Querying Yellow Pages in Natural Language: a Corpus-Based Modelling

Violaine Prince and Didier Pernel

LIMSI-CNRS
P.O. Box 133
F 91403 Orsay Cedex
email: prince@limsi.fr, pernel@limsi.fr

1. Introduction

This paper presents an experimental modelling conducted during the ESPRIT project PLUS (P5254). The project purpose was to provide a prototype for an Information Seeking Man-machine dialogue in unconstrained natural language.

The basic idea was that the system had to be as robust and flexible as possible, allowing queries with some aspects of vagueness. The goal of the system was to demonstrate that a heavy pragmatic setting would help any dialogue system to fulfill its task of understanding the user's intention and offering him/her the best answer. The chosen domain application was Yellow Pages (YP). People who are looking through YP have a particular need. Conducting a whole dialogue with the user relies on a precise model of users' intentions, knowledge and beliefs.

In order to understand how users would behave in front of a computer with intelligent front end capabilities, the PLUS consortium decided to collect corpora in three different languages, English, French and Swedish. We were responsible for the French corpus collection and have thus gathered 46 dialogues with the Wizard of Oz (WOZ) technique, where the subject (a human being) interacts through a terminal with what he/she believes to be a computer system. This particular method was preferred over natural man-man dialogue collection because we wanted to know what would be the features of an interaction in natural language through a keyed input and with the computer as a medium. This preference is discussed in section 3, where we try to compare man-man dialogues with the WOZ man-machine dialogues, in terms of relevance to our particular task. Thus we collected 30 man-man dialogues in a setting very close to the WOZ, and both are described in section 2.

About corpus analysis methods, we considered the number of gathered dialogues not sufficent to extract data with statistical validity. However, we tried to put forth, whenever we could, quantitative aspects as good clues to some emerging properties. A preliminary analysis of the gathered data in terms of relevant variables made us notice several interesting characteristics. A relevant variable is for example, the dialogue length, the type of dialogue opening, closing, criteria marking topic shifting, changes in goals, system/user conflicts in knowledge, etc. These variables are grouped in classes that are explained in section 4. Among the interesting findings, we had, for instance, the following:

(i) Dialogues were rather long, ranging from 7 up to 49 turns. This enhanced the idea that the keyboard was not so obviously an impediment to interaction, as one would have thought in the first place.
(ii) Interactions respected the formal features of natural dialogues: openings (with as many possible variations), thematic developments, formal closings.
(iii) Queries tended first to be awkward in formulation, and then through interaction, began to acquire more characteristics of a man to man conversation. Many dialogues started with a quasi key-word expression and ended with a pure natural language form.
(iv) Several interactions showed multiple topics, complex queries, multiple goals interventions, ambiguities in formulation, and so forth.

Many other more technical aspects appeared during analysis, and enabled us to suggest a dialogue model emerging from data, and not the other way round. This led us to examine

the opportunity of a corpus-based modelling of information seeking man-machine interactions.

Our aim in this contribution is to describe how corpora were collected, and the properties of the analysis method that we designed for these corpora. We do not define here the content of the resulting model, which is going to be addressed in another contribution. We confronted our analysis in terms of dominant variables with model-based methods, promoted by our partners in the PLUS consortium. These methods are such that quantitative aspects were not dealt with, and any dialogue was taken as an example and modelled with an already asserted set of properties. We argued that many of our gathered dialogues were sufficiently unpredictible: they did invalidate some modelling rules, such as thematic completeness (in Grosz and Sidner 1986), proper topic shifting (in Mc Coy and Cheng 1990) and adequate relevance of information provided by users in their interventions (in Sperber and Wilson 1989). We do not claim that *a priori* models are not fit, but that one can find enough exceptions that may impede their systematic application.

We will describe our experiment and its results by first providing the frame of our corpus collection (in section 2) then making some preliminary remarks about salient items (section 3). We will indicate how we did detemine our analysis variables (in section 4). Last , we will conclude about the results of our experiment.

