An Environment for Dialogue Corpora Collection (ENDIACC)

Zygmunt Vetulani
Adam Mickiewicz University
Department of Computer Linguistics and Artificial Intelligence
ul. Umultowska 87, PL-61614 Poznań , Poland
http://main.amu.edu.pl/~vetulani
vetulani@amu.edu.pl

Abstract

The novelty of the proposal presented in this paper is a free, easily accessible, language independent software platform to provide an experimental setting for written dialogue corpora collection (ENvironment for DIAlogue Corpora Collection). This platform will be particularly well adapted to the collection of corpora of chat-like dialogues with multimodal elements.

1. Introduction

In this paper we present a software platform[1] for dialogue corpora collection (ENDIACC) being a part of our long-term program consisting in development of methodology and tools to design systems with Emulated Language Competence (ELC systems). The ELC systems are those able to communicate interactively with their human users in the human language. The key point of our methodology is creation of an environment for systematic observation of the user (interacting with a "machine"). The methods of acquisition of the initial linguistic knowledge, necessary at the early steps of the design of ELC systems, continue to be systematically implemented for Polish language at the Adam Mickiewicz University. This research program has been outlined in (Vetulani & Marciniak, 2000). The element we focus on in this presentation is an open experimental setting to generate empirical data about NL dialogues in the form of dialogue corpora.

2. Former research

The dialogue corpora has become a subject of increasing interest since the 70ties (cf., e.g., Grosz in (Walker, 1978)). The key idea of the empirical approach within the domain of man-machine communication is enclosed in the following statement by Alphonse Chapanis from the Hopkins University. "If we are to know how to built computers so that they can converse with their human users in simple, human-like terms, we need to know how people naturally communicate with each other" (Chapanis, 1973). The main problem with the corpus-based empirical approach consisted in absence of an easy and inexpensive way of collecting naturally generated dialogue recordings. A (partial) remedy was in designing experiments where quasi-natural dialogues could be recorded. Experiments of new type, called the wizard-of-Oz experiments have been invented. In such experiments the (human) participants were persuaded to be in communication with an ELC system, whereas the ELC system was simulated by some other human participant (hidden). The major problem of these early trials was the lack of appropriate tools to help making simulations representative of natural situations. We address this issue further in this paper where we present a new, free software platform ENDIACC for dialogue corpora acquisition.

2. Identification of current needs

The corpus based methodology in AI, although started already at the beginning of the last quarter of the 20th century, is still considered as valid and fruitful, as shown at the Question Answering Roadmap worked out at the Workshop Question Answering: Strategy and Resources (Maybury, 2002). The objective of this workshop (affiliated to the LREC 2002) was to encourage people (ca. 40 participants) to discuss essential methodological issues concerning the question answering domain. The results were presented by Mark Maybury in his *Report on the Workshop* where the reader can find the list of "characteristics that distinguish QA environments". We quote this list (in a slightly simplified form) below:

- nature of query, including the question form (e.g., keyword(s), phrase(s), full question(s)), the question type (e.g., who, what, when, where, how, why, what-if), and the intention of the question (e.g., request, command, inform),
- level of complexity of the question and answer,
- characteristics of the source(s) and/or supporting corpora (...),
- potential for answer use,
- properties of the domain and/or task (...),
- degree of performance required (...),
- nature of the users (...),
- purposes of the user (...),
- nature of supporting knowledge sources (...)
- reasoning requirements (e.g., inference required for question analysis, answer retrieval, presentation generation),
- degree of multilinguality or crosslinguality (...),
- user model (e.g. stereotypical vs. individualized),
- task model (...),
- type of answers provided (e.g., named entities, phrases, factoid, link to document summary),
- nature of interaction (e.g., user reactivity, mixed initiative, question and answer refinement, answer justification).

[1] The ENDIACC platform has been designed by Zygmunt Vetulani and implemented by Paweł Konieczka in Java 1.4.1. Mr Konieczka participated also as experimenter in corpus collection. The platform is accessible for non-commercial purposes at http://main.amu.edu.pl/~zlisi.

When aiming at a high quality QA system, one has to take into account all these characteristics. If one expects the systems to perform as well as the humans do, then observing humans is the proper way to acquire the necessary knowledge. It is however generally true that direct observation is impossible in most cases (no physical access to real data, high costs of observations, legal problems). A partial solution consists, at least for some of the above listed features, in gathering data through experiments. In the list above, we have marked by **bold** characters those characteristics that can be best provided on the ground of the experiment-based methodology.

3. ENDIACC: an ENvironment for DIAlogue Corpora Collection

In what follows we present the ENDIACC environment (cf. Fig. 1, below) allowing system designers to generate experimental data which may help designing

