TR-IT-0229

Everything You Always Wanted to Know About EMMI Experimentation*

*But didn't know who to ask

Laurel Fais
July, 1997

This report details the history of EMMI experimentation, and gives a general summary of the results for each experiment, and for each measure examined across all five experiments done in the EMMI interface. More importantly, it gives a detailed description of the location and nature of EMMI data, available for use in future telecommunication studies. It concludes with recommendations for re-thinking the multimedia interface issues faced by ITL.

Contents

1. The Motivations

1.1 Understanding EMMI

Research is a messy thing. Although in the papers that we write and the presentations that we make, we like to make it look as if we actually had a specific question, developed an experiment to find the answer and then found the answer, we rarely do. Instead, we have a hunch, run an experiment that we think will give us some clues, find that the results are not what we thought, pursue what the results seem to be telling us either by looking elsewhere in the results or running another experiment, and then discover a completely different phenomenon from the one we set out to investigate. Then we write a paper as if that were the question we wanted answered all along. Never mind all the floundering around we actually did in the middle while the ideas were coming clear.

In the course of the four or so years that I have worked with the EMMI (Environment for MultiModal Interaction) interface at ITL, I have written a number of fairly straightforward, question-experiment-answer sorts of papers; those are readily available and appear in the bibliography. I have done a good deal of floundering as well. The papers represent the distillation of points of interest that we have gleaned from our experimentation in EMMI. This Technical Report will chronicle the floundering, which, though much less straightforward and much harder to "sell," is also much more accurate and interesting.

So, if you are looking for a clear statement of the results from these experiments, this is probably not the best place to look. Instead, you should refer to ITL Technical Report #TR-IT-0172, "EMMI Progress Report: An Evaluation of Research Done with the First EMMI Interface," and a submission to the Journal of Natural Language Engineering, "When to put the 'face' in 'interface': Determining the effects of multimedia and a persona on communicative behavior" (available from the first author), as well as other various references cited throughout and listed in the bibliography.

In addition, this report will give something of the history of EMMI, and a chronicle of the experimentation done in that interface. It will not, however, give an engineering account of the design of the environment. That is not my area of expertise, and I am not qualified to write that history. So this report will focus on the *experimentation* done in the EMMI environment. I realize that this sort of thing is not in the area of expertise of many of the researchers at ITL; I hope that this report will help them to understand the motivations behind the kind of research that was done, as well as the advantages and the difficulties. Understanding the nature of experimentation and understanding specifically the background to the EMMI experimentation, is essential to true appreciation of the results obtained.

1.2 Making the data available

There is another equally important motivation for writing this report: to make available to any interested researcher information concerning the location and nature of the data from these experiments. A truism in our field is that there are vast differences between read and spontaneous speech, and between scripted and spontaneous dialogues (but if you need a reference, see Campbell, 1995). If our goal is truly to be able to process and translate natural speech, then it is critical for our work to be *based* upon natural speech. The collected EMMI dialogues form an invaluable resource for the understanding of natural speech, one which I hope researchers from many different areas will make use of.

So, where is all this wealth of data concerning human communicative behavior? Below we describe the location of the various transcription and other files which are available for use. Here we would like to stress the fact that they are "available." These collected conversations provide a rich source of information concerning spontaneous and natural human interaction in various contexts. They can and should form a basis for work on dialogue modeling for translation systems, prosodic modeling, task and user modeling, development of parsing mechanisms and speech recognition language models, to name only a few, broad areas.

In the following pages is given a diagram of the location of the various versions of the transcription files for the EMMI experiments and of the XWaves-related and speech waveform files as well. The original DAT tapes are located at the time of this writing in a cabinet in the EMMI experiment room, across from the reading/coffee area of ITL.

2. The History

The first experiment we ran, with real human beings coming in to sit in front of the computer interface and have a conversation, generated a great deal of excitement, interest, and untold political ramifications within ITL. The second as well was a rather unusual event in our lab. But by the time we did the third, then the fourth, then the fifth... the experiments began to blur together; nobody took much notice except to be inconvenienced by strangers in the coffee room, and experiments became viewed as, if not a regular occurrence, at least one of not much note. I would like to focus that blur a bit, to make clear why ITL researchers had to give up their places at the coffee room tables, and walk softly down the hall when these people off the streets were participating in an experiment. What exactly was going on?

2.1 The beginnings

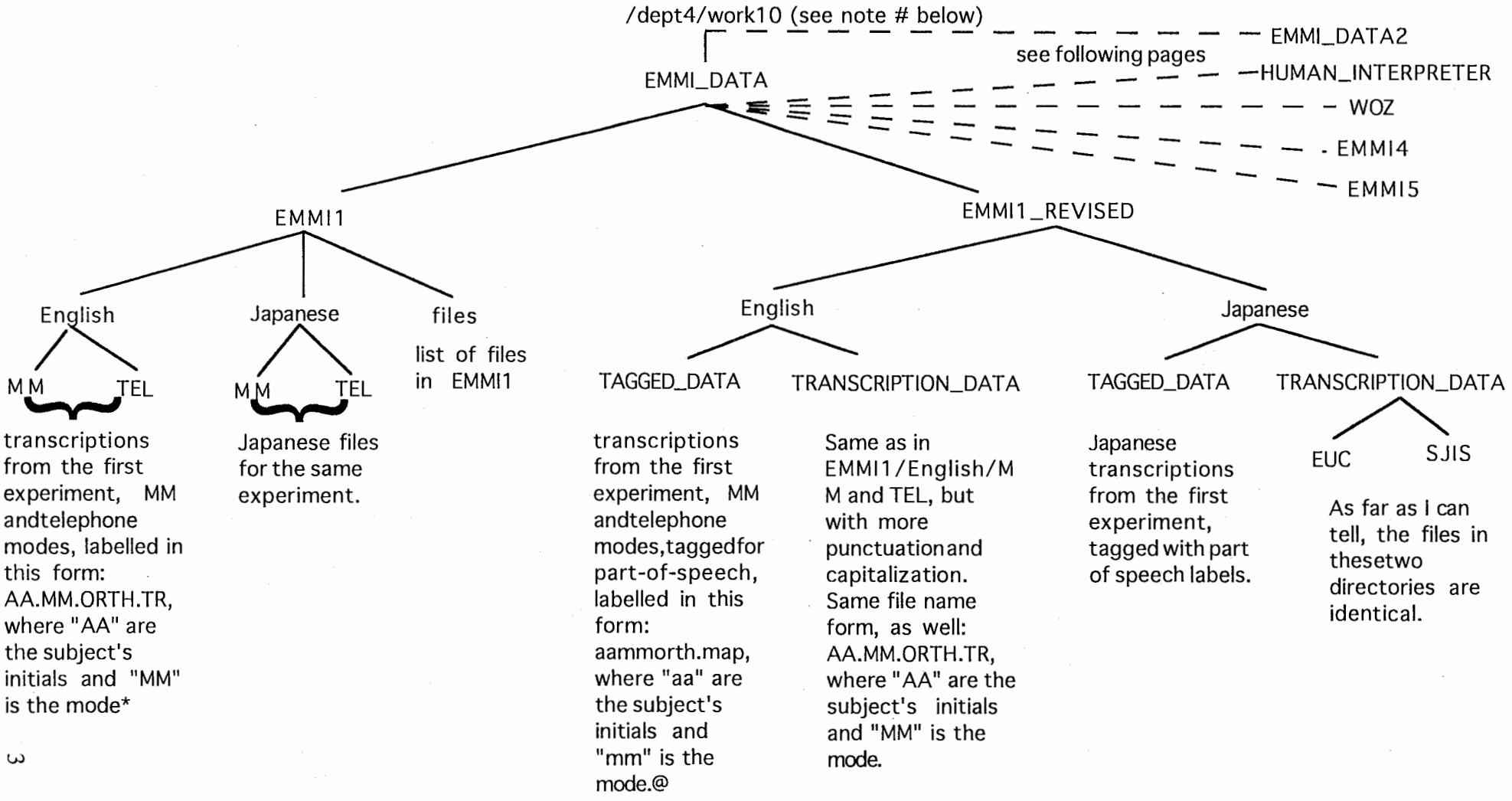
It all began in 1993 when Kyung-ho Loken-Kim designed a multimedia interface that could be used to support giving directions and, later, making hotel reservations. Exactly how and why that happened is not part of this story; I came in on the scene after the interface had been constructed. At that point, Loken-Kim was at a bit of a loss as to what to do with it, but it seemed to him that we ought to try to see if we could ascertain "how well it worked."

Meanwhile, I had been attempting to work on discourse structure and had been looking at some conference registration transcriptions in the ATR database. However, those conversations were staged, interpreted conversations and often were "discourse disfluent." Since I wanted to study *natural* discourse processes, this type of data was clearly not sufficient. I welcomed the idea of recording spontaneous conversations between native speakers of the same language as a way of getting the kind of natural speech data I needed to do real discourse analysis.

2.2 EMMI1: Human-to-human interaction (September, 1993)

The first EMMI experiment was conducted between native speakers of English and between native speakers of Japanese. Ten English speaking participants took part; however, due to "technical error" (I forgot to turn the recording machines on), data from only eight and a half participants were collected (each participant had two conversations; one participant got one recorded and one not recorded; another got nothing recorded at all). The participants were English-speaking friends and colleagues; some were researchers at ATR. They were solicited to do the experiment by me, and paid for their time. In the conversations they had with one another, they asked for and gave directions to a fictitious conference center, via telephone and our multimedia (MM) interface.

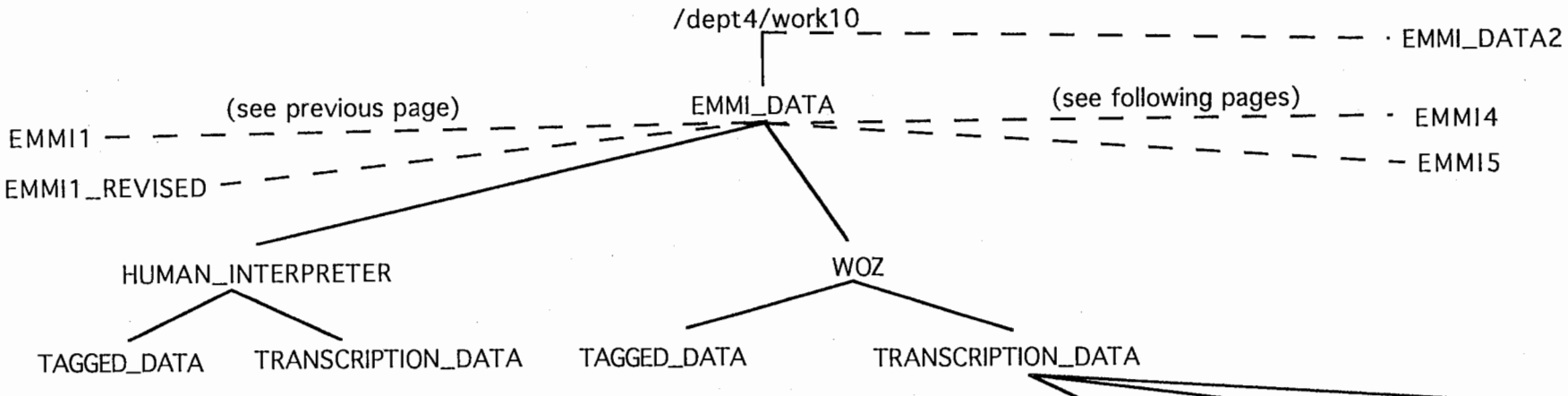
But the biggest question once we had collected the data was, what were we going to do with it? Original memos from the first planning sessions mention looking at the structures of speech used and also the speech acts used, reflecting my interest in syntax and discourse. It was clear that the purpose of the experiment was to evaluate; however, we had no idea *how* to evaluate the interface based on the data we had collected. Under



#Most of the collection and organization of these files is due to the efforts of Suguru Mizunashi, whose work is gratefully acknowledged here. This list of files and directories is not comprehensive, but it does cover all the important sources of EMMI data.

*The labelling of these files was unfortunately more opaque than it needed to be. In most cases, especially in the first three experiments, files have completely unnecessary extensions. In later experiments, however, the extensions are generally meaningful and transparent.

@Descriptions of the part of speech tags that were used are in /dept4/work10/EMMI_DATA2/EMMI.POSlist. These labels were taken from the tags used for the ATR Treebank; the full list of those is in /dept4/work10/EMMI_DATA2/Treebank.POSlist.



Transcriptions from the second EMMI experiment, MM and telephonemodes, labelled with part of speechtags.(See note "@", previous page.) Files are labelled in the form: bbmm.map, where "bb" are the subject's initials and "mm" is the mode.

Transcriptions from the second EMMI experiment, MM and telephone modes. Files are labelled in the form: bb.mm.orth.tr, where "bb" are the subject's initials and "mm" is the mode.

