音声言語情報処理　9－7
（1995. 12. 14）

How many words is a picture really worth?

ローレル　ファイス　キュンホ　ローケンキム
ATR音声翻訳通信研究所
〒６１９-０２ 京都府相楽郡精華町光台２丁目２番地
（０７７４）９５-１３３０
fais, kyungho@itl.atr.co.jp

あらまし　人間は視覚的な情報を用いて発話を簡略化すると思われる。そこで、マルチメディア環境で用いられる視覚的なイメージと対話の中の単語の数の関係を電話対話とマルチモーダル対話を比較することによって調査した。その結果、予想に反して被験者は視覚的なイメージの存在によってより多くの単語を使用したことがわかった。ここでは、イメージと発声単語数の関係について報告する。

キーワード　マルチメディア、翻訳会話、メータ会話、自動翻訳スツテム

How Many Words is a Picture Really Worth?

Laurel Fais and Kyung-ho Loken-Kim

ATR Interpreting Telecommunications Research Laboratories

2-2 Hikaridai Seika-cho Soraku-gun
Kyoto 619-02
(0774)95-1330
fais, kyungho@itl.atr.co.jp

Abstract　Subjects communicating in telephone and multimodal environments do not replace speech with visual images in the multimodal environment. Instead, they use more words in this environment. We discuss the trade-off between words and images, addressing this surprising result. Results from experiments in multimodal interpreted dialogues indicate that several factors are involved: a high number of redundant visual gestures; "meta-mode" conversation addressing the communication media themselves; and a slightly greater amount of information conveyed in the multimedia setting. Suggestions concerning integration of a multimedia system with automatic translation are made.

key words　multimedia, interpreted conversation, meta-conversation, automatic translation system

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, if the location of a building is to be conveyed, it is generally thought that a visual image such as a map will convey that information better than a verbal description such as "three hundred meters west of the intersection of Elm and Vine, on the north side of the street." A single picture seems to convey at once what a fairly complex grammatical structure takes longer to express.

This first concept plays a major role in a number of systems that are concerned with the automatic construction of multimedia presentations [1, 2, 6]. These systems contain rules which match features of the information to be conveyed with the corresponding capabilities of the various 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 one medium.

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.

It is natural, then, that when we turn from multimodal presentation generation to the *use* of multimodal systems, we assume that humans will operate in much the same way. We might predict 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 in some other way, for example, by typing or speaking the information. We might suppose that users will employ pictures (where that is the appropriate medium) instead of words.

This view is attractive in the field of natural language processing. Fully automatic real-time language processing has proven to be too difficult a goal to pursue for the near future. For that reason, systems designers have turned toward the use of multimedia options as a way to supplement language processing systems. If users in fact do employ the non-speech options available in such integrated systems instead of using (only) speech, that would reduce the amount of language processing necessary. It may then be possible to build a language processing system capable of handling such a reduced load.

The Environment for Multimodal Interaction (EMMI) designed and built at the Advanced Telecommunications Research Institute (Japan) is an example of such an effort. EMMI is a multimodal communication environment in which 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.

The "one picture is worth a thousand words" approach is applicable to the integration of EMMI with real-time, automatic machine translation. If users convey information in modes other than speech, the translation system's job becomes that much easier. That is, if locations and directions are presented visually, and basic personal information is typed in, less information is presented in speech form and the speech translation system will carry less of the burden of communication. This provides strong motivation for an integrated multimedia translation system.

Our initial hypothesis is, then, that the integration of non-speech media in a communication environment will reduce the amount of speech used, when compared to a speech-only (telephone) setting. Below we report on the results of three experiments conducted in EMMI to test this hypothesis. The experiments comprised three different interpretation conditions and two conversational settings: telephone-only and multimedia. By comparing the communicative behavior of subjects across varied interpretation conditions, we were able to assess the contribution made by the use of speech and that made by the use of non-speech media to each condition. By comparing communication behavior in the telephone and multimedia settings, we were also able to determine whether, in fact, the use of non-speech media in some sense *replaces* speech in conveying information, and results in a reduced amount of speech in a multimedia setting.

