *instead of* describing it in words. The adage implies that, where you can use a picture, you don't need the words.

When we turn from multimedia presentation generation to the utilization of multimedia systems by actual users, we often assume that humans will operate in the same way. We think that if users have the *option* of presenting the location of a building on a map, they will use that option *instead of* presenting the information by typing or speaking the information, for example. We assume that, where appropriate, users will employ pictures instead of words.

This view is attractive in the field of natural language processing. If instead of using speech, users in fact do employ the non-speech options available in a language processing system integrated with various media options, that would reduce the amount of language processing necessary. This, in turn, might make language processing systems more effective.

The Environment for Multimodal Interaction (EMMI) designed and built at the Advanced Telecommunications Research Institute (Japan) is part of an effort to build such a system, in which various communication media are integrated with machine translation. Within EMMI, users can speak, draw on a map, type to a form, or type unrestricted messages in order to perform a direction-finding task and a hotel reservation task [5]. EMMI can accommodate same-language interaction or bilingual interpreted interaction with either human or "machine" interpretation.

Below we report on the results of three experiments conducted in EMMI to test our initial hypothesis: that the availability of non-speech media in a communication environment will reduce the amount of speech used, when compared to a speech-only (telephone) setting.

## 2. METHODS

In the first of the three experiments, subjects acting as "clients" got directions to a conference site by engaging in a cooperative dialogue with the "conference agents." In this first experiment, both "clients" and "agents" were native speakers of American English, and their interaction was human-human (mediated by various technologies; see below). In two further experiments, native American English-speaking "clients" interacted with Japanese-speaking "agents." In one of these experiments, speech was interpreted by human translators; in the other, by a simulated automatic machine translation system ("Wizard of Oz" style; we will refer to this condition below as the "machine-interpreted" condition or as "human-machine interaction"). In all three experiments, subjects interacted in two different communication settings: via a standard telephone, and via a computer-based, multimedia environment [EMMI] in which subjects could freely interact by voice, by typewritten text, by drawing on a visual image (a map), and by typing to a form [3, 4, 7]. The acoustic data for all three experiments were recorded on DAT tapes and transcribed; the visual data (drawing or typing by both agent and client) were recorded directly from the computer screens and noted on the speech transcriptions [8].

We made three measures of the linguistic data. First we counted the number of words in each conversation. Second, we identified and labeled conversation concerned with the mode of presentation itself (see below). Third, a task analysis of the "information units" conveyed in the conversations was made, and those units were counted for each conversation. Examples of such "information units" are: location of the client in Kyoto

Station; location of the bus stop; length of bus ride; amount of train fare; and so on.

## 3. RESULTS

If the use of non-speech media *replaces* speech, we should find coordinated changes in the amount of information conveyed by speech and that conveyed by visual means (i.e., drawing or typing) as illustrated in the hypothetical Figure 1; as the amount of visual information increases, the number of words should decrease.

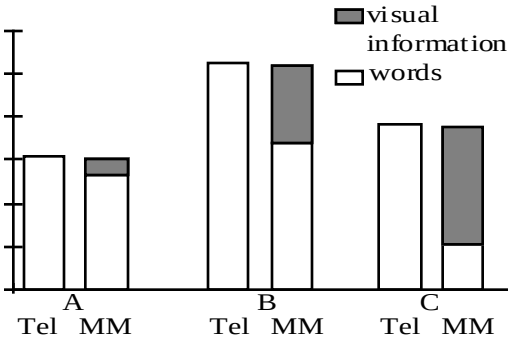


**Figure 1:** Hypothetical contributions of visual information and words in telephone and multimedia settings.

However, we found this not to be the case. A comparable graph reflecting the actual results of the experiments is shown in Figure 2.

Clearly, Figure 2 implies some notion of the worth of visual information relative to words. Natural language descriptors for some of the visuals (drawings or marks on the map) were constructed and the average number of words per descriptor was determined. This gave us a rough idea of how many words a drawing might replace; it turned out to be eight. However, note that this was done for illustration only.