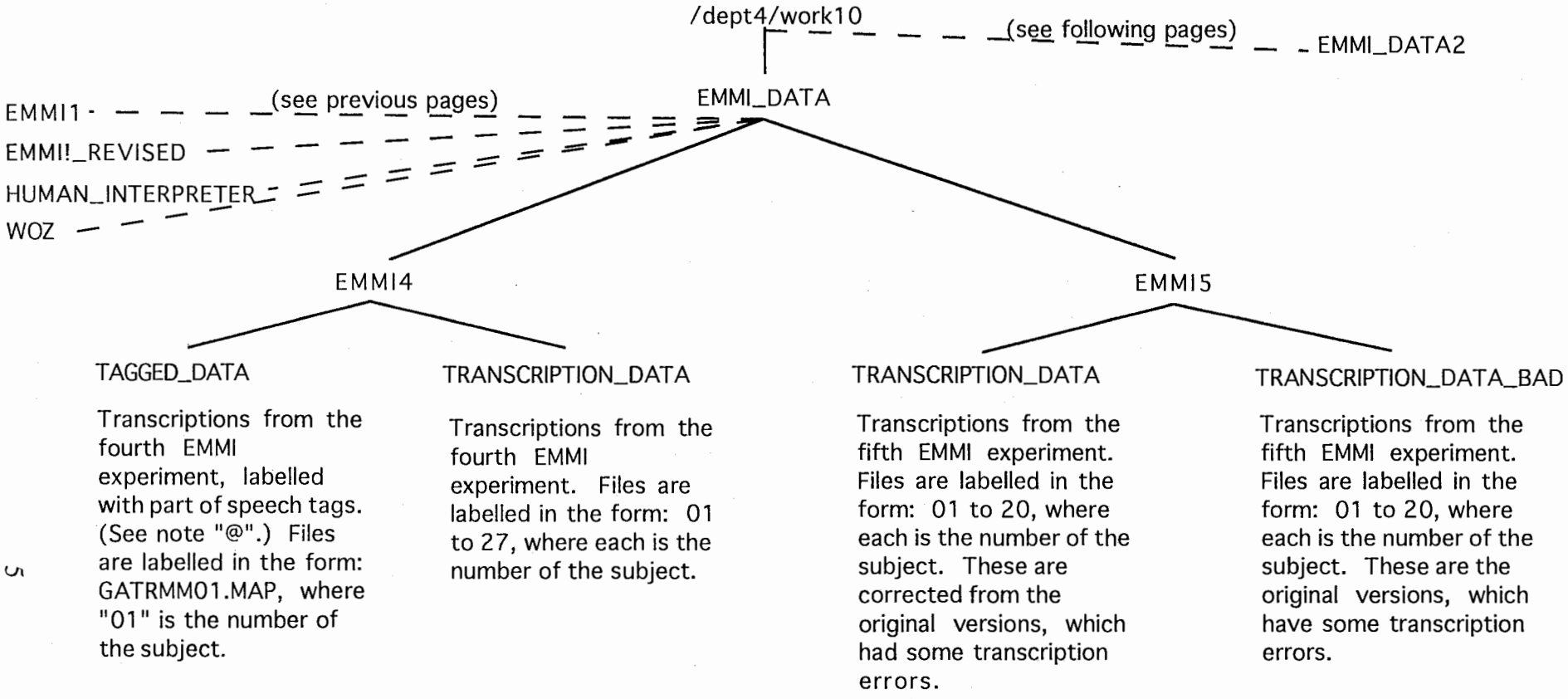
Transcriptions from the third EMMI experiment, MM and telephone modes, labelled with part of speech tags.(See note "@", previous page.) Files are labelled in the form: acmm.map, where "ac" are the subject's initials and "mm" is the mode.

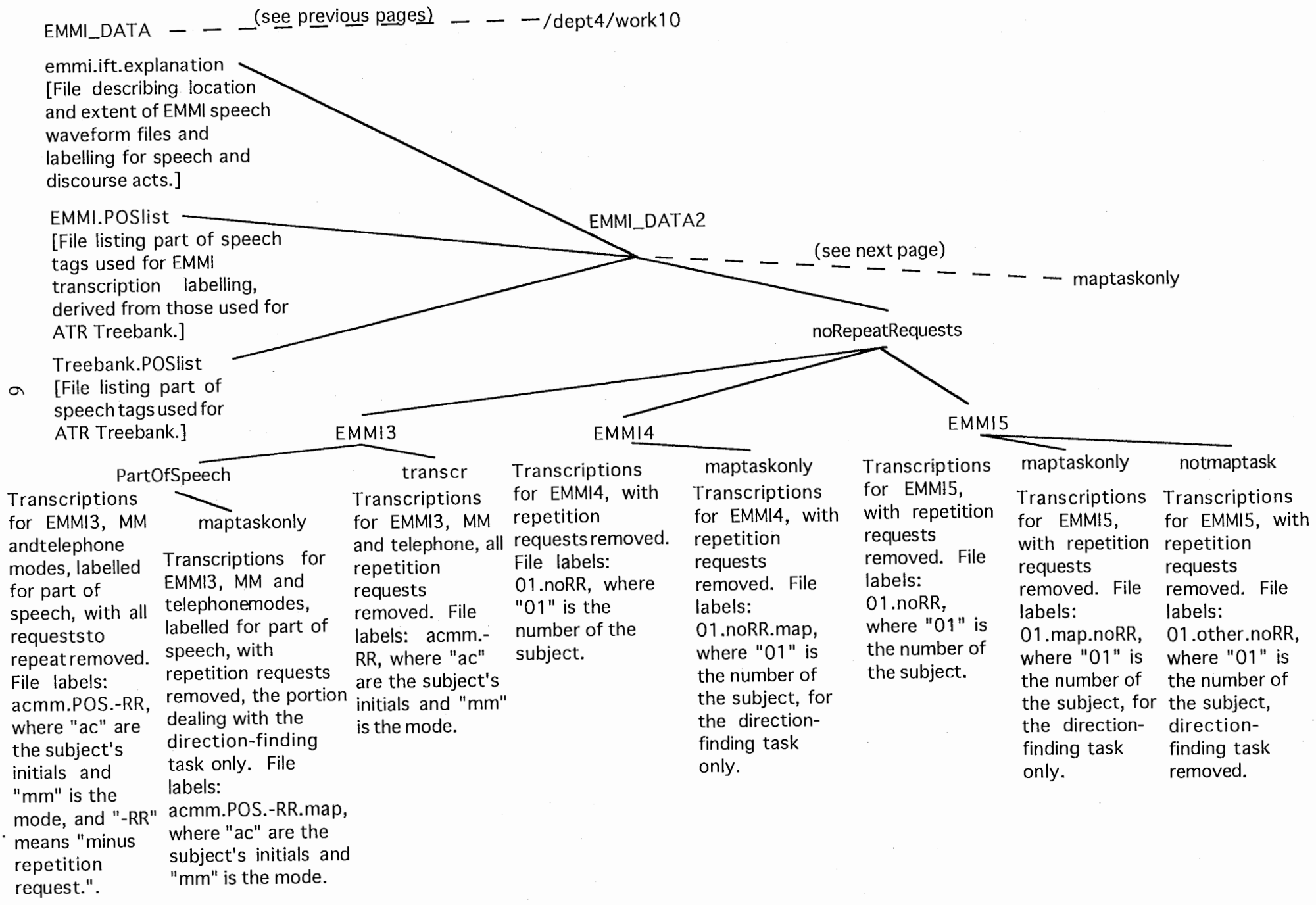
Transcriptions from the third EMMI experiment, MM and telephone modes. Files are labelled in the form: acmm.txt, where "ac" are the subject's initials and "mm" is the mode.

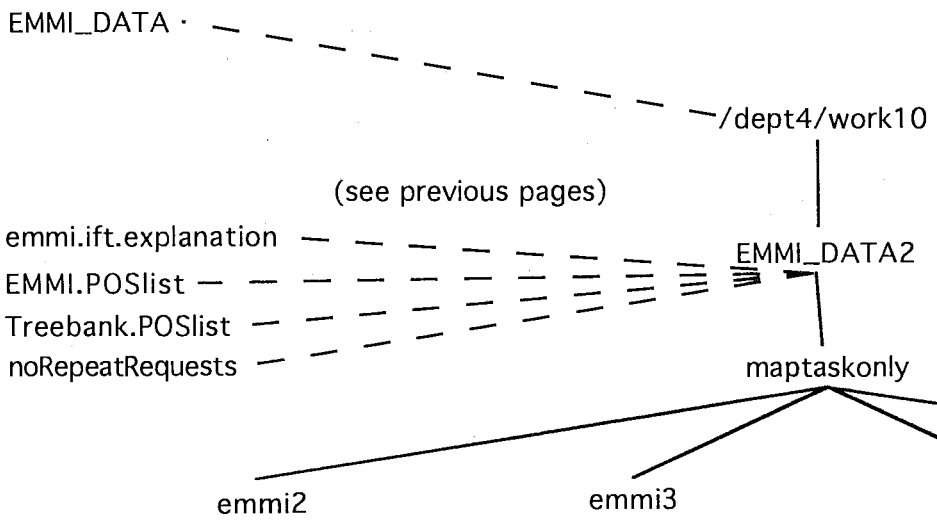
I believe these to be the same files as those in the TRANSCRIPTION _DATA directory (same file name form).

I believe these to be the same files as those in the TRANSCRIPTION _DATA directory (same file name form), but with gesture (i.e., drawings) annotations?.

I believe these to be the same files as those in the TRANSCRIPTION _DATA directory (same file name form), but with annotations for "please repeat"?


/dept4/work10
(see following pages)
EMMI_DATA2
EMMI_DATA
(see previous pages)
EMMI1
EMMI!_REVISED
HUMAN_INTERPRETER
WOZ
EMMI4
EMMI5
TAGGED_DATA
TRANSCRIPTION_DATA
TRANSCRIPTION_DATA
TRANSCRIPTION_DATA_BAD
Transcriptions from the fourth EMMI experiment, labelled with part of speech tags. (See note "@".) Files are labelled in the form: GATRMM01.MAP, where "01" is the number of the subject.
Transcriptions from the fourth EMMI experiment. Files are labelled in the form: 01 to 27, where each is the number of the subject.
Transcriptions from the fifth EMMI experiment. Files are labelled in the form: 01 to 20, where each is the number of the subject. These are corrected from the original versions, which had some transcription errors.
Transcriptions from the fifth EMMI experiment. Files are labelled in the form: 01 to 20, where each is the number of the subject. These are the original versions, which have some transcription errors.


EMMI_DATA
(see previous pages)
/dept4/work10
emmi.ift.explanation
[File describing location and extent of EMMI speech waveform files and labelling for speech and discourse acts.]
EMMI.POSlist
[File listing part of speech tags used for EMMI transcription labelling, derived from those used for ATR Treebank.]
Treebank.POSlist
[File listing part of speech tags used for ATR Treebank.]
EMMI_DATA2
(see next page)
maptaskonly
noRepeatRequests
EMMI3
EMMI4
EMMI5
PartOfSpeech
transcr
maptaskonly
notmaptask
Transcriptions for EMMI3, MM andtelephone modes, labelled for part of speech, with all requeststo repeatremoved. File labels: acmm.POS.-RR, where "ac" are the subject's initials and "mm" is the mode, and "-RR" means "minus repetition request.".
maptaskonly
Transcriptions for EMMI3, MM and telephonemodes, labelled for part of speech, with repetition requests removed, the portion dealing with the direction-finding task only. File labels: acmm.POS.-RR.map, where "ac" are the subject's initials and "mm" is the mode.
Transcriptions for EMMI3, MM and telephone, all repetition requests removed. File labels: acmm.-RR, where "ac" are the subject's initials and "mm" is the mode.
Transcriptions for EMMI4, with repetition requestsremoved. File labels: 01.noRR, where "01" is the number of the subject.
maptaskonly
Transcriptions for EMMI4, with repetition requests removed. File labels: 01.noRR.map, where "01" is the number of the subject, for the direction-finding task only.
Transcriptions for EMMI5, with repetition requests removed. File labels: 01.noRR, where "01" is the number of the subject.
maptaskonly
Transcriptions for EMMI5, with repetition requests removed. File labels: 01.map.noRR, where "01" is the number of the subject, for the direction-finding task only.
notmaptask
Transcriptions for EMMI5, with repetition requests removed. File labels: 01.other.noRR, where "01" is the number of the subject, direction-finding task removed.



/DB/PI/data/emmi

Location of speech act and discourse labels for EMMI1, as well as waveforms. See also /dept4/work10/EMMI_DATA2/emmi.ift.explanation

Transcriptions for EMMI2, MM and telephone modes, labelled for part of speech, direction-finding task only. File labels: bbmm.POS.map, where "bb" are the subject's initials and "mm" is the mode.

Transcriptions for EMMI3, MM and telephone modes, labelled for part of speech, with repetition requests removed, the direction-finding task only. File labels: acmm.POS.-RR.map, where "ac" are the subject's initials and "mm" is the mode.

Transcriptions for EMMI4, with repetition requests removed, direction-finding task only. File labels: 01.noRR.map, where "01" is the number of the subject.

Transcriptions for EMMI5, with repetition requests removed, direction-finding task only. File labels: 01.map.noRR, where "01" is the number of the subject.

"effectiveness" we had written: "until we see how all this goes, right now, subjective. Goal of conversations is for Client to get information--assess each condition for efficiency." It was the right idea; however, it took us until the third experiment to take it seriously. For the first two experiments we opted, instead, to look at things that could be counted: words, turns, disfluencies, and then later, accommodation. Analyzing syntactic and discourse structures at first seemed like an enormous and very subjective task, and counting words and such, a much easier one. Our rationale evolved along with the decision to measure things that were easily measurable: we were looking for Client speech that would reduce the burden on a language processing system such as machine translation. We figured that less speech, less disfluent speech, and more predictable speech (see Fais, 1996a), would do just that. We hoped that an MM environment would encourage such speech from naive users.