2. Describing Corpus Collection

2.1 User/experimenter communication

A person (called "user") was seated in front of a computer. He/she was told to use a new system of Yellow Pages querying. He/she was also told that he/she was testing a new software understanding sentences in unconstrained natural language. In reality, he/she was linked to an experimenter, who answered by typing on a terminal, and this typing was echoed on the user's terminal. User and experimenter never saw each other. They only communicated via computers. The communication software was developed by Cap Gemini Innovation, PLUS prime coordinator. The experimenter's sentences instantaneously appeared on the user's screen. 46 dialogues were collected thus, where users were supposed to believe dealing with a system alone. We wanted to know the impact of such a belief on dialogue structure and contents. Would people converse differently if they believed they dealt with an automatic system from what they would have said if they believed it was a person to person conversation via a keyed input? Therefore, we collected 30 dialogues with people whom we informed that there was a person at the other end. In order not to introduce technical biases we used the same communication software.

2.2 Dialogue Scripts for Yellow Pages Enquiry

One could issue as many possible requests to a YP application as one has got different trades in a list of all professionals and, in a free interaction, one could ask any question even remotely linked to a YP database true content, that is, a number of headings (ex.: plumbers, doctors, car hiring agencies...) and a number of professionals (names and addresses). To have a good point of comparison between the different dialogues, we gave the users scripts. Experimenters had to know the specific dialogue task that the subject was confronted with: this knowledge allowed experimenters to request relevance precisions, i.e. possible subheadings of the database (e.g. a more precise desired location, a wanted particular subclass of a given class of trades, etc.). We had 3 scripts agreed on: restaurant seeking, car hiring and insurance agents determination. For the French corpus, and because of the database size, we restricted location to Paris only (Paris is subdivided into twenty sectors, and these were considered large enough to play the role of a specific city each!). The following table gives a more precise description of the distribution across scripts and socio-professional groups for the man-machine corpus:

	Car hirer	Insurance	Restaurant	TOTAL
Student	9	6	10	25
Reseach worker	1	3	0	4
Secretary	6	4	7	17
TOTAL	16	12	17	46

2.3 Corpus collection biases and goals

We respected the requirements for WOZ corpora (Fraser and Gilbert 1991). The most important problem was the unreal motivation of the users. We suggested problems to solve. This motivation lack explains some particular dialogues. A limited number of eccentric users tried to investigate the system limitations: they seeked advice from system, they asked questions about the system itself (its location, its design, etc). Nevertheless, many played their part seriously. Another bias was not about corpus collection conditions but about our own point of view concerning gathered data. This corpus was supposed to fill some expectations, that is, we believed that corpus study would provide useful information for test and evaluation tasks. We assumed that corpora were representative enough. One of the limits is that designing the system behaviour so as to make it fit our gathered dialogues was possibly a way from preventing it to be able to accept any kind of dialogue to come. This was the main objection that was made to our corpus-based approach. Our counter- argument is the following: it is true that gathered corpora would constrain the system capabilities, however, one must implement a model flexible enough, and this is possible through an incremental approach. A first corpus serves as a basis for modelling, a second,and this time, real validation corpus, is used to modify the design to make it as efficient as possible. This is feasible provided that one does not implement an over-specified dialogue model. As a consequence, we ran our emerging model on other corpora (in other information seeking applications). A first result (Pernel 1994) showed that the model resisted at least other corpora, gathered with other conditions and requirements.

3. Preliminary results and comments

Some of our observations are interesting enough to be stated as preliminary comments.

3.1 Users' beliefs

Strangely enough, all the WOZ subjects believed that they were 'talking to a computer'. Even though the wizard sometimes used very sophisticated and human-like language, the users accepted this as being an indication of the new advanced prototype system that can understand natural language. On the other hand, in a couple of cases, when the wizard made obvious mistakes like sending the wrong type of information, the users simply regarded these as system errors that commonly occur when one is dealing with computers. Thus, features that professional computer scientists may consider too human-like actually supported the subjects in their belief that it was a computer that they were talking to.

3.2 Users' expectations

Interestingly, the users seemed to be more concerned about quick and exact replies than fluent dialogues. Obviously, because fluent dialogues were something that they did not expect a computer to master; the capabilities of a computer to understand and produce such a human-like language actually surprised them. However, even if the subjects acknowledged the naturalness of the system responses, they regarded the long response time as the main factor in distinguishing their computer conversations from real human-

human conversations. Thus, in building a robust system, considerable amount of attention is to be paid to fast and efficient algorithms.