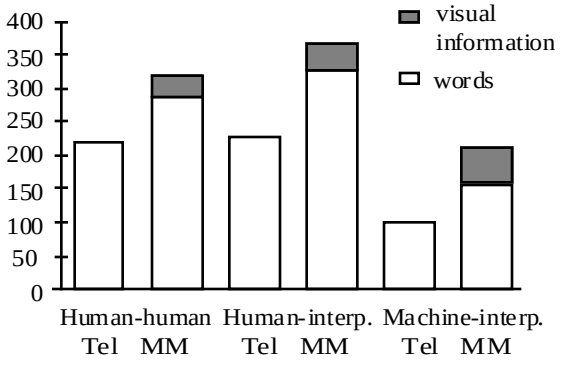


**Figure 2:** Actual contributions of visual information and words in telephone and multimedia settings.

The actual weight assigned to the visual information is irrelevant. The important point to note is this: in all three conditions, there was a significantly higher number of words in the *multimedia* setting than in the telephone setting. The use of visual information contributed to the conversation above and beyond this greater number of words. This is opposite to what our hypothesis suggests.

### 3.1 Meta-media conversation

How can we explain these results? A closer examination of the conversations in the multimedia setting revealed examples like these:

1a. A(gent): I'm circling the station...

1b. C(lient): I'm at the Kintetsu line. I'm putting a mark where I'm standing

2a. A: and now we'll show you where you're goinu go

2b. C: yes I was going to type in a message on the bottom

3a. A: and I can draw up a schematic of the bus station if you would like...

3b. C: OK should I tell you also or just type it

4a. A: the Shinkansen is here I circle it for you can you see

4b. C: can you see my location now I'm by the Shinkansen concourse

In (1), the speakers talk about what they are currently doing with the media; in (2), they talk about what they will do next; in (3), they ask their partners what they should do next; in (4), they confirm their partners' understanding of what they have just done. Each of these examples includes meta-conversation specifically concerned with managing the media available. We surmised that it was the addition of such "meta-media" conversation like this that was responsible for the unexpected increase in the number of words in the multimedia setting. Meta-media conversation is virtually absent from telephone conversations, but is present in significant amounts in the multimedia setting (Figure 3).

We then eliminated the meta-media conversation from our evaluation of the words used in each setting. However, even when meta-media conversation was subtracted out, the multimedia setting still showed a higher number of words than the telephone setting. This difference is no longer statistically significant for the human-human experiment, but it is significant for the human-interpreted and machine-interpreted conditions (Figure 4).

This suggests that, while meta-media conversation accounts for the "extra" words in the human-human condition, some additional factors are at work in the human-interpreted and machine-interpreted conditions.

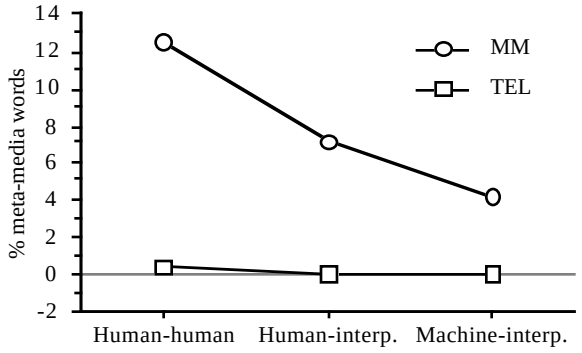


**Figure 3.**: Percent of meta-media conversation in telephone and multimedia settings for all three experiments.

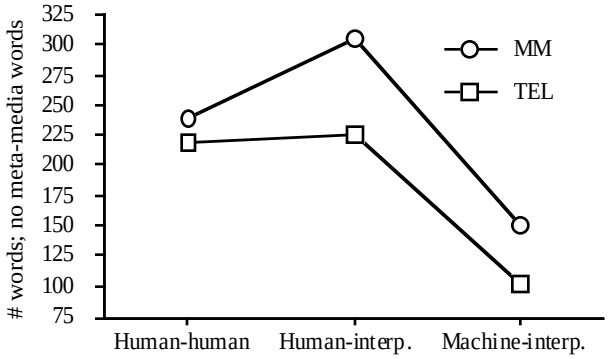


**Figure 4:** Number of words in telephone and multimedia settings for all three experiments, with meta-media words removed.