The person playing the role of Agent, a second person from outside ATR, and I transcribed the first experiment. This was an extremely valuable experience. I highly recommend to anyone doing experimentation, to follow through every part of it personally at least once before contracting the work out. On many future occasions, after we had begun having an external company do the transcriptions, there were problems with the quality of the work that could only be understood and solved by someone who had actually done the work herself. Since I had had that experience, it was possible for me to be extremely explicit in instructions to the transcribers, to anticipate and check for problems that could cause difficulties for the analysis, to deal in an informed way with the problems they encountered, and to suggest effective ways for the transcribers to solve these problems. First lesson learned.

In addition, doing the transcriptions meant that I had intimate knowledge of the way the conversations went. Much of the later analysis was done by word search-and-count, which requires no knowledge of the real nature of the discourse. It also *reveals* very little about the real nature of the discourse. The only way to understand well what was happening in these sessions was to become immersed in the tapes and transcriptions. Second lesson learned.

2.3 EMMI2: Human-interpreted interaction (June, 1994)

The first EMMI experiment was intended as a sort of benchmark against which to measure future variations on the theme. Since the first experiment was human-to-human, we assumed that that experiment captured basic differences in how humans use telephones and multimedia interfaces. (See Section 3.1 below for an account of what we actually found.) We then wanted to compare this behavior to behavior in a human-interpreted setting, and then in a "machine" interpreted setting, hidden-operator style. Although I already had collected the data I was interested in for discourse work, I remained a part of the "team" as Young-duk Park, a researcher on loan from ETRI for a year, took over the organization of these next two experiments.

This time, the participants were recruited by a company, which also took responsibility for doing the transcriptions. The participants, for some reason, had an unfortunately good command of Japanese; three of them at least had grown up in Japan. Although they all did a masterful job of pretending they didn't know what the Japanese-speaking Interpreter was saying, it is probably from this point that I can date my reluctance to take very seriously some of the aspects of the discourse management of *these* conversations, especially. Would the participants have conducted the conversations the way they did had they not understood Japanese the way they did? It was not a question I had intended to ask from this experiment, and one for which the data were rather sparse and never analyzed. It goes down as a slight, nagging suspicion about the data from this experiment, and the third lesson learned. It's always a good idea to monitor the efforts of outside agencies in carrying out tasks assigned to them. In fact, this same question reared its head again in the fourth experiment.

A second unfortunate twist to the history of EMMI dates from this second experiment as well: the addition of another task, that of hotel reservation. The addition of such a task was actually a welcome one; the task was sufficiently different to elicit new kinds of speech and MM behavior, and it also increased the amount of data we were able to

collect. However, forever after, it required all kinds of manipulations (some of which were very hard to keep track of) in order to compare later experiments (with the two tasks) to the first experiment (with only one). It also meant that we had no "baseline" for human-to-human behavior for the hotel reservation task. Perhaps this was lesson number four: plan ahead.

2.4 EMMI3: "Machine-interpreted" interaction[1] (October, 1994)

With a company responsible for hiring participants and for transcribing, and with Park at the helm, we sailed through the second and third experiments. In the third experiment, we hired two translators, one a native speaker of English and the other of Japanese, to simulate machine translation. Another wrinkle was added to the experimental results in this situation. In order to make the simulation more lifelike (everybody knows there is no such thing as *really* efficient machine translation yet), we instructed the Wizards (named after the Wizard in the "Wizard of Oz" who was actually just a "man behind a curtain") to ask the participants to "Please repeat" when the participants' utterances were too long, too disfluent or too complex. For the first two conversations, Loken-Kim listened in and cued the Wizards; after that they took the initiative and prompted the participants on their own. We were lucky to have two intelligent and thoughtful translators who not only translated well, but also prompted participants consistently (thus eliminating the need for another nagging doubt about how well we controlled the rate of repetition requests) and offered insights and suggestions that were very helpful in designing and running that experiment and later hidden-operator-style experiments.

2.5 Intermission: Protocol experiment (July 1995)

Once the data was in and transcribed from those two experiments, it was time to re-think where we were and where we were going. Results up to this point weren't very clear; we didn't seem to be getting the kinds of answers we were looking for. We expected participants to be using fewer words in the MM condition; they weren't. We hoped for lower disfluencies; we didn't find them. But there were some interesting things going on with the responses that the participants made to the requests to repeat (Fais, Loken-Kim, and Park, 1995), and with how Clients were accommodating to the Agents (Fais, 1996a), and I did some detailed analyses of these areas.

I also decided to tackle the question of the greater number of words head on--the fact the Clients used a greater number of words in the MM condition was counter to intuition, and in fact, counter to claims made by Oviatt and other researchers (Oviatt and VanGent, 1996). Clients should be letting the graphics take up some of the information and shouldn't be expressing so much in language. It was at this point that I finally returned to the notion of information exchange and efficiency, which seemed to hold a clue for where the real results concerning MM were going to lie.

By this time, a few problems with the interface were apparent. Some required only minor corrections, but some were more serious. One major difficulty lay in the fact that there was no starting sequence; the Client simply typed in a phone number and was plunged into the system without instruction. (Of course, each Client was instructed before his participation in the experiment; however, the system itself had no built-in instruction, which would be necessary in a "real" stand-alone system.) At the same time, we were also searching for possible reasons for the lackluster results we had gotten thus far. Not only were the numbers not showing us much that was particularly interesting, but also, our participants weren't making as much spontaneous use of the media as we hoped they would. We suspected that one problem lay in the fact that Clients using the system were not comfortable working with such a MM system, or possibly even with a computer, for that matter. So we asked five ATR researchers, who were extremely proficient computer

---

[1]The quotation marks around "machine-interpreted" will be eliminated below; however, keep in mind that the "machine" interpretation was always simulated.

users, and some even proficient users of MM interfaces, to walk through the tasks using the current EMMI design in a sort of "protocol experiment", critiquing as they went.

2.6 EMMI4: MM and a Persona (March, 1996)

Based on their comments, we did an overhaul of the system, including more polished front- and back-ends, greater initiative built into the system itself, and a face to represent the Persona of the computer (Fais, Mizunashi, Loken-Kim, and Kurihara, 1996). Then, because we suspected that the MM use experience of the participants made a difference in how well they were able to handle the media, we designed our fourth experiment to include participants with little experience and participants with a great deal of experience working with computers (we felt it would be impossible to find a sufficient number of participants with specific, extensive MM experience, so we settled for computer experience).

In this new set-up, Clients could hear everything--both sides of the translation, as well as the speech of the Agent. For this reason, I had requested participants who had a low command of Japanese. Thus began a comedy in communication. The supervisor of our project in the company finding our participants for us didn't speak English; my Japanese was not good enough to communicate clearly with him. So, we both went through intermediaries. Imagine the following chain of questions and answers passing up from some unknown quantity at the company (the person actually doing the recruiting), through several layers there, over to another layer at ATR and then down to me:

The company: What constitutes "a little Japanese?"
Me: Minimum, day-to-day ability and no more.
The company: How can we tell? Please give us some criteria.
Me: I'll try. Not married to a Japanese person; been in Japan less than six months, didn't study Japanese before coming.
The company: OK.

The first participant was fine. The second could clearly understand the Japanese, and in fact, reacted to the Agent's information before it was translated to her. This was a problem for a number of reasons. It led to the same problems with discourse naturalness that we had in experiment two; in addition, since the utterance exchanges in the machine-interpreted setting were so slow (we had the Wizards pause for a while before translating, much as a real machine translation system would do), anticipation of the answer by the Client *before* the answer arrived in English would also affect things like disfluency and overlapping speech. After it became clear that participants three and four were also conversant in Japanese, I initiated more conversation with the company through our torturous chain. The end result was that these students, as they turned out to be, had in fact been in Japan less than six months and were not, in fact, married to Japanese spouses, but had spent their (barely) less than six months here in an intensive study of Japanese. Determined not to compromise the integrity of the data we were collecting, I tried to specify to the company that this was also not acceptable, and they pleaded for more time to find appropriate participants.

It was at this point that we had a stroke of luck. I called to apologize to one of the canceled participants, and she happened to mention something about the description of the experiment that "her friend, who had recruited her" told her. I asked who this friend was, discovered she was an American working for the company and had in fact been in charge of the recruitment, and got her phone number. When I called her and was able to short circuit our previous roundabout route of communication, things went smoothly. With direct communication possible, she could clearly understand my requirements, and exercise her own (excellent) judgment in meeting them. We had learned lesson five: direct communication works best.

Meanwhile, we had begun to let our results be our guide; instead of dictating where we hoped we could get positive results, we started looking at where we actually *were* getting positive results. One such area showed up in the responses to questionnaires after the experiments: participants enjoyed the MM setting, and felt added confidence in the

information because they received it visually as well as verbally. So, for this fourth experiment, we decided to use an attitude scale to attempt to measure and verify the positive attitude our participants reported having toward the MM environment.

Another positive result had been in the area of information exchange. Since we had begun to see higher levels of information exchange in the MM environment, the analysis of the data collected in the fourth experiment focused more closely on that area than it had before.

2.7 Maze experiment (April, 1996)

What is our ultimate goal in all this research? To determine the optimal design for an effective MM interface for machine translation. But what makes such an interface "effective?" Presumably, an interface is effective if it allows the user to give and receive information with some measure of efficiency and clarity. To this point, our focus had been on *giving* information: using the touchscreen to draw a route or circle a location; using the keyboard to type a spelling. But what about receiving information?

We designed a small experiment to test the relative contributions of language and drawing in receiving route information (Fais, 1996b). "Small" in that it was a simple task; participants sat at a computer and listened to a description of a route in various types of wordings and sometimes saw the route drawn as well. (Actually, though, with six different conditions and ten participants per condition, it was probably responsible for the most imposition on Department members' coffee room space.) While the results were interesting, and somewhat unexpected (as so much of the results of these EMMI experiments were), we weren't really ready to take this line of research farther. It ended up as a slim technical report (Fais, 1996b) and a subsubsection in a lengthier paper (Fais and Morimoto, submitted).

2.8 EMMI5: Filling in experimental gaps in the machine-interpreted paradigm (January, 1997)

Well, as is clear from the results discussion below (Section 3.5), experience of the participant didn't seem to be the problem, either. So now we had data from two different machine-interpreted studies (which was good), but the interfaces differed fairly drastically in design (which was bad). The results were still not directly comparable. In what has turned out to be the final effort to understand the causes of the results we were getting, we designed the fifth and last EMMI experiment. This one fills in the gaps in the two-by-two design comparing the presence and absence of MM and the presence and absence of the Persona face. We already had data from the absence of both (i.e., the telephone data from the third experiment), and from the presence of both (the fourth experiment). The fifth experiment contained two more conditions: face alone and MM alone (of course, we had MM alone in the third experiment, but with a different system; we wanted data from the same interface so that it would be maximally comparable.)

With our vast experience behind us, this experiment went smoothly: the same recruiter located good participants for us; I knew enough to have helpers whose entire responsibility was to turn on the recording machines; our Wizard and Agent were experienced and needed minimal practice. Now that we were at the end of the series of experiments, we knew how to do them properly.

3. The Results, by EMMI

I will approach this section in two parts. In the first, I'll give general results for each experiment. Details can be found in the various reports associated with each and given as references. My intention here is not to describe results in detail; rather, I would like to discuss the motivations behind the results, reasons that are sometimes too "unscientific" to be mentioned in the "scientific" accounts of the results found in the bibliography. Thus, please do not use this Technical Report as a reference for results. Much of interest has been smoothed over or ignored in the presentation below and the complete results are available in the references given in each section, which should be the sources cited.

In the second part, I'll give graphs of the results from all five experiments plotted together, with a brief discussion of the peculiarities of each.

3.1 EMMI1: Human-to-human

As alluded to above, the results from the first experiment were somewhat disappointing (Fais and Loken-Kim, 1994). Mostly, they had to do simply with the effects of learning: Clients got more disfluent and more efficient in their second trials. And the Clients who participated in the telephone condition first used a greater number of words and exchanged larger amounts of information.

With Park's arrival, Loken-Kim was also interested in building a drawing recognizer that could not only recognize the shapes of the drawings made by the users, but also link them to the referring phrases they matched in the utterances. As a part of this work, I analyzed the use of deictic expressions and the like in this data (Fais, Loken-Kim and Morimoto, 1996). Despite the claims that we made about using this data to model the users, i.e., to better define the roles of Client and Agent, it was never actually used for that purpose. In fact, given our scenario, it would only have been useful for modeling the Client, since we considered our Agent to be trained, a known quantity, and not in need of "modeling" from his utterances. We had data in search of a reason for being.

In any event, the drawing recognizer that Loken-Kim and Park eventually developed was probably the most interesting thing to come out of the work with the first EMMI (Loken-Kim *et al.*, 1995). Certainly, we could not demonstrate any real differences between participant behavior in the telephone setting and that in the MM setting.

The other interesting result, though we didn't see it that way at the time, was the first indication that our intuitions concerning the use of visual information in an MM setting to substitute for words, were wrong. We tried to show that they were right; and did some rationalization to allow us to report only the condition (the first condition in which the participant performed the experiment) in which these intuitions did in fact hold. The rationalizations were not deceitful; however, they did not tell the whole story. But then, at that time, we didn't know the whole story. That came later.

3.2 EMMI2: Human-interpreted

Since EMMI3 came hot on the heels of EMMI2, I never did a separate analysis of the data from EMMI2. However, that data figured large in subsequent comparisons of the data from the first three experiments. As can be seen below, this experiment stands out from the others in many ways. It certainly was the most verbal of all the experiments, having the highest number of words, overlapping speech, and accommodation. I will characterize this experiment further in comparison to the others in Section 4 below.