3.3 Comparing man-man with WOZ dialogues

There was only one difference between what we called man-man (via terminal) and WOZ dialogues: it was the user's belief about the system status. Comparing a true man to man (without terminal) dialogue and a complete WOZ key-board based input was introducing a heavy bias in analysis. If the user was to interact with a machine through a keyboard then the difference in communication mode is such when compared to a vocal human interaction that comparison cannot stand. In our opinion, modelling a man-machine written interaction on the basis of an oral human dialogue analysis is not accurate enough. This is why we restricted the difference between the two collecting methods to the sole psychological variable of belief about the system status. We assumed that this belief could introduce remarkable differences in dialogue structure and contents. More precisely, we thought that, considering that they were 'talking to a machine', subjects would:

(i) make very short sentences, very close to a keyword expression
(ii) make enquiries with a unique goal
(iii) not multiply enquiries (stick to one topic as much as possible)
(iv) not ask for particular precisions
(v) not deal with conceptual knowledge (knowledge about relations between objects and trades and people who perform them and so on)

and we supposed that it was to be the other way round with people who knew that it was a person they were linked with through the terminal. Practically, not only our expectations were largely infirmed (not always, but in numerous cases) in the WOZ dialogues, but also we could not pinpoint any real or subtantial difference between WOZ and human-human (via terminal) dialogues! This enhanced our impression that the WOZ technique, although limited in its possibilities, is reliable enough in this kind of well-defined and rather restricted type of interaction.

4. Corpus analysis

The analysis method we designed for this corpus, which was not large enough to be analysed statistically, is trade-off between quantitative and qualitative methods. We defined criteria as large categories of items. Each item is an assumption about what one is supposed to find in a dialogue (example : a request for information, a confirmation…). Each item is formalised through a relevant variable, i.e. a property that is supposed to be recurrent in different dialogues. Occurrences of a variable were numbered and these values are given in result tables.

4.1 Criteria

The first category deals with dialogue structure. We recognize the following relevant items: dialogue length, dialogue opening and dialogue closing. From this first analysis, we modelled the dialogue structure according to (Schegloff and Sacks 1973) and to (Moeschler 1985). The second analysis axis is the concept of metadialogue:it concerns every statement about the dialogue itself, like evaluating the system's behaviour, expressing satisfaction or discouragement, commenting the system's capabilities. The last part referred to dialogue strategies: changes in goals, change of mind, conflict between system and user knowledge. The mentionned dialogue numbers refer to the corpus text published in (Pernel 1991).

4.1.1 Dialogue Structure

4.1.1.1 Dialogue length

Values about length are given in § 4.2.1.1. Some dialogues are quite long (50 turns): users changed strategy during interaction; they processed several scripts or modified parameters of their main request. In our corpus, long dialogues did often present a violation of thematic completeness (Grosz and Sidner 1986) (i.e. a brutal change of request without waiting for the system's answer) or of the relevance principle (Sperber and Wilson 1989) (providing much unrelevant information for the system's task).

4.1.1.2 Dialogue opening

We defined several kinds of openings (several variables) according to the human dialogue openings. By convention, we will give an abreviation of the variable in capital letters between brackets. This abreviation will be used as a heading in the tables in §4.2. There are two major types of openings:

- formal human explicit openings (FHO), such as "good morning" or "hi" (dialogues n° 9, 13).
- implicit openings: these are characterised by the lack of an explicit marker. They are dispatched in 4 subclasses:
- direct polite opening (DPO): "I would like the address of..." (n° 12).
- direct laconic opening (DLO): "Which insurance compagnies are good for Brasil?" (n° 11).
- personal opening (PO): "I need a car" (n° 6), "I am looking for a car to hire" (n° 1).
- keyword opening (KWO): "car hire" (n° 16, 26) or "insurance" (n° 20).