## 3.2 Information units

What if clients are requesting and receiving more information in the multimedia setting of these conditions? In fact, there does tend to be a higher number of information units in the multimedia setting for all three experiments (although this result is not significant). However, this number is not high enough to account for the greater number of words used in the MM setting; when we examine the number of words used per information unit, we still find a significantly higher number of words per information unit in the multimedia setting across all three experiments.

## 3.3 Meta-media and information units

So far, we have examined the effects of meta-media conversation and the words-per-information-unit separately. Finally, we analyze the joint effect of these two factors: we extract the meta-media conversation from the number of words, and then determine the number of words used per information unit. We find that the modal difference is no longer statistically significant. That is, if we ignore the meta-media conversation in the multimedia setting, the numbers of words used per information unit in both multimedia and telephone settings are equivalent (Figure 5).

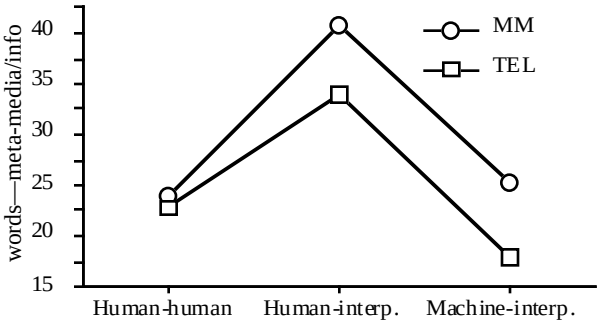


**Figure 5:** Words per information unit (with meta-media words subtracted out) for telephone and multimedia settings for all three experiments (differences are not significant).

# 4. DISCUSSION

Do we need to re-think the old adage? It would appear so. Our results reveal that subjects use an equivalent number of words to convey information whether they are communicating by telephone or via multimedia.[1] Subjects do not, in fact, use visual images to replace speech.

If we think about the everyday use of visual images, we realize that this is a reasonable result. It is rare that we allow images to *replace* words in everyday life. Newspaper, magazine, and book illustrations are invariably accompanied by captions; grandparents displaying pictures of their grandchildren never allow the picture alone to carry the message.

These anecdotal observations are supported by experimental evidence. In analyzing the visual information in these experiments, we found that it was presented in two ways. In the first, visuals accompanied deictic expressions in speech, and conveyed information about those expressions in a visual *instead of* verbal fashion. These accounted for only half the visuals used, however. The other half accompanied speech that contained no deictic expression. These were simply visual correlates of the information that was *also* being expressed verbally. Thus, approximately half the time, the use of visuals conformed to the old adage; the other half of the time, it did not.

There were some revealing differences in use of visual information across interpreting conditions. In the human-human interaction, the presence of additional meta-media conversation alone accounted for the greater number of words in

[1]This result is achieved only by ignoring the meta-media conversation that typically accompanies dialogue in the multimedia setting. See Section 5, below.

the multimedia setting. On the other hand, words-per-information-unit was significantly higher for the multimedia setting than for the telephone setting.

Neither meta-media conversation alone nor calculating words-per-information-unit was enough to explain the greater number of words in the multimedia setting of the two interpreted conditions. However, ignoring the presence of meta-media words, the additional amount of information conveyed in the multimedia setting of these conditions did account for the greater number of words.

In addition, the use of visual information increased in both interpreted conditions (Figure 2). These results are reflected in the post-experiment interviews. Subjects uniformly reacted in a positive way to the multimedia setting [3, 7]. They cited the presence of the map and seeing directions marked on the map as having a positive influence on their ability to understand and enjoy the task. Subjects' greater confidence in and enjoyment of the multimedia setting is probably correlated with the increased amount of information and use of visual information in that setting.