3.3 EMMI3: Machine-interpreted

The hint that people perhaps did *not* replace words with the use of visual information in an MM setting that we first saw in EMMI1, became a certainty once the data from EMMI2 and EMMI3 were added (Fais, 1996a). And the disappointing suspicion that disfluency would not be affected by MM was also confirmed. On the other hand, the results for accommodation were somewhat hopeful: while accommodation in the machine-interpreted environment of EMMI3 was not as high as that in the human-interpreted environment of EMMI2, it was still higher than in the human-human environment of EMMI1. But still, the only thing we had to fall back on for explanation was the possibility that experience was playing a part. And the only really clearly positive result we could show was that, according to the questionnaires that our participants filled in, they enjoyed using the MM interface and thought it was useful.

3.4 Design research

At this point we stepped back a bit and tried to understand what might be the best way to re-design the interface. We had a number of goals and achieving them dictated some

major decisions. We wanted to incorporate more ITL technology in order to come closer to an actual functioning system, so we decided to use the CHATR synthesizer to output the English (Campbell, 1996a; Campbell, 1996b). We wanted to try to cut down on the amount of meta-media speech used by the Clients, so we incorporated instructions for use of the media into a stand-alone front end for the interface in which the computer translator took more discourse initiative than before. We wanted to see if we could increase the level of accommodation on the part of the Clients; we also wanted to reduce the amount of overlapping speech by helping them understand more clearly what was going on in the conversation. For both those reasons, we included in the interface, the face of the computer translator "persona." We felt that participants might accommodate more to a human-like computer translator, and we could use movements of the head of the Persona to indicate whose turn it was to talk (Fais, Mizunashi, and Loken-Kim, 1996). The end result, then, was the interface for the fourth experiment.

3.5 EMMI4: Experience of the user, and interface

As hinted at above, we found in this experiment that the experience of the user, at least in the form that we were testing it, had little effect on the user's linguistic or paralinguistic behavior in EMMI (but see Section 4.11 below). This was a welcome result in that it allowed us to lay to rest a suspicion that had been nagging us since the first experiment. But it still left us without a very satisfactory explanation for why users didn't take more advantage of the "benefits" of MM, why they used more words in the MM environment, why they didn't behave the way we expected them to. Of course, there is another explanation: our expectations were wrong.

We did find, however, that we were able to encourage more information exchange and more efficient information exchange and that our Clients needed less prompting to type in the new interface. The last, especially, seems to have been a direct outcome of the online instructions; Clients shown a form and told simultaneously that they can type on it are more likely to do so than those whose instructions came before they even began the task.

However, our hopes for less meta-media speech, for lower disfluency, and for greater accommodation were not borne out.

3.6 Maze experiment

In this experiment, we tested how much information participants had gotten from watching a map and hearing a route description by asking them to draw the route and describe it in words afterwards. Some results were expected: participants wrote the best descriptions after hearing the most detailed description (regardless of seeing or not seeing a drawing). Some results were interesting: hearing *any* language at all, even the most simple, boosted participants' scores for drawing the route later. And some results were surprising: those who drew the best routes were those who had heard a language description of the route but had *not* seen it drawn.

These results all point to language as the best vehicle for conveying information, *even spatial information* such as a route description. What did we make of this in the context of an MM interface? Well, in fact, we made very little of it. We did not even follow our own advice and construct rich language descriptions for our Agents when we pre-scripted their utterances in EMMI's 4 and 5. The latter was an oversight at the time. On the other hand, it could be exploited now. Having scripted directions in a real EMMI task with minimal language accompaniment and deictic drawings (see Section 4.11) in EMMI's 4 and 5, we could experiment with participants' information gathering in the same conditions, but with rich language descriptions and redundant drawings. Maybe EMMI6.

3.7 EMMI5: Effects of the Persona

With the results of this fifth experiment, we were able to compare the relative effects of the presence of the Persona and that of multimedia. Again the results were equivocal: the presence of the Persona reduced the amount of meta-media speech Clients used, and

increased the amount of information. All to the good. But it was the *absence* of the Persona that yielded the least words-per-turn (a desirable result for speech recognition). All other effects of the Persona tended to be positive, but were not significant.

It was in the *absence* of multimedia that participants tended to have lower disfluency and higher accommodation. But the *presence* of MM yielded significant fewer words-per-turn, and somewhat less overlapping speech. The most resounding positive effects of MM were still in the areas of information and attitude: both amount and efficiency of information exchange were increased significantly and attitude was (also significantly) favorably affected in the presence of MM. (Fais and Morimoto, submitted)

4. The Results by Measure.

Over the period of time from the planning of the first experiment to the analysis of the last, we have examined a vast number of different measures, depending upon what kind of information we needed. Although we had a certain core of measures we were interested in, the results always lead to new questions. Did the Clients using the MM environment first in EMMI1 behave differently from those using the telephone condition first? Did the Clients in the Persona experiment exchange more non-typical information than those in the EMMI3 experiment? Sometimes we could investigate these questions, and the answers, in fact, helped us to understand the results of the experiment. Sometimes they dead-ended and we were left without being able to explain our results very satisfactorily. The fruitful paths are documented in the references listed for each experiment. The ones leading nowhere gather virtual dust on my Macintosh desktop.

Again, however, it is the major results that we report below, the results that, no matter the experiment, we analyzed time after time. For each measure, we give a graph of the results across all five experiments and relevant media. Speculations as to the reasons for the differences follow.

4.1 Words

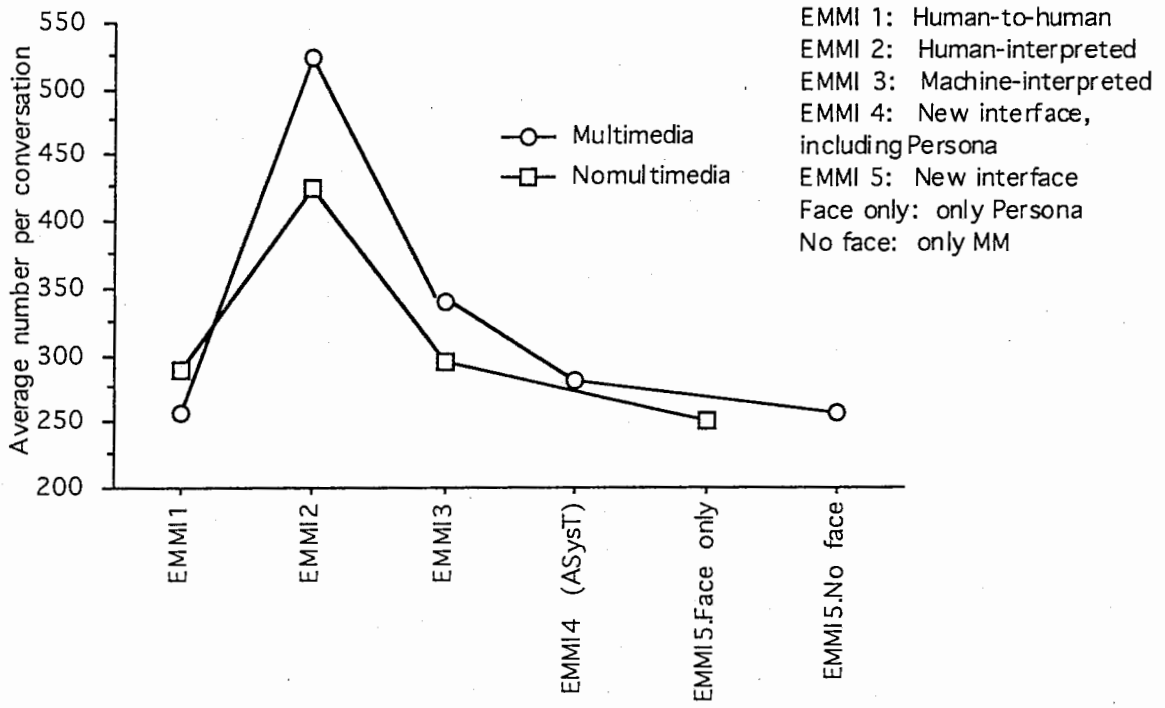


Figure 1. Average number of words per conversation for each EMMI experiment condition, for Client only.

(Recall that EMMI1 involved only one task; the numbers of words for that experiment have been adjusted to take this into account.)

Across all five experiments and all the various experimental conditions, with the exception of the human-interpreted experiment, the number of words used is about the same. That this should be so across such a wide variety of communicative settings is striking--participants simply did not change the number of words they used to complete these tasks, no matter what the situation.

As we said, with the exception of the human-interpreted experiment. As mentioned above, there was never any analysis done that focused explicitly on that experiment. However, a close reading of the transcripts yields some suggestions. There was a small amount of exchange between the Client and the Interpreter (in addition to the standard exchange between the Client and the Agent), which could account for some of the increase. That is, the conversation was not always a truly two-sided, interpreted conversation. The number of words was also inflated in this condition by a propensity for Clients to spell their names aloud to the Interpreter, resulting in exchanges that went something like this:

Client: B
Interpreter: B
Agent: B?
Interpreter: B?
Client: Yes, B.
Interpreter: Hai, B.
Agent: So desu ne, B.
Interpreter: I see, B.
Client: A...

etc. Each of the utterances of the letter was counted as a word. This kind of exchange was not limited to this experiment, but seemed to be more prevalent here. This was one reason why we wanted to encourage Clients to use the typing option by including more explicit instruction. It seemed highly inefficient for Clients to spell their names through an Interpreter. However, although that was easy for us to see, Clients still persisted in spelling their names, despite the fact that the typing option would have been much more efficient. Lesson six: Participants simply do not know and sometimes do not recognize what the experimenter considers to be the most efficient or best options for communication (Fais, Loken-Kim, and Morimoto, submitted). The participant simply does what he is used to, or what seems like the most effective, or what happens to strike him at the time. The motivations behind media use by participants are still extremely mysterious.

The picture for words used by both Client and Agent combined was different. This picture caused us to question the notion that the use of visual media to convey information would result in fewer words used.

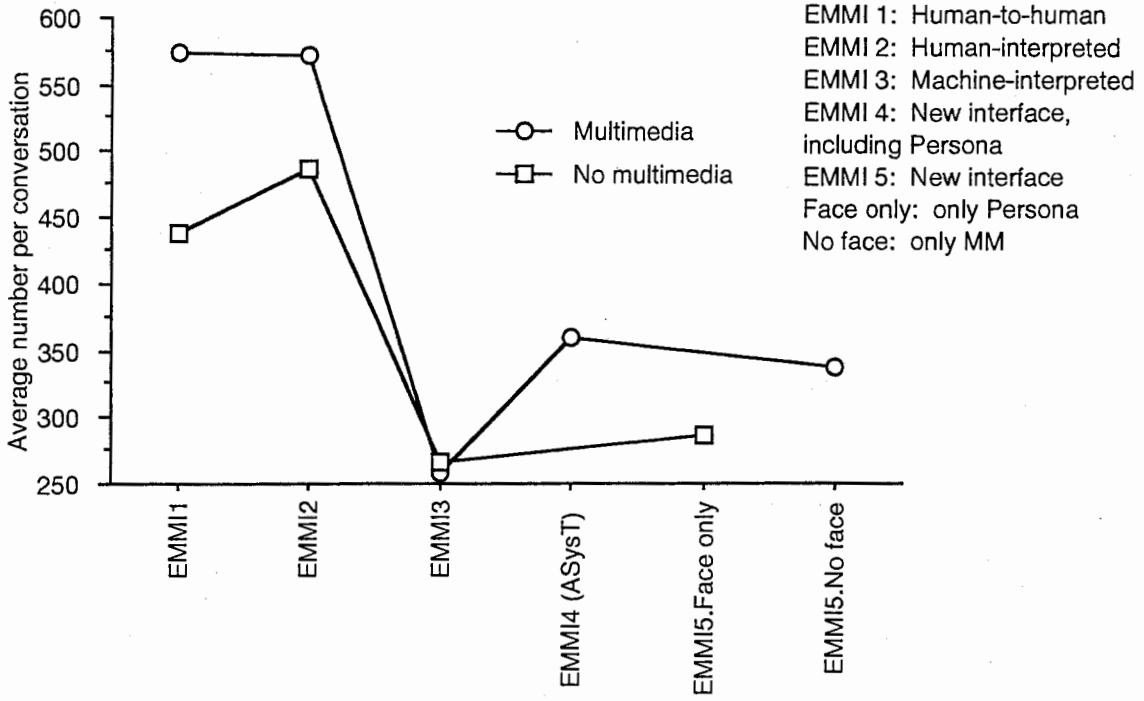


Figure 2. Average number of words per conversation for each EMMI experiment condition, for Client and Agent combined.