4.1.1.3 Dialogue closing

In the same way, one may define the following variables:

- formal human explicit closings (FHC), "goodbye" (n° 24, 29), "thank you, bye" (n° 1, 12)
- direct polite closing (DPC): "thank you" (n° 10, 11, 13)
- direct laconic closing (DLC): a negative answer to a demand of a new task (n° 18)
- personal closing (PC): the user expresses his needs, reasons or justification in order not to pursue interaction, such in n° 31.
- brutal stop (BSC): the user leaves the yellow pages system without informing the system, such in n° 20.The system shuts down once (n° 26).

4.1.1.4 Dialogue Structure

We define the dialogue structure on the basis of the adjacent pairs concept (Schegloff and Sacks, 73). The most important point is the incidental pair, called *sub-exchange*. We analyse subexchanges according to the function each of them performs in the dialogue:

- specification of the value of a parameter: "I need a car", "do you want to buy one?", "yes" (n° 6)
- clarification of concept when misunderstanding occurs such as: "are you looking for a mutual or classical insurance?", "what is the difference?", "A mutual insurance chooses people who may join and prices are smaller." (n° 9)
- prerequisite of a question to verify partner's knowledge: "do you know Elysee?", "Yes, it is an official building", "do you know its phone number?" (n° 6)
- precision of a previous answer in order to obtain more information: "here is the list you wish…", "What is price for a Mercedes 190D", "Sorry, I have no information about prices, You should phone" (n° 37)

4.1.2 Metadialogue

Metadialogue is conversation about the running dialogue itself. Its largest part is related to the partners' behaviour and to user's beliefs. We study morphological features, user's behaviour and system's behaviour.

4.1.2.1 Morphological features (MF)

This kind of metadialogue is only issued by the user. It is a question about word spelling. For example, in n° 6, the user asks: "how do you spell Elysée?"

4.1.2.2 User's behaviour

The user's behaviour is expressed by the system (UBS) or by the user himself (UBU). The user states his/her intention about the forthcoming intervention, e.g. "I repeat" (n° 15) or "I want to precise my question" (n° 14). The system may ask the user to give an answer or a request. Examples are given in t n° 9, 14 "You may express your request", or in n° 32 "Say me, I listen to you".

4.1.2.3 System behaviour

Making remarks about the system's behaviour is related to response time. The user (SBU) complains of the slowness: "quicker" (n° 47). The system (SBS) tries to make the user be patient: "have a little more patience, I am looking for it in the database". The other type of system's bahaviour metadialogue is related to the system's capabilities: the system precises its limitations. E.g "I can only process one question at once." (n° 10).

4.1.3 Dialogue Strategy

We recognize four types of dialogue strategies: purpose alteration, reformulation, precision, contradiction / insistance.

4.1.3.1 Purpose alteration

This item gives birth to three variables:

- goal change (GC): it is a change of script or more generally a change of main request (professional) can be marked as: "Now I am looking for a restaurant"[1]. (n° 16) or not: "Do you know the phone number of Elysee ?"[2] (n° 6).
- change of mind (MC): it is characterized by a change of parameters: the user first wanted a mutual insurance company then a classical insurance (n° 9). Dialogue 12 gives us an other example: "I repeat my request, but for a restaurant situated in Champ Elysee."
- subgoal change (SGC). A subgoal is a goal still related to the current request. The difference with the previous kind (GC) is the satisfaction or not of the user's expectation. In this case, the user is never satisfied, by the given answers, therefore he/she changes his/her subgoal in order to receive a satisfactory answer.

4.1.3.2 Reformulation (UR)

Reformulation can be formally marked, such as "I repeat: do you know an attractive restaurant ?" (n°15). When the system is too slow, the user reformulates his/her utterance in order to receive an answer.

4.1.3.3 Precision (SP)

The system asks the user for precision in all the dialogues that we collected. This is caused by the fact that the user does not instantiate the set of task parameters and there are

[1]In this example, goal shifting occurred before the current goal was achieved.
[2]This example respected topic shifting.

too many professionals satisfying the user's request. For exemple, in dialogues 1, 9 and 16, the system says: "Can you tell me where you live?" (location is a good restricting factor).