Methods

In the first of the three experiments, subjects acting as "clients" were instructed that their task was to get directions to a specific place (the site of a conference they were "attending") by engaging in a cooperative dialogue with the "conference agents." In this first experiment, the subjects, both "clients" and "agents," were native speakers of American English, and their interaction was human-human. In two further experiments, native American English-speaking clients interacted with Japanese-speaking agents, both to get directions and to make a hotel reservation[1]. 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;" keep in mind, however, that interpretation was done by trained translators mimicking a computer-based system.). The two settings, which were the same for all three experiments, were via a standard telephone, and via a computer-based, multimedia environment 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]. Descriptive statistics for all three experiments appear in Table I.

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 on video and

[1] For purposes of comparison with the first experiment, the hotel reservation portions of the second and third experiments were not included in this analysis.

	English subjects	Japanese subjects	Task	Turns (#), English	Words (#), English
human/human	12	0	direction-finding	1283	12,342
human-interpreted	9	5	direction-finding; hotel reservation	1233	9,513
machine-interpreted	10	10	direction-finding; hotel reservation	1919	12,636

Table I. Description of the three experiments.

noted on the speech transcriptions to get an approximate idea of the relative timing of these gestures with respect to the speech. In the human-interpreted and machine-interpreted conditions, drawing and typing were also recorded directly from the computer screens [8].

In the analysis of the conversations, we made three measures. First we counted the number of words in each conversation. Second, we observed that in the multimedia conversations, there was a noticeable amount of conversation concerned with the mode of presentation itself (see below). We identified and labeled this kind of conversation. Third, we examined transcripts for "information units." A task analysis of the direction-finding portion of the experiment was made and a list compiled of all the possible "information units" that could be conveyed. Examples of such "units of information" are: location of the client in Kyoto Station; location of the bus stop; length of bus ride; amount of train fare; name of train line, and so on. Each conversation was analyzed to discover how many of those units it contained.

Results

Before we turn to the actual results of the experiments, let us look at what our hypothesis implies the results should be. If the use of non-speech media such as drawing on a map or typing replaces speech, we should find coordinated changes in the amount of information conveyed by speech and that conveyed by visual gestures (i.e., drawing or typing) as illustrated in the hypothetical Figure 1.

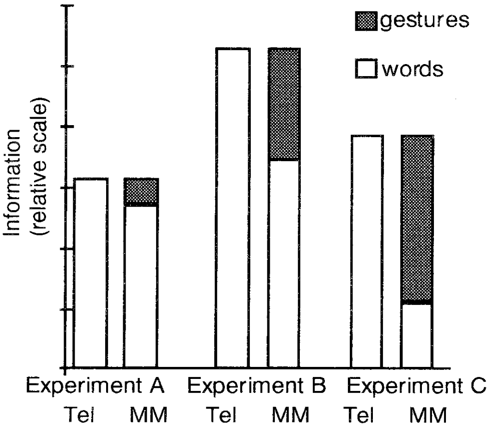

Figure 1. Hypothetical contributions of visual gestures and words in telephone and multimedia settings.

If our hypothesis is correct, the amounts of information conveyed by gesture and by speech should be inversely proportional: as the number of gestures increases, the number of words should decrease.

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

A short digression to explain the construction of this graph is in order. Clearly, Figure 2 implies some notion of the worth of gestures relative to words. Although it is meant as illustrative only, and is not meant to make a claim about the relative weight of gestures and speech, the weights were not assigned completely arbitrarily. Natural language descriptors for a number of the gestures found in our experiments were constructed and the average number of words per descriptor was determined. These were cases of what we call "deictic" gestures in which the gesture was related to a deictic expression in the speech of the gesturer (see below). An example would be "I am standing *here*" accompanied by a mark. This gave us a rough idea of how many words a deictic gesture might replace; it turned out to be eight. However, only about half of the gestures observed were deictic; the rest did not replace speech but instead were redundant. Thus, each gesture was weighted as four words. However, note that this was done for the purpose of illustration only.[2]