4.2 Words-per-turn

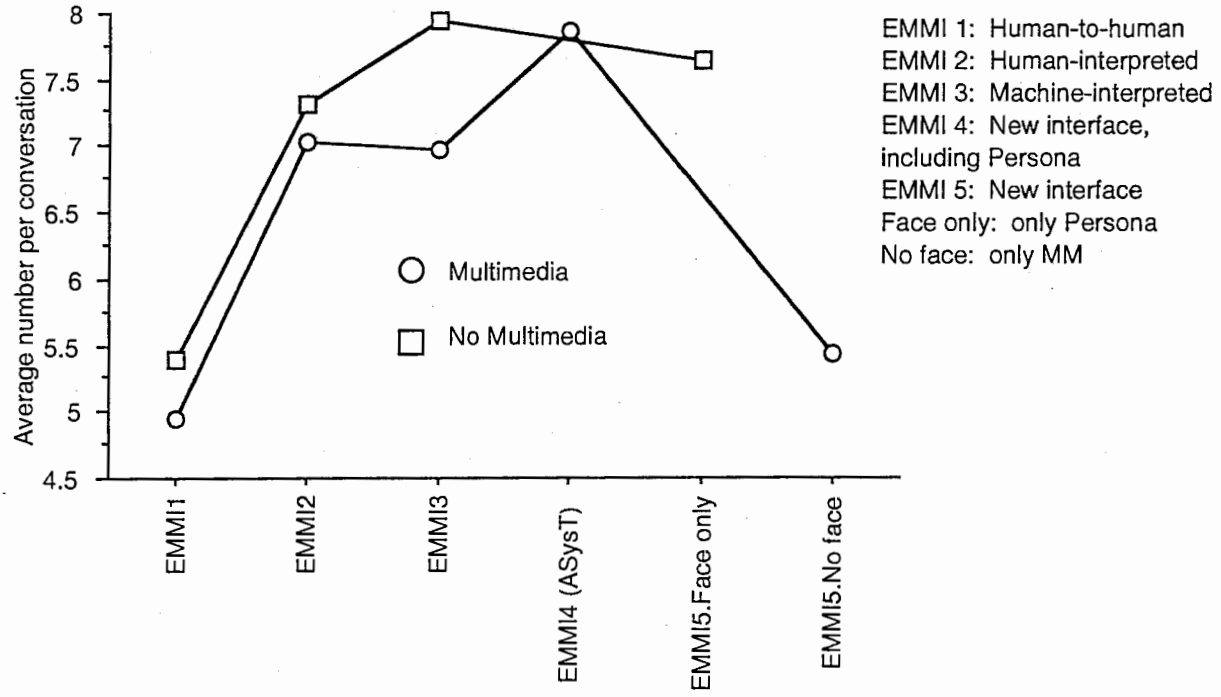


Figure 3. Average number of words-per-turn per conversation for each EMMI experiment condition.

The most easily processed speech for current speech recognizers, comes in small packages. Our smallest packages came in EMMI1 and the MM-only condition of the new interface.

There are a couple of points of interest here. Let's ignore the results for the EMMI5.No face condition for the moment. In all other conditions, we have a sharp contrast between the short utterances in EMMI1 and the longer utterances in all the other experiments. In EMMI2, turns sometimes included conversation directed to the Interpreter as well as to the Agent, which might have accounted for some of the extra length. In addition, in all mediated conversations, the Client and Agent both tended to give longer explanations to be sure that the Interpreter understood. Further, in the machine interpreted conditions especially (EMMI's 3 through 5), there were long pauses while the machine was "translating." In order to make those pauses worth the wait, Clients tended to pack as much into an utterance as they thought was reasonable, rather than engaging in quick back-and-forth exchanges, simply because the exchanges could never be "quick." This is especially apparent at the beginnings of tasks, where Clients sometimes tried to give all relevant information in one utterance:

Client: Hello, My name is Smith and I'm going to attend the conference tomorrow and I've just arrived at Kyoto Station and I would like information on how to get to the Conference Information Office from Kyoto Station

...somewhat longer than the apparent average of nearly eight words per utterance. It is interesting to note that utterances such as these got a "Please repeat" response from the machine translator. Almost always, the Client did not take that literally, but simplified his/her request in some way. Across a number of repetition requests, then, Clients were, in effect, trained to speak in shorter utterances. However, that still did not have the effect of yielding average words-per-turn that were as short as those in "normal" human-to-human speech.[2]

Now, let's return to the EMMI5.No face situation. What was so different about this situation? The instructive comparison here is between EMMI5.No face and EMMI3. These were both MM interfaces, with no Persona. However, the EMMI5.No face interface also had no video image of the Client or Agent, whereas EMMI3 MM condition did. Thus, the EMMI5.No face situation was the only MM situation in which the Client saw no visual image of any interlocutor, whether Agent or machine translation Persona. In EMMI1, the Client saw a video image of the Agent; likewise in EMMI2 and EMMI3. In EMMI4, the Client saw both the Agent and the Persona and in EMMI5.Face only, the Client saw nothing but the face of the Persona and that of the Agent. We could conjecture that the absence of any visual image of an interlocutor contributed to lower numbers of words-per-turn.

But what of the telephone conditions for EMMI2 and EMMI3? The Client saw no visual images of interlocutors in those conditions either. Well, perhaps the mediation effect described above was responsible for the longer utterances in the telephone conditions of EMMI2 and 3.

Now do you see what I mean about these sorts of explanations being difficult? If we rely on the mediation effect, it explains the differences between EMMI1 and the rest, except for the short utterances in EMMI5.No face, which *was* mediated. If we explain the longer utterances on the basis of the presence of a video image, then we haven't explained the longer utterances in the telephone conditions of EMMI2 and 3 or the shorter utterances in the MM condition of EMMI1, but we have accounted for the results from EMMI5. No face.

This is a clear example of the kind of question that fuels research projects like this one for years--if we have faith that a rational explanation exists, which as research scientists

---

[2]Clients tried a number of strategies for avoiding repetition requests. The most common strategies were to speak more loudly and more slowly. They also simplified their expressions, or changed vocabulary items that they thought the machine might not understand. It is my impression, having watched nearly 60 of these machine-interpreted conversation, that reducing utterance length is the last strategy for Clients to try and to learn.

we must, then we cannot be satisfied with this account. We must look further into the interplay between user behavior in contexts with mediation and in contexts with video images of the interlocutor in order to understand this phenomenon sufficiently. And my guess is, that the results to the next, carefully designed study to do just that, will simply lead to further puzzling outcomes.

4.3 Disfluency.

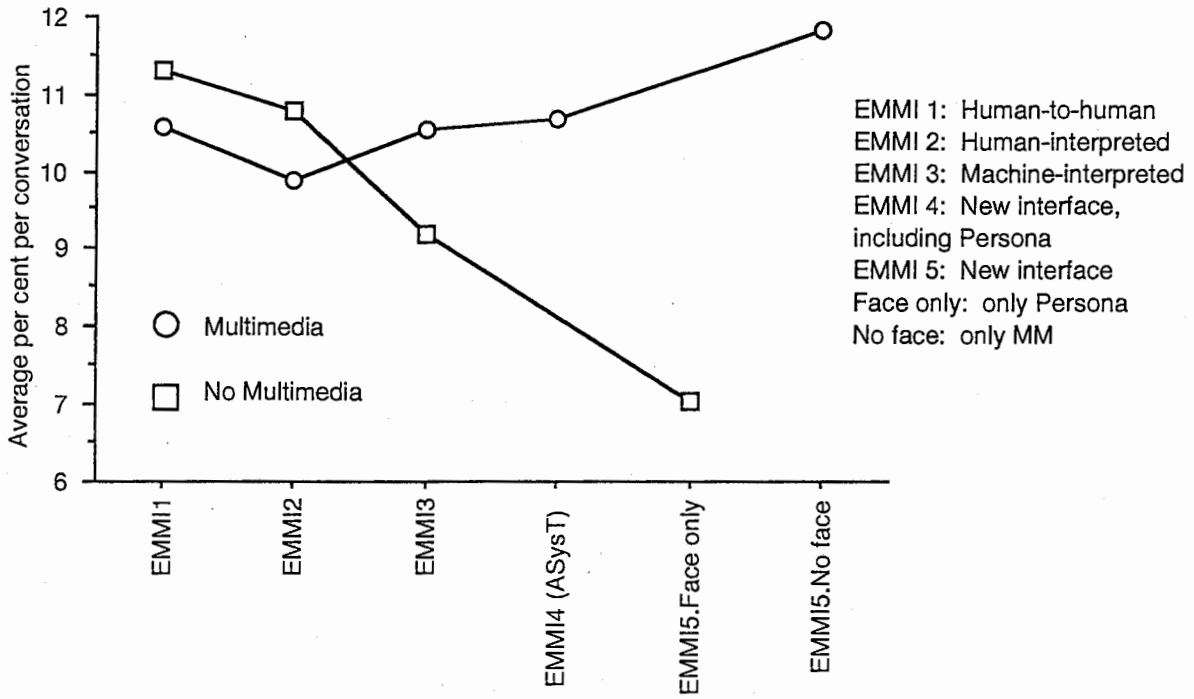


Figure 4. Average per cent disfluency per conversation for each EMMI experiment condition.

It is easy to see the disappointing results we were getting initially for disfluency rates; in the first three experiments, there is virtually no difference, except for what ends up being a hint of a trend in EMMI3 (but only in hindsight, when combined with the results for EMMI's 4 and 5) towards greater disfluency in the MM cases. Basically, what we have here are virtually equivalent results for all conditions of all experiments except for EMMI5.Face only, which shows a marked decrease in disfluency. This time, it's easy to pinpoint the "cause," since this condition is the only one of its kind: the only one with mediated conversation, via a non-telephone channel, with simply the visual images of the interlocutors.

Of course this time, we are plagued with too much information. Is it the fact that it was a *machine* interpreter in this situation that had the most effect? That is, would we have gotten the same results if the mediator had been a human being? Would we have gotten the same results if the participant had spoken into a *telephone* while viewing the images on the screen? What if there had only been the images of the Client and the Agent? Or only the Client and the Persona?

At least this time, we have some ideas to go on. Clearly, disfluency is not low in the presence of MM. Every result is consistent with that. It is also not low over the telephone. It *is* low when viewing the faces of the other interlocutors, but only the faces, i.e., no other visual material. It seems consistent with all these results to say that MM options might distract the speaker, making utterances more difficult to plan, and leading to greater disfluencies (Fais and Morimoto, submitted). To confirm that, we might want to investigate the questions raised above. At the very least, we've discovered that MM options are not going to be helpful in lowering disfluency rates.

4.4 Simultaneous speech

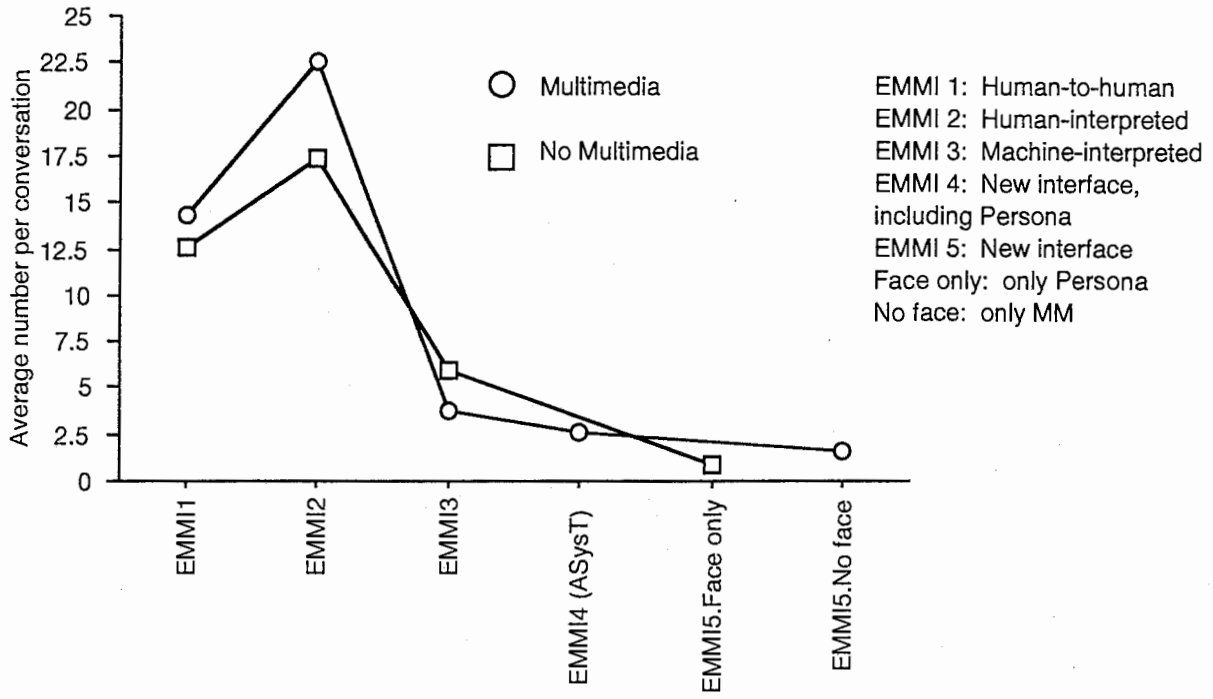


Figure 5. Average number of instances of overlapping speech per conversation for each EMMI experiment condition.

Now here, finally, is a pretty graph, one we can more or less draw some conclusions from. Clearly, rates of simultaneous speech fall off sharply in the machine-mediated conversations. One question does present itself, however. Is that because it is a *machine* "translator" or because of the very long pauses in between utterances (one is inclined to suspect the latter). Would we get the same results if the machine translator worked as quickly as the human Interpreter? This could, of course, be a future study.