Thus, despite the fact that the presence of the visual channel seems to have had little effect on reducing the amount of speech and thus the processing burden on an automatic speech processing system, it is still a worthwhile addition to such a system by virtue of the benefits it offers to the understanding and enjoyment of users, and to their ability to convey more information.

## 5. FUTURE DIRECTIONS

Of course, we still have not dealt adequately with the phenomenon of meta-media conversation. Meta-media conversation is, in fact, an integral part of communication in a multimedia setting. Consider everyday experience with multimedia devices. The first thing most speakers do when speaking at a microphone, for example, is to make some comment on the medium itself: "Is this working?" or even, "I don't like to use mikes, but..." This corresponds to the meta-media conversation we found in the multimedia setting of these experiments.

Despite its naturalness, the presence of meta-media conversation represents an extra burden on an automatic language processing system. Our future work will focus on reducing meta-conversation in the multimedia setting by encouraging subjects to use non-speech options more effectively. We have already initiated improvements that will make the interface more system-driven, including more on-line instruction to increase users' confidence in the system and in their abilities to operate the system.

Furthermore, there is also the possibility that experience plays a role in the effective use of non-speech options [3]. Perhaps experienced users of multimedia will not show such a high rate of meta-media conversation. This is another area of future research.

Finally, we would like to explore the dialogue function of visual information that is redundant to the speech it accompanies. Walker [9] proposes that redundant *statements* in dialogue strengthen the underlying assumptions that must be shared by conversants in order for them to establish a set of mutual beliefs. Redundant visual information may serve the same function; if so, then it is not truly redundant but necessary to the process of establishing mutual beliefs. This leads us to conclude that the function of visual information may not be to replace speech, but to supplement speech in the essential aspects of the mutual construction of dialogue .

## 6. REFERENCES


1. André, E. and T. Rist. 1993. The design of illustrated documents as a planning task. In *Intelligent Multimedia Interfaces*, M. Maybury, ed., pp. 94-116. Menlo Park, CA: AAAI Press/MIT Press.

2. Arens, Y., E. Hovy, and M. Vossers. 1993. On the knowledge underlying multimedia presentations. In *Intelligent Multimedia Interfaces*, M. Maybury, ed., pp. 280-306. Menlo Park, CA: AAAI Press/MIT Press.

3. Fais, L. 1994. Effects of communicative mode on spontaneous English speech. Technical Report of the IEICE, **NLC94-22,**(1994-10), Tokyo, Japan, pp. 1-8.

4. Fais, L., Loken-Kim, K.H., and Park, Y.D. 1995. Speakers' Responses to Requests for Repetition in a Multimedia Cooperative Dialogue, in Pro., International Conference on Cooperative Multimodal Communication, (Eindhoven, The Netherlands, May 24-26, 1995), pp. 129-144.

5. Loken-Kim, K.H., Yato, F., Kurihara, K., Fais, L., and Furukawa, R. 1993. EMMI - ATR environment for multi-modal interactions. ATR Technical Report TR-IT-0018, ATR ITL, Kyoto, Japan.

6. McCaffery, F., M. McTear, and M. Murphy. 1995. Designing a multimedia interface for operators assembling circuit boards. In Pro., International Conference on Cooperative Multimodal Communication, (Eindhoven, The Netherlands, May 24-26, 1995), pp. 225-236.

7. Park, Y., K.H. Loken-Kim and L. Fais. 1994. An experiment for telephone versus multimedia multimodal interpretation: methods and subjects' behavior. ATR Technical Report TR-IT-0087, ATR ITL, Kyoto, Japan.

8. Park, Y., K.H. Loken-Kim, L. Fais, and S. Mizunashi. 1995. Analysis of gesture behavior in a multimedia interpreting experiment; human vs. Wizard of Oz interpretation method. ATR Technical Report TR-IT-0091, ATR ITL, Kyoto, Japan.

9. Walker, M. 1992. Redundancy in collaborative dialogue. In Proceedings of 14th COLING, pp.345-351.