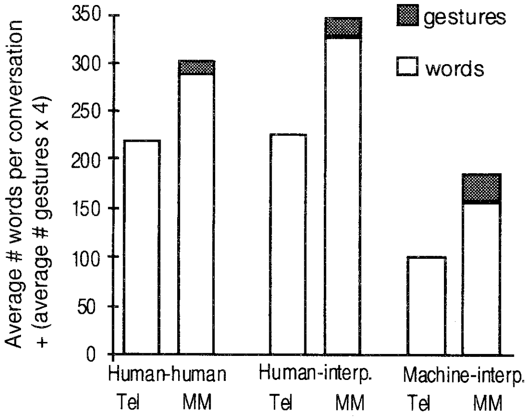

Figure 2. Contributions of visual gesture and words in the telephone and multimedia settings of the three experiments.

[2] Besides, "A picture is worth four words" just doesn't have the right ring.

The actual weight assigned to the gestures used is irrelevant. Whatever the weighting system, the important point to note from this graph 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 gesture made a contribution to the information in the conversation above and beyond this greater number of words. This trend is opposite to what our hypothesis would suggest.

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

1a. Agent: I'm circling the station...

1b. Client: I'd like to do whoop sorry whoop /laugh/ I'm sorry I remember something about you need to go up [uh] it's a little different cause that other one you can go up and /typing/ OK so and return

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

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

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

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

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

4b. Client: 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 meta-conversation like this, what we call "meta-media" conversation, 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; however, there is a significant amount of this kind of conversation in the multimedia mode (Figure 3).

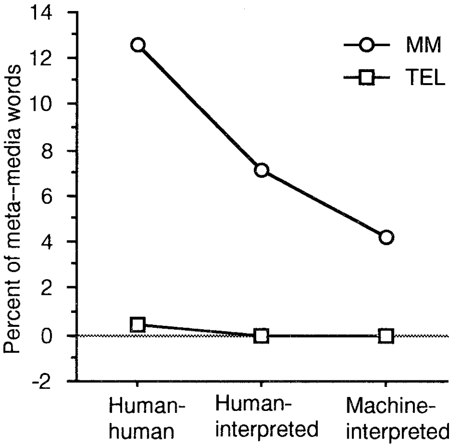

Figure 3. Percent of meta-media conversation in telephone and multimedia settings for all three experiments.

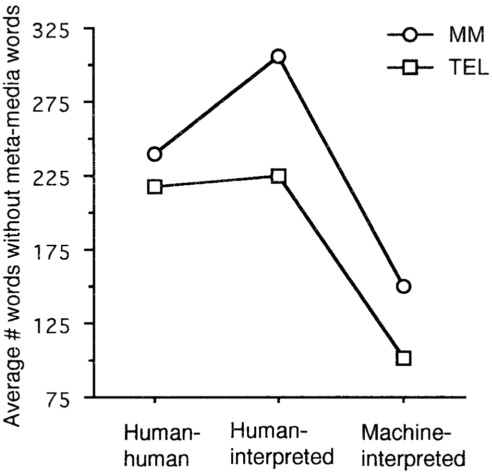

Figure 4. Number of words in telephone and multimedia settings for all three experiments, with meta-media words removed.

We then eliminated the meta-media conversation from our evaluation of the relative contributions of speech and visual gestures. 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 still 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.

Information units. What if clients are simply requesting and receiving more information in the multimedia setting of these conditions? This would have the effect of increasing the number of words used. In fact, there tends to be a higher number of information units in the multimedia setting for all three experiments (Figure 5), although this difference is not significant for any of the experiments. Because of this lack of significance, a greater amount of information cannot explain the higher number of words in the multimedia setting of the two interpreted experiments.