4.5 Accommodation

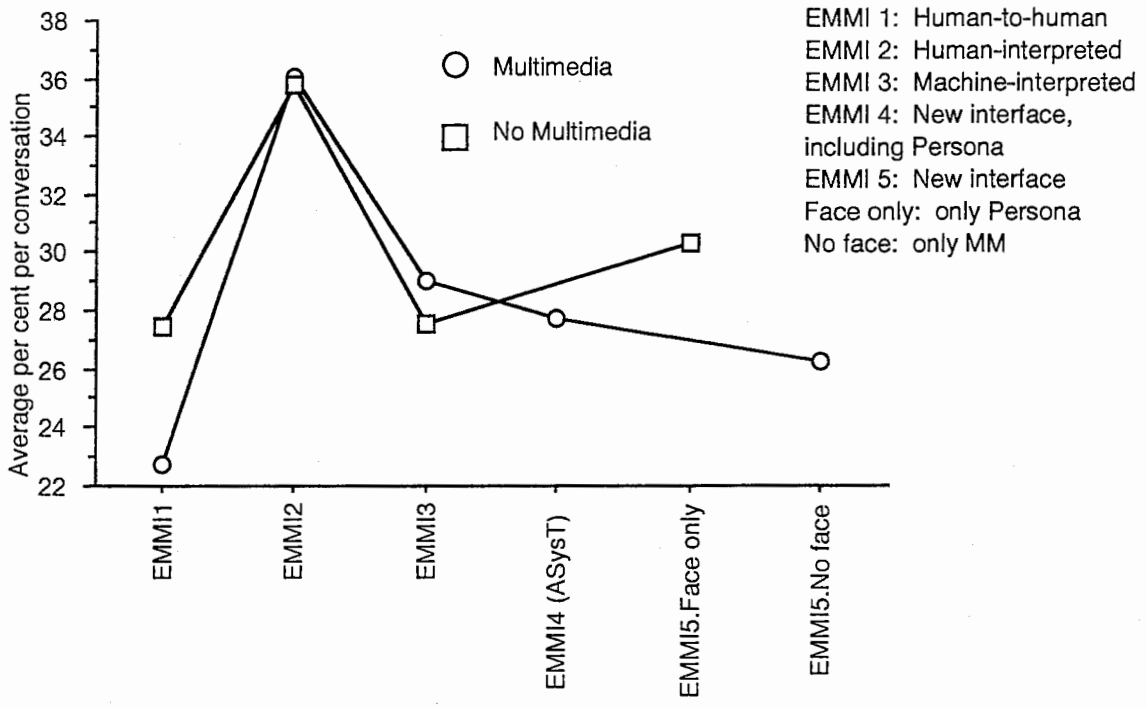


Figure 6. Average per cent accommodation per conversation for each EMMI experiment condition.

With results from the first three EMMI's, we did an extensive examination of the phenomenon of accommodation (Fais, submitted). In that study, we concluded that efforts on the parts of *both* the Interpreter and the Client to "help each other out" by adopting common lexical items were responsible for the very high rate of accommodation in EMMI2. On the other hand, efforts on the part of the Client to make himself understood to the machine by using words that the machine had used previously were responsible for the somewhat higher average rates of accommodation for EMMI3 than for EMMI1 (especially in the MM condition where the fact that the interlocutor is a machine is strongly visually reinforced).

Recall that we added the Persona to the interface in the express hope of capitalizing on the greater tendency to accommodate with a human-interpreter. We hoped that this would extend to human-*like* interpreters. This hope was ill-founded in the case where the Persona was buried amidst the rest of the MM options available, as in EMMI4. Where the Persona was central, however, that is, in EMMI5.Face only, we did get higher levels of accommodation. It seems we were on the right track after all, except for the fact that the presence of the Persona has no effect unless it is extremely salient. How much is "extremely?" Would increasing the size of the image make a difference? What about changing the placement? Next study...

4.6 Meta-media speech

In investigating the balance between the use of speech and the use of visual channels (drawing and typing) to exchange information, we were struck by the fact that the *expected* relationship between these two does not, in fact, exist. We might think that where participants use more speech, they will use less visual information, and where they use more visual information, they will use less speech. This, in fact, not the case. Instead, where participants use more visual information, they also use more words, both meta-media speech to manage their media use, and informative speech to accompany the visual information (Fais and Loken-Kim, 1996).

One of our very first hopes was that the inclusion of MM options in a machine translation interface would lower the number of words participants used. When we discovered that this did not happen, we, of course, wanted to find out why. The phenomenon of meta-media speech was the first culprit to be blamed, though later it became apparent that even the greater amount of meta-media speech in MM conditions did not account for the greater number of words there; participants were also using drawings redundantly with speech (Fais and Loken-Kim, 1996).

Having happened upon the existence of meta-media speech, our first reaction, in accordance with our initial goal to reduce the amount of speech used, was to attempt to eliminate it.

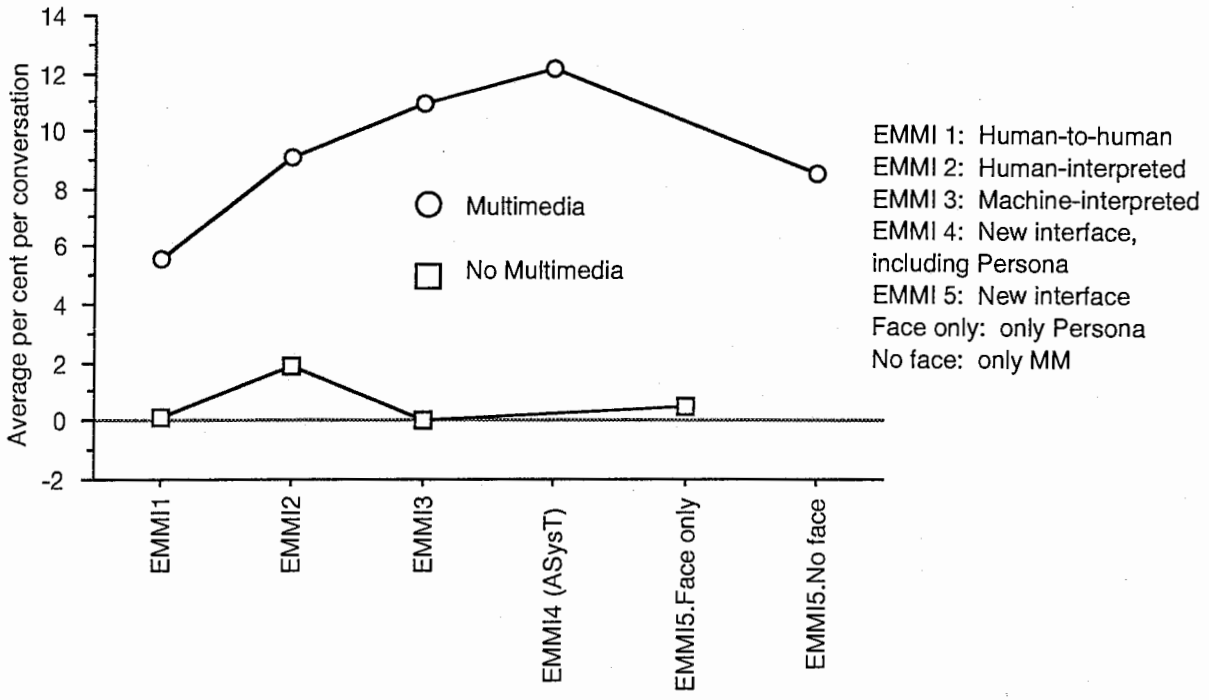


Figure 7. Average per cent meta-media words per conversation for each EMMI experiment condition.

This graph shows merely the obvious: that participants talk a lot more about using MM than they do about using the telephone or, in the case of EMMI5.Face only, than they do about using speech and watching an image. It also shows, however, that we did not in fact succeed in our goal of reducing meta-media speech with the new interface (EMMI4 and EMMI5.No face).

We had hoped to do that by including instructions that appeared online, automatically, in the course of the participants' use of the interface. However, as can be seen from the slightly elevated levels of meta-media speech in EMMI4, the need for instructions seems not to have been the issue. Instead, what seems to be the problem is how complex the interface is. EMMI4 represents the peak of complexity, as well as the peak of meta-media speech: it includes MM options, online instruction and the Persona. Eliminating the face, as in EMMI5.No face seemed to help, as did eliminating the instructions as in EMMI3. Eliminating the machine aspect helped, too (EMMI2) and eliminating mediation helped the most of all (EMMI1).

This is yet another indication that the complexity of the interface was a drawback to achieving our goals.

4.7 Information

If those goals are linguistic, that is.

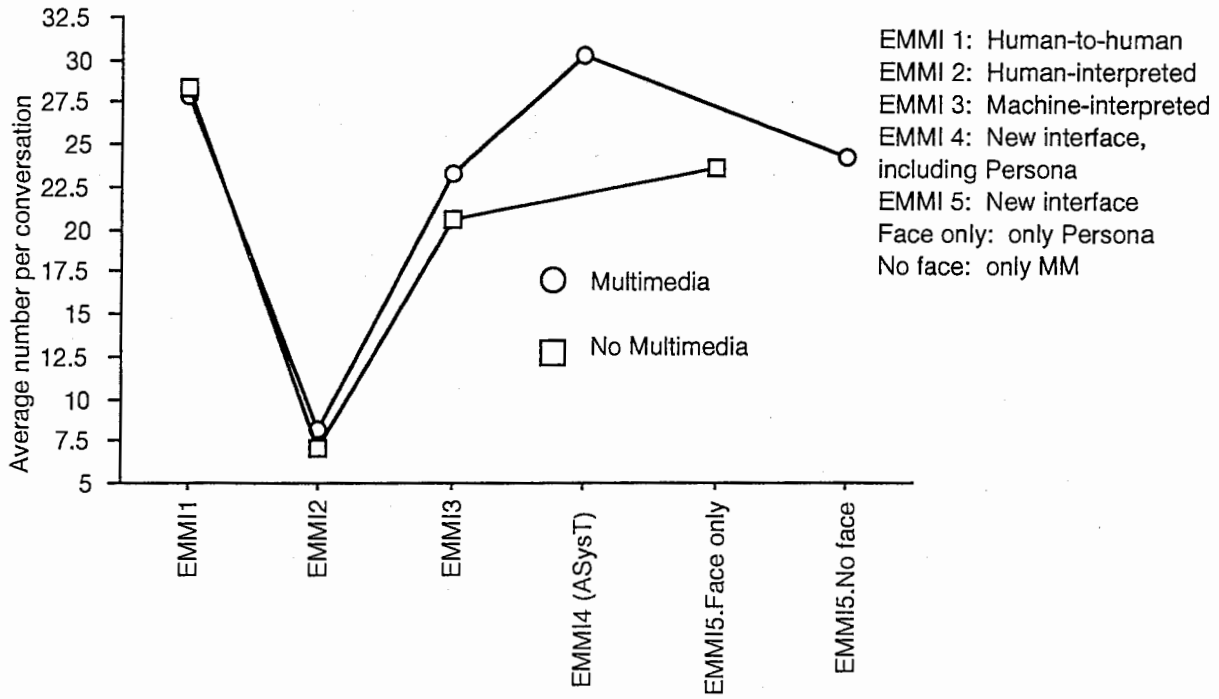


Figure 8. Average number of information units per conversation for each EMMI experiment condition.

These are the results for information exchange (the notion of information units is extensively discussed in Fais, Loken-Kim and Morimoto, submitted). Participants exchanged the most information in the most comfortable, natural situation (human-to-human) and also, what do you know, in the most *complicated* interface, the EMMI4 interface with instructions, Persona and MM options. It is clear that for the machine-interpreted situations, the presence of MM options is a distinct advantage to encouraging information exchange. It is interesting that it requires the context of machine interpretation to make this advantage apparent. We've conjectured elsewhere (Fais and Morimoto, submitted; Fais, Mizunashi and Loken-Kim, 1996) that participants simply enjoyed the MM conditions more and so were happy to use it long enough to get more information (see also Section 4.9 below). But why, then, does that explanation not hold for the human cases (EMMI's 1 and 2) as well? Perhaps the participants' involvement with the other human partner somehow precluded focusing on the MM aspect of the interface. The human being was more salient than the interface itself. On the other hand, it might be that the MM options of the machine-interpreted interfaces are more salient than either the machine interpreter (who is, after all, not human) or the human Agent (who you are not really talking to directly and, furthermore, can't understand).

So, sigh of relief, a clear, strong argument for the inclusion of MM options in an interface. It was a long time in coming. But this is also a result that is not apparent from initial studies with human conversational partners, one that is specific to the area of machine translation.