4.1.3.4 Contradiction / Insistance

Contradicting the system answer (KBC) seldom occurred. Users contested the system knowledge three times. For example, in dialogue 47, the user writes: "the Palais Imperial is Chinese" whereas the system has given it as a Japanese restaurant. Moreofen, users asked for the same professionals again (UI): "do you know a special insurance company for foreign countries" (U_{26} i.e. turn number 26) and "what is the list of special insurance companies for foreign countries" (U_{42}) (n° 13). Last, the system insists by asking the user a confirmation (SI), such as "you wish to rent a car, don't you ?" (n° 10.8).

4.2 Results

Here are the quantitative results for the different variables.

4.2.1 Dialogue Structure

4.2.1.1 Dialogue length

Script \ Value	Minimum	Maximum	Mean
Car Hirer	7	49	22,25
Insurance	8	47	20,00
Restaurant	9	48	22,12

There is no remarkable length difference between the three scripts.

4.2.1.2 Dialogue opening

FHO	DPO	DLO	PO	KWO
5	7	3	15	12

Personnal openings ('I would like to hire a car…') and key-word openings ('Care Hire') were the most frequent. Users hesitated about how th address the system and preferred to go directly to the point.

4.2.1.3 Dialogue closing

FHC	DPC	DLC	PC	KWC
11	13	15	1	2

Conversely, users were impressed about the human-like features of the system and a tendancy for formal and polite closings was enhanced. However, many let the system take the closing initiative and shut dowm their conversation by default (DLC = 15).

4.2.2 Metadialogue

MF	UBU	UBS	SBU	SBS
3	4	6	5	46

There were in total many occurrences for metadialogue variables. The number of system's interventions was caused by the fact that the wizard had to browse a database in order to find answers.

4.2.3 Dialogue strategy

Purpose alteration			Reformu-lation	Precision	Contradiction / insistance		
GC	MC	SGC	UR	SP	KBC	UI	SI
8	13	7	5	46	3	19	2

Users often change their mind and were responsible for many shifts in goals and thus in planning the system's answering task. Notice the high value of user's repetition of his/her goal, as if he/she considered the system as unable to remember his/her request.

Conclusion

In this contribution we tried to present our results in corpus analysis for a Man-machine information seeking dialogue within the frame of a YP query application. The most striking aspects resulting from our experiments could be summarised as following:

(i) the subjects, playing the part of users, were pleased with a system accepting unconstrained natural language input and answering as such. They tended to converse for a rather long time with the device and although they were given scripts, they tended not to feel too restricted in their actions by these scripts.

(ii) The collected dialogues did contain the conventional features characterising human conversations: therefore the forthcoming system has to interpret users' utterances according to their illocutionary function (Searle 1969).

(iii) Principles about topics and relevance constraints in task-oriented conversations were not systematically observed. The forthcoming system has to be flexible about changes of mind and brutal modifications of the user's demand.

The designed variables helped us design a model accounting for the different dialogue strategies (table in §4.2.3), the existence of metadialogue (tables in §4.2.2) and confirm the structural-oriented modelling that we undertook, inspired from (Moeschler 1985).

References

Grosz, B., and Sidner, C. 1986. Attention, intentions and the structures of discourse. *American Journal of Computational Linguistics*, 12(3), 175-204.

McCoy, K., and Cheng, J. 1990. Focus of attention: Constraining what can be said next. In M. Swartout (Ed.), *Natural Language Generation in Artificial Intelligence and Computational Linguistics*. Kluwer Academic Publishers.

Sperber, D.,and Wilson, D.1989 *La pertinence. Communication et cognition*. Paris:Les éditions de Minuit.

Fraser, N., and Gilbert, N. 1991 Simulating speech systems. *Computer Speech and Language*, 5,1. 81-99.

Pernel, D. 1994. Gestion des buts multiples de l'utilisateur dans un dialogue Homme-machine de recherche d'information. PhD Dissertation of the University of Paris XI. Orsay, France.

Schlegoff,E.A. and Sacks, H. 1973. Opening up closings. *Semiotica*, 7,4. 289-327.

Moeschler, J. 1985. *Argumentation et conversation, éléments pour une analyse pragmatique du discours*. Paris: Hatier.

Pernel, D. 1991. *Le Corpus "PLUS"* (Internal Document). LIMSI.

Searle, J. 1969. *Speech Act*s. Cambridge.