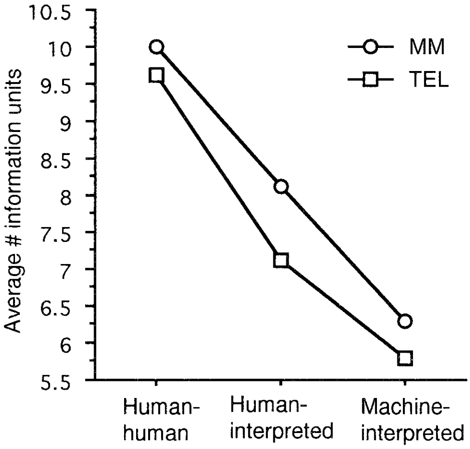

Figure 5. Number of information units in telephone and multimedia settings for all three experiments.

What might be more telling, however, is not the raw number of words used or of information units achieved, but the relationship between the two. Perhaps the number of words per information unit conforms to the expected trend (lower for multimedia; higher for telephone). When we examined the number of words used per information unit, however, we found the by-now familiar pattern: there is a significantly higher number of words used to achieve information units in the multimedia setting across all three experiments (Figure 6). We are faced with the same dilemma: in requesting and receiving information, subjects use more words in the multimedia setting (per information unit) than in the telephone setting.

Relationship of meta-media conversation and information units. We have examined the effects of meta-media conversation and the words-per-information-unit separately. What if we analyze the joint effect of these two factors on subjects' linguistic behavior in the telephone and multimedia settings? When 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 which takes place in the multimedia setting, the numbers of words used per information unit in both multimedia and telephone settings are equivalent in the two experiments (Figure 7).

Discussion

Do we need to re-think the old adage? It would appear so. Comparing experimental results across the telephone and multimedia settings reveals that subjects use about the same number of words to convey information whether they are communicating by telephone or via multimedia. (And this result is achieved only by ignoring the meta-media conversation that typically accompanies dialogue in the multimedia setting.) Subjects do not, in fact, use visual images to replace speech.

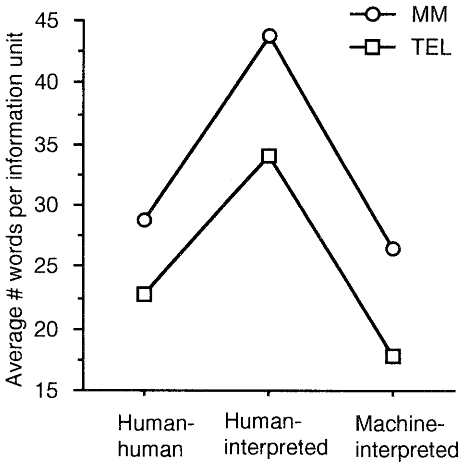

Figure 6. Words per information unit in telephone and multimedia settings for all three experiments.

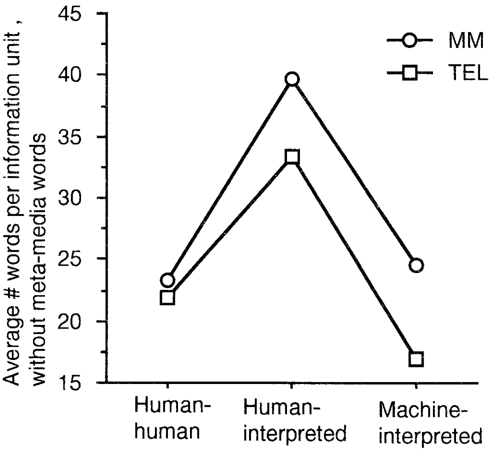

Figure 7. Words per information unit (with meta-media words subtracted out) for telephone and multimedia settings for all three experiments (differences are not significant).

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 drawings made by agent and client in the direction-finding task in these experiments, we found that these visual gestures were of two types. The first type, what we call "deictic" gestures accompanied a deictic expression in the speech of the subject, and did, indeed, convey information about that expression in a visual rather than verbal fashion. These gestures did tend to replace longer verbal descriptions. They accounted for only half the visual gestures used, however. The other half were what we call "redundant" gestures. These accompanied speech that contained no deictic expression and were simply visual correlates of the information that was *also* being expressed verbally. While the use of deictic gestures seems to support the adage, the use of redundant gestures contradicts it. In the latter case, subjects express information verbally despite the fact that the information is available visually.