4.8 Efficiency in information exchange

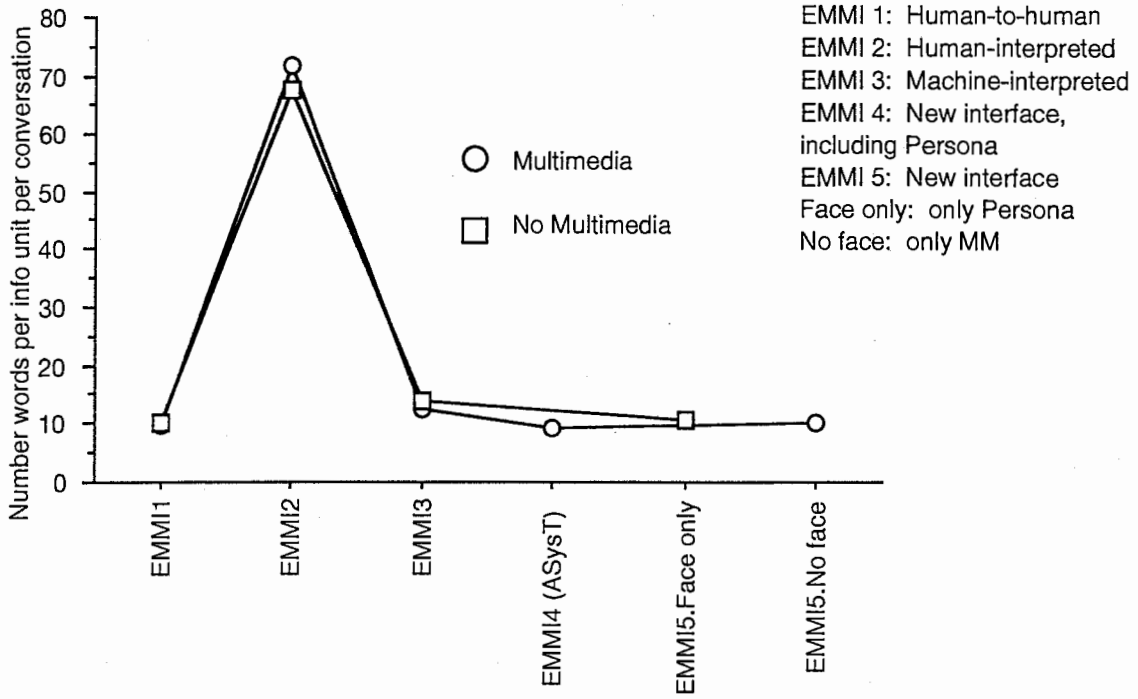


Figure 9. Average number of words-per-information-unit per conversation for each EMMI experiment condition.

And here we turn the picture upside-down. In this case, of course, we are looking for low numbers: in order to be *efficient* in information exchange, participants should use a low number of words per information unit exchanged. Not only was the human-interpreted situation bad for *amount* of information exchange, it was also bad for efficiency. Participants were much more efficient either human-to-human or in any of the machine-interpreted situations. Possibly participants were efficient in the former setting by virtue of habit and familiarity and in the latter through care and design. Note, however, that this time MM options had no effect--apparently, *enjoying* information exchange doesn't imply greater efficiency.

4.9 Attitude

And some of our clearest results in all these experiments were that the participants enjoyed using the MM interface. From the smiles on their faces, to the tossed-off "Wow's" when they saw the system, to the enthusiastic comments on questionnaires, right through to the significant results on the more quantifiable attitude scales we tested them on in the later experiments, it has been clear that participants *like* using multimedia. The significant results we saw in the final experiment were gratifying confirmation of what had been apparent all along; further, they made somewhat more specific the aspects of the system that won approval from users (Fais and Morimoto, submitted). The charts on the following page give the ratings for the final system for all fifteen of the adjective scales. In the first chart, scores for the MM conditions are compared to those for the no MM conditions. In the second chart, scores for the conditions with Persona are compared to scores for those without Persona. The addition of MM had a greater positive effect on users' attitudes than did the addition of a Persona. (Fais and Morimoto, submitted) contains a detailed discussion of the significance of these charts.

Adjective	Com-puter	Persona only	p <...	MM only	p <...	Persona and MM	p <...
forgiving	3.3	**4.3**	**0.04**	**4.9**	**0.03**	**5.2**	**0.07**
systematic	5.9	**5.4** (less)	**0.03**	5.3	0.91	6.0	0.49
cooperative	4.5	**5.4**	**0.02**	5.1	0.53	5.7	0.13
flexible	3.6	3.9	0.89	**4.9**	**0.08**	5.1	0.14
obliging	4.4	4.3	0.81	**5.1**	**0.08**	5.4	0.37
helpful	6.1	5.3	0.28	**5.2**(less)	**0.03**	5.6	0.77
simple	3.7	4.6	0.14	4.5	0.78	3.8	0.72
easy	4.2	4.3	0.46	4.7	0.89	5.3	0.14
obedient	5.3	4.9	0.51	5.5	0.71	5.4	0.89
unthreatening	5.0	5.1	0.81	5.6	0.55	5.3	0.66
pleasing	5.3	4.7	0.22	5.1	0.87	5.6	0.87
satisfying	4.5	3.6	0.34	4.3	0.75	4.6	0.90
calming	3.7	3.6	0.74	4.0	0.84	4.5	0.24
intelligent	4.6	3.7	0.32	4.5	0.76	4.6	0.90
personal	3.9	4.3	0.31	4.6	0.40	4.3	0.79

Adjective	Persona only	p <... <——>	Persona and MM	p <... <——>	MM only
flexible	**3.9**	**0.008**	5.1	0.34	4.9
intelligent	**3.7**	**0.02**	4.6	0.27	4.5
obliging	**4.3**	**0.03**	5.4	0.61	5.1
forgiving	**4.3**	**0.05**	5.2	0.37	4.9
satisfying	**3.6**	**0.05**	4.6	0.39	4.3
pleasing	**4.7**	**0.07**	5.6	0.43	5.1
easy	**4.3**	**0.08**	5.3	0.26	4.7
systematic	5.4 (less)	0.18	6.0	0.15	5.3
cooperative	5.4	0.68	5.7	0.37	5.1
simple	4.6	0.50	3.8	0.55	4.5
obedient	4.9	0.54	5.4	0.68	5.5
unthreatening	5.1	0.44	5.3	0.88	5.6
calming	3.6	0.12	4.5	0.46	4.0
personal	4.3	0.54	4.3	0.88	4.6

4.10 Comfort

In going over the initial results from the first EMMI experiment, I came across a measure that we used for that experiment and then did not use in subsequent experiments. Responding to this notion that users felt more at ease in the MM setting, we tried to quantify what we called at that time "comfort" (Fais and Loken-Kim, 1994). We reasoned that, if users were comfortable in a MM setting, they would take longer to complete their task(s) and use more words. At the time we had only the barest hint of the phenomenon that would be one of the main results of these studies, namely that in fact, in the MM settings, participants *did* use a greater number of words, a result which *was* correlated with greater comfort. When this result became clear (around about EMMI3), we looked for the reason in many places (see Sections 3.1 and 4.1 above). Perhaps it is simply the case that participants used more words because they were comfortable, enjoyed the setting, and simply wanted to prolong the conversation.

One reason we dropped this notion was because it was difficult to defend in any quantifiable way. Perhaps participants use more words in MM conditions because they were having trouble and needed more words to get through the task? Why should greater number of words be indicative of comfort? Despite the fact that the two phenomena were *correlated* (both appearing in MM settings), we had no principled reason for ascribing *cause-and-effect*. So we discarded the notion. Perhaps it is time to bring it back, to make a greater effort to more clearly define the notion of "comfort" in our context.

4.11 Drawing

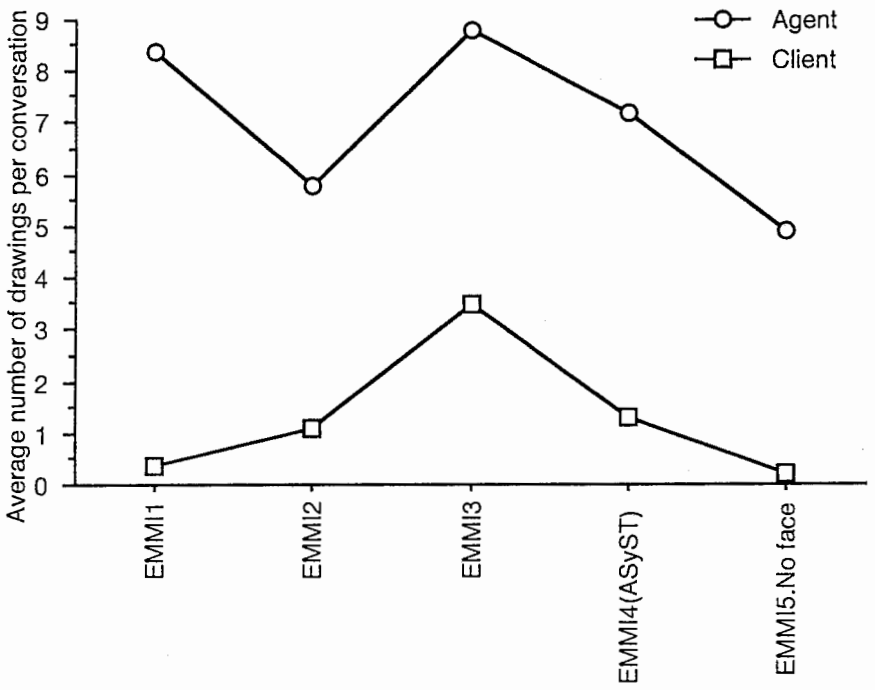


Figure 11. Average number of drawings per conversation for each EMMI MM experiment condition.

In all our concern with manipulating participant linguistic behavior through particular interface design, and with the attitude and information efficiency of participants, let us not forget that the fundamental elements of the interface, what make the interface "multimedia" are the non-speech options available. It is important to look, then, not only at linguistic and attitudinal behavior, but also at what the participants actually do with the media available to them.

In general, not as much as we would have liked. Somehow, with all the attention that the word "multimedia" gets these days, we expected to see participants merrily drawing and typing away, taking full advantage of all that we had made available to them. But in fact, while Agents seemed to find the drawing option useful for *giving* information, Clients found it less useful for actively *receiving*. it.

The high point of Client (and Agent) drawing was in EMMI3. This was the first machine-mediated interface; it contained no online instruction. The later two experiments did; messages like "you can draw on the map with your finger" appeared in print and were output through CHATR the first time the map appeared on the Client's screen. In fact, this sort of stand-alone design was the only major difference between EMMI3 and EMMI5 (the condition without the Persona, which is what is shown in Figure 10). Why then should there be such a marked decline in drawing as the system gets more "polished"?

Possibly, the fact that the system took more initiative in instructing and guiding the Client through the initial stages of the interaction robbed the Client of his own sense of initiative, so that the Client was less willing to "take charge" of the system by actively modifying it with drawing. (A similar feeling of someone else being "in charge" might have accounted for the drop in drawing in the human-interpreted situation, where the Interpreter might have been perceived to be the one guiding the conversation.)

Interestingly enough, the use of drawing was the one area affected by the experience of the participant. Experienced users drew more than inexperienced users, and especially drew more deictic drawings. Perhaps experienced users were not as affected by the appearance that the system was taking charge, as we would not expect them to be.

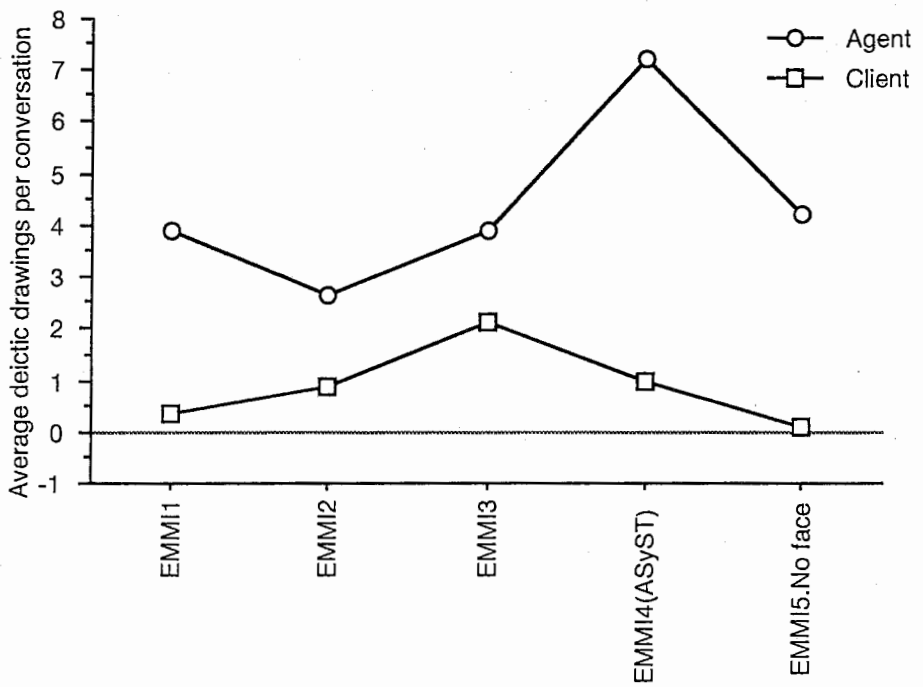


Figure 12. Average number of deictic drawings per conversation for each EMMI MM experiment condition.

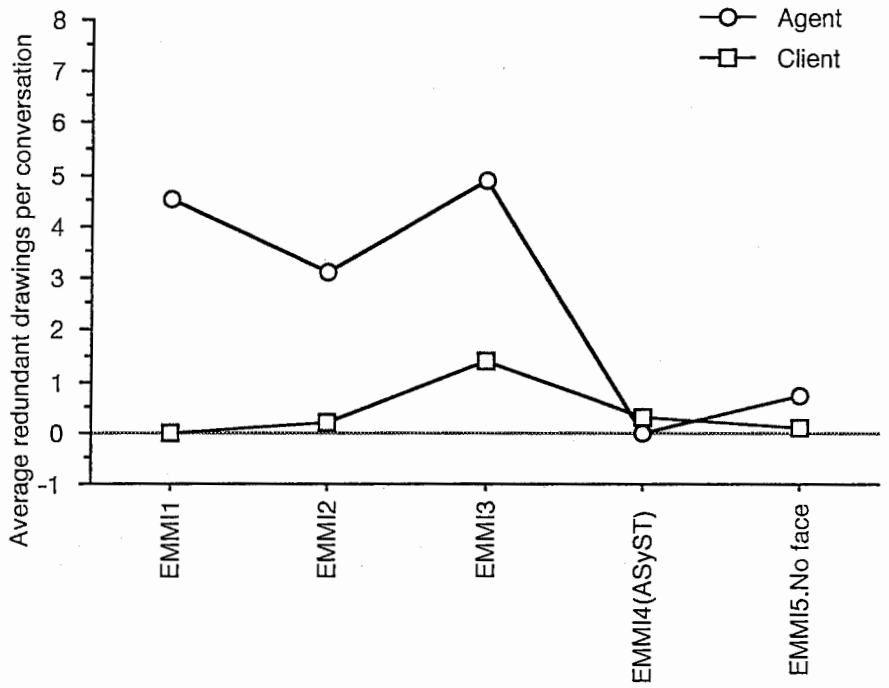


Figure 13. Average number of redundant drawings per conversation for each EMMI MM experiment condition.

The stereotypical expectation of the use of drawing and speech is that the drawer will use the drawing to substitute for some part of the speech. She might draw a route and say "go this way." In fact, we noticed that our Agent, especially, used drawing frequently to supplement rather than to substitute for speech. We did a detailed examination of this phenomenon (Fais and Morimoto, submitted). The graphs above tell the story. The figures for the deictic and redundant drawings for the Agents in EMMI's 4 and 5 were design choices; we had pre-planned most of the Agent utterances, and designed them predominantly as deictic utterances to be accompanied by deictic drawings. However, notice that this behavior differs sharply from the more natural behavior found in EMMI's 1, 2, and 3, in which the Agents' speech was not constrained. In these cases, the Agent uses slightly greater numbers of redundant than deictic drawings. We take this to be the natural case. Given this striking result, we would change the utterances of the Agent in future designs to include far more redundant drawings, in order to better simulate a natural information-giving situation.

Notice that Client drawing behavior did not seem to be affected by the change in the proportions of deictic and redundant drawings in the Agents' interaction. Clients always used slightly more deictic drawings than redundant drawings.

4.12 Typing

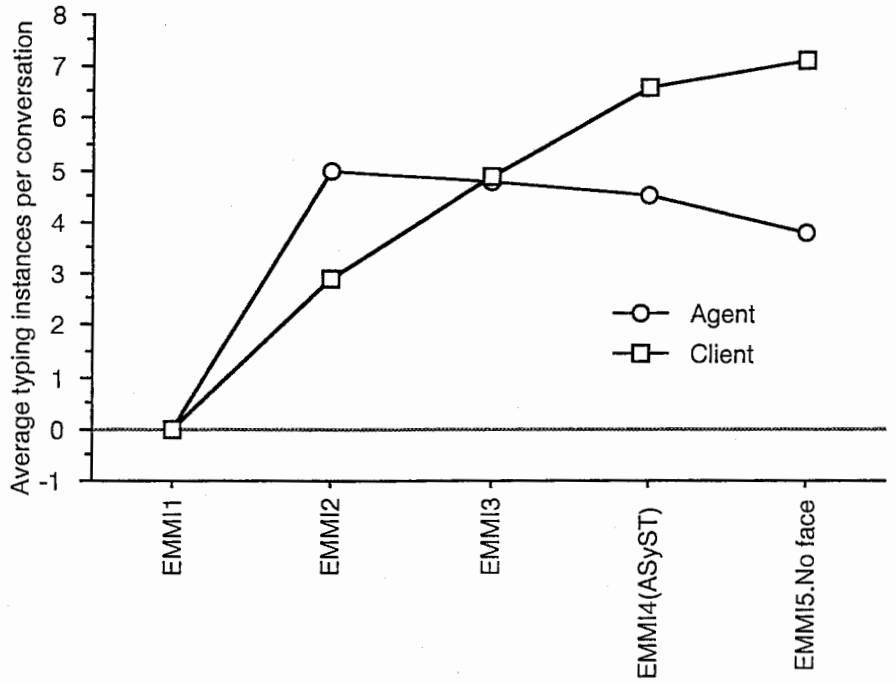


Figure 14. Average number of instances of typing per conversation for each EMMI MM experiment condition.

Results for instances of typing are a striking contrast to those for drawing.[3] Clients tended to let the Agent type in the human-interpreted setting; they typed as much as the Agent did in the less-directed EMMI3 setting. But as the instructions for the use of the system got more explicit (EMMI's 4 and 5), Clients typed more.

Why should instructions to type ("You may type on the form. Touch the slot first, then type.") have had a positive effect on Client typing while instructions for drawing ("You may draw on the map with your finger") had a negative effect? Possibly the answer lies in the fact that, especially in the hotel reservation task, the information to be given belonged to the Client, and had to be given to the Agent. The Client was empowered by the instructions to type in the information belonging to him. On the other hand, the Client had little information to be given to the Agent in the direction-finding task. The Client's use of drawing in that situation centered more around the use of drawing to point out places about which the Client had a question. Since this was not necessarily vital to the task, the Client did less drawing.

Or it could have been a straightforward reaction to the wording of the instructions. "Touch on the slot with your finger; then type," was meant as an instruction concerning how to type. But it is couched in the imperative form, and might have been construed by the Clients as an instruction that they should type. "You can draw on the map with your finger" is simply a statement of possibility and does not imply immediate action on the part of the Client.

In any event, increased typing by the Client is a welcome result, the kind of result that we hoped to achieve with the new, more system-driven design of the interface.

---

[3]Recall that EMMI 1 did not include the hotel reservation task; although Clients in EMMI1 could type messages, they did not. One of the main reasons for typing in later experiments was to give or receive the spellings of names and Japanese words; this was not necessary in EMMI1 and this the typing option was not used.

5. The General Conclusions and Recommendations

In a sense, we have been somewhat disappointed with the results from these experiments because we've been comparing the results from machine interpreted systems to those from human-human studies. We hoped that we could mimic human-human interactions in a human-machine-human setting.

But this sort of goal is ill-conceived. Direct communication is qualitatively different from mediated communication, as a number of the results above have shown. Thus, our benchmark should be, not human-human conversation but human-mediated conversation. If two conversants do not speak the same language and their conversation must be mediated, can we do that as "well" with a machine-interpreter as with a human interpreter?

The answer seems to be a solid "yes." In the human-interpreted situation, our participants exchanged much less information and did it much less efficiently than in the machine-mediated situations. Further, in the latter situations, participants used fewer words, fewer words-per-turn (in one condition), and less overlapping speech, all results that characterize easier speech to process in an automatic language processing system. Thus, machine-mediated situations, and certain configurations of MM in addition, seem to be successfully designed for efficient exchange of information, as compared to human-mediated situations. In addition, as we have seen, users modify their speech in machine-mediated situations in a number of ways that make that speech easier to process automatically.

This describes the results that we have gotten, given our aims and assumptions up to this point. But we would like to suggest a re-thinking of the EMMI interface, in light of the results obtained over the last four years of research, that might make better use of its features and take better advantage of its benefits for mediated communication.

In some sense, research concerning the EMMI interface has been done backwards. The interface was designed before we had a sense of what a good design might be. The tasks were chosen before we had a sense of how people might use the interface to achieve goals.

And so the results have been equivocal, but only from the point of view of the particular design and the particular tasks. That is, if we assume, as we have, that this interface is going to be situated say, in a train station, and that people using it are going to be primarily naive users, then our results indicate that the interface is too complex for them to use while at the same time maintaining fluent speech.[4] In addition, they won't make the best use of the MM options, though it seems they will enjoy using the interface and will get a large amount of information efficiently.

On the other hand, what if we had envisioned this project differently? What if we had thought of it, not as a kiosk in a train station, but as a computer accessory that consumers buy and use attached to their personal or office computers? This changes the scenario quite a bit. Our results show only what first time users of EMMI do with the interface. What if users had a chance to become accustomed to the system? To practice extensively with the media options? Then, I think, we would see far different results. As users become more accustomed to the system, their speech becomes more fluent; their use of media becomes more directed, effective and frequent; and their interactions are accomplished not only with informational efficiency but also with linguistic efficiency. It is exactly the effects of accommodating to the characteristics of the machine and the type of interaction it allows that we did *not* see in the sort of one-shot experiments we ran, based on the scenario for the use of the EMMI interface that was originally envisioned.[5]

It is not apparent what the future of the EMMI interface will be in the context of the current goals of ITL. Certainly, however, the lessons learned in developing the interface

---

[4]Though we haven't tested this with a real speech recognizer, I would guess that this disfluent speech may cause problems in recognition that offset the advantages that follow.

[5]With the exception of the use of the media. We have support already for the claim that as users gain more experience they will use the media options more, derived from results with experienced users, see section 4.11.

are broad ones that are applicable to any system incorporating multimedia options. I would make the following suggestions concerning this broader goal of incorporating multimedia options into interfaces:

1. Keep it simple if the system is to be used by novices who have limited opportunity to learn the system.
2. For naturalness and best transfer of information, full language descriptions should be accompanied by redundant drawings.
3. Have system goals and parameters clearly in mind (is it stand-alone? desktop accessory? for individual use? for use by a trained Agent?) when designing an interface.
4. Similarly, keep in mind the results above concerning which aspect is effective in what area. For example, MM options are good for information exchange, but are bad for disfluency. The Persona is good for meta-media speech, but is not effective for words-per-turn.

Just as we couldn't know how users would behave in our interface before we tested it, we still don't know how users will behave in a system when machine translation lag times are closer to actual human conversation. This area requires investigation if MM options are to be incorporated into a machine translation setting. The same is true of : the effects of the nature of the face and the quality of the voice; the placement of various options, including the persona on the screen; animation of the face; wording of instructions; use of color and font in the maps and forms, etc. The list of possible parameters requiring testing seems endless.

While this may be nice for the continued job prospects of MM researchers, it does not bode well for companies who have to produce a marketable system. And systems incorporating MM options and even Personas are already available. How did they do it?

By creating systems that are interesting and fairly usable, and taking advantage of the fact that most users' multimedia preferences, as we saw so clearly above, are far from formed. Presenting users with a system that is appealing in some way (a cute Persona, flashy graphics), they ensure that the system will be bought, and then *users adapt their behavior to the system.* Should ITL want to produce a marketable product, we already know enough about interface design to follow this course of action.

References

Campbell, W. N., 1995. From read speech to real speech. *Proc. ICPhS,* Stockholm, Sweden, Vol.2, pp. 20-27.

Campbell, Nick, 1996a. CHATR: A high-definition speech re-sequencing system. *Proc ASA/ASJ Joint meeting,* Hawaii, pp. 1223-8.

Campbell, W. N., 1996b. Synthesizing spontaneous speech, in *Computing Prosody,* Sagisaka, Campbell & Higuchi (eds.). NY: Springer Verlag, pp. 165-186.

Fais, Laurel, 1996a. Lexical accommodation in machine-mediated interactions. *Proceedings, COLING-96,* The 16th International Conference on Computational Linguistics, Copenhagen, Denmark, pp. 370-375.

Fais, Laurel, 1996b. Multimedia route descriptions: Experimental results. ATR Technical Report TR-IT-0173. Kyoto, Japan: ATR Interpreting Telecommunications Research Laboratories.

Fais, Laurel, submitted. Lexical accommodation in human- and machine-interpreted dialogues. *International Journal of Human-Computer Studies.*

Fais, Laurel, and Kyung-ho Loken-Kim, 1994. Effects of communicative mode on spontaneous English speech. *Technical Report of the Institute of Electronics, Information, and Communication Engineers,* NLC94-22 (1994-10).

Fais, Laurel, and Kyung-ho Loken-Kim, 1996. How many words is a picture really worth? *Proceedings, ICSLP 96,* International Conference on Spoken Language Processing, Philadelphia, USA, pp. 2179-2181.

Fais, Laurel, Kyung-ho Loken-Kim, and Tsuyoshi Morimoto, 1996. Linguistic and paralinguistic differences between multimodal and telephone-only dialogues in English and Japanese. *Journal of the Acoustical Society of Japan (E)* **17** (5), pp. 229-238.

Fais, Laurel, Kyung-ho Loken-Kim, and Tsuyoshi Morimoto, submitted. How many words is a picture really worth? *Speech Communication.*

Fais, Laurel, Kyung-ho Loken-Kim, and Young-Duk Park, 1995. Speakers' responses to requests for repetition in a multimedia language processing environment. *Proceedings,* International Conference on Cooperative Multimodal Communication, CMC/95, Eindhoven, The Netherlands, pp. 129-143.

Fais, Laurel, Suguru Mizunashi, Kyung-ho Loken-Kim, and Kazuhiko Kurihara, 1996. EMMI progress report: An evaluation of research done with the first EMMI interface. ATR Technical Report TR-IT-0172. Kyoto, Japan: ATR Interpreting Telecommunications Research Laboratories.

Fais, Laurel, and Tsuyoshi Morimoto, submitted. When to put the "face" in "interface:" Determining the effects of multimedia and a persona on communicative behavior. *Journal of Natural Language Engineering.*

Loken-Kim, Kyung-ho, Suguru Mizunashi, Mutsuko Tomokiyo, Laurel Fais, and Tsuyoshi Morimoto, 1995. Analysis and integration of multimodal inputs in interpreting telecommunications. Proceedings, IPSJ Workshop, 1995.

Oviatt, Sharon and R. VanGent, 1996. Error resolution during multimodal human-computer interaction, in H. Fujisaki, ed., *Proc. 1996 Internat.Symp. on Spoken Dialogue*, 2-3 October 1996, pp. 117-120.