The experimental design allowed us to examine the balance of speech and visual gestures across interpreting conditions as well. Doing so pointed up some revealing differences. In the human-human interaction, the presence of additional meta-media conversation alone accounted for the greater number of words in the multimedia setting. There was only slightly more information exchanged in that setting and analyzing words-per-information-unit had little effect on the relationship between the results for the telephone and multimedia settings.

However, in the interpreted conditions, the results were different. Note that the use of gestures increased in these settings (Figure 2). At the same time, meta-media conversation itself was not enough to account for the greater number of words; instead, it was necessary to consider the additional amount of information conveyed in the multimedia setting of these conditions. This suggests that in cases where the communication process is complicated by interpretation, subjects make greater use of visual gestures, and convey slightly greater amounts of information. It is only by considering the interaction of meta-media conversation and higher levels of information in the multimedia setting that the words used in that setting and in the telephone setting become equivalent.

These results are illuminated by the results of post-experiment interviews. In these interviews, subjects uniformly reacted in a positive way to the multimedia setting [3, 7]. They cited the presence of the map and the capability of seeing directions marked on the map as having a positive influence on their ability to understand those directions and on their enjoyment of the task. Human beings vary in their ability to absorb information through visual, auditory and tactile channels, and strong visual learners, especially, appreciate the presence of the visual medium in a direction-finding task such as this. Subjects' greater confidence in and enjoyment of the multimedia setting is probably correlated with the increased amount of information conveyed 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.

Future Directions

A number of issues remain to be explored. The dialogue function of redundant gestures has yet to be fully explained. The presence of redundant visual information in multimedia communication may, in fact, contribute to the support of assumptions underlying the mutual beliefs of the conversation participants. 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 gestures may serve the same function. If this is the case, these gestures are performing a function similar to that performed by repeated phrases in dialogue, and as such, are not truly redundant but a necessary part of the process of constructing dialogue.

In addition, we still have not dealt adequately with the phenomenon of meta-media conversation. With some reflection, it is clear that meta-media conversation is, in fact, an integral part of communication in a multimedia setting. Consider everyday experience with multimedia devices. We could think of the use of a microphone by a public speaker to be a limited case of the use of multimedia. We are all aware that the first thing most speakers do when speaking at a microphone is to make some comment on the medium itself: "Is this working?" or "Can you hear me?" 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 efforts to reduce the meta-conversation used by subjects in the multimedia setting by encouraging them to use the non-speech options more effectively. We hope to achieve these results while maintaining the benefits of increased information communication in the multimedia setting found in these experiments. We have already conducted a protocol experiment in which expert multimedia users negotiated the two tasks in the multimedia setting while being videotaped and interviewed. These users commented on problems and suggested improvements in the system and interface, improvements which we hope will make the system more usable, especially for novices. The interface will be more system-driven initially, and it will include more on-line instruction to increase users' confidence in the functioning of the system as well as in their own abilities to operate the system. We are currently in the process of implementing and testing these improvements.

There is also the possibility that experience plays a role in the effective use of non-speech options. Perhaps, for example, experienced public speakers do not have to confirm that their audience can hear them through a microphone. Similarly, perhaps experienced users of multimedia will not show such a high rate of meta-media conversation or of redundant gestures. An experiment comparing the communication behavior of experienced and non-experienced multimedia users in the improved EMMI is planned.

References

1. André, E. and T. Rist. 1993. The design of illustrated documents as a planning task. In *Intelligent Multimedia Interfaces*, M. Mayberry, 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. Mayberry, 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 Institute of Electronics, Information and Communication Engineers, **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 Proceedings of the 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 Interpreting Telecommunications Research Laboratories, Kyoto, Japan.

6. McCaffery, F., M. McTear, and M. Murphy. 1995. Designing a multimedia interface for operators assembling circuit boards. In Proceedings of the 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 Interpreting Telecommunications Research Laboratories, Kyoto, Japan.

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

9. Walker, M. 1992. Redundancy in collaborative dialogue. In Proceedings of Coling92.