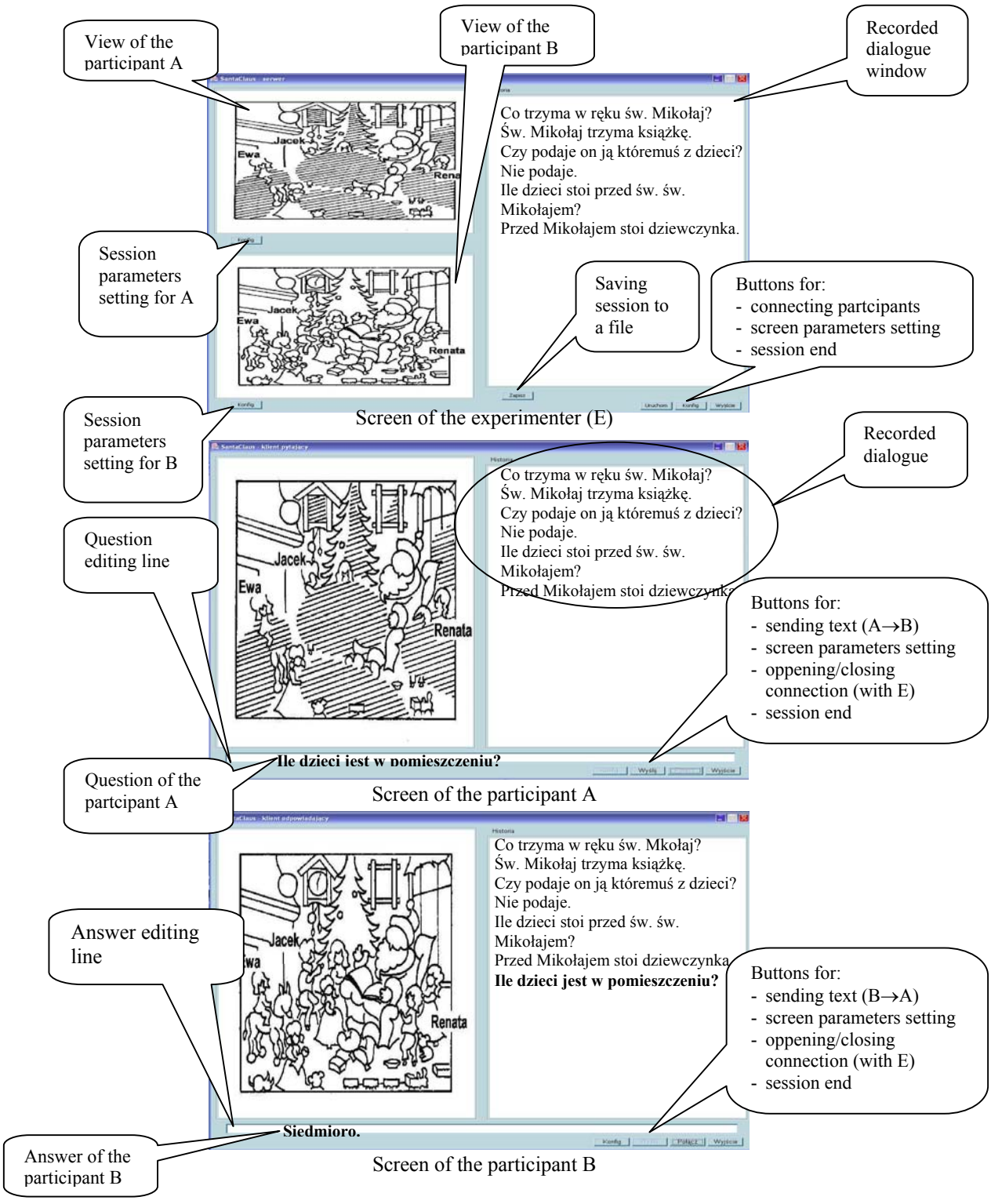


Fig. 1. Screens of the ENDIACC environment (new St Claus experiment)

man-machine dialogue systems (at least at the early designing stage).

The environment ENDIACC is particularly appropriate to support dialogue sessions between two participants (in future one of them may be substituted by a system with emulated language competence) communicating through internet in a classical way, i.e., *chatting* using keyboard. An important feature of the environment is that additional information may be provided to the participants in form of a picture(s) displayed in participants' windows. This gives to the platform a multimodal dimension (cf. access to information in text mode combined with MMS-like technologies). A **typical** usage of the environment consists in arranging sessions between an information provider and an information seeker with the objective for the information seeker to complete his/her knowledge about the dialogue domain. The dialogue is recorded at the session and may be stored in a file. In fact, the recorded dialogue text is considered as the intended program output. (A sample of dialogue record may be seen at Fig. 1 displayed in the "recorded dialogue window".)

The environment is structured as a client-server architecture involving the *experimenter* (working on a *server*) and participants (*clients*). The experimenter is allowed to spy and to record the session. It may set session parameters (e.g. relative to visualisation) and interact with participants (sending messages). In particular, wizard-of-Oz experiments may easily be supported by this architecture. In this case the experimenter may play the role of the system. Also cross-language experiments may be performed.

The ENDIACC platform provides structured screens to the participants and to the experimenter. Participant's screen includes: the editing line (to ask questions or/and give answers), dialogue window where the already recorded text is presented, an additional, graphical window (supporting JPG format) where extra information (typically pictures) may be displayed by the experimenter. The screen of the experimenter is similar but with the graphical windows of both participants displayed. The screens presented in Fig. 1 are doted with several function buttons necessary for session execution or configuration.

The ENDIACC system may work in several language versions (Polish, English, French and German) and may easily be localised to other languages by the user. Also cross-language experiments may be designed.

The system and hardware requirements are easy to satisfy. ENDIACC may easily be installed under any graphical operating system (Windows, Linux) with a Java language interpreter. It requires internet access, otherwise may be installed in any properly configured local network.

4. Corpora collection experiments

In what follows, we will focus on two already performed experiments. We start with reminding our former experiment done in a very traditional way (paper-and-pencil) which may seem technically obsolete. Then we present an exemplary analysis of the data obtained recently in the ENDIACC environment. The results of both of them are similar which is an argument in favour of both techniques.

4.1. The first St Claus corpus

The usefulness of the corpus based research for design of an ELC system was practically verified by our former contributions consisting in design and implementation of human-human interaction experiments and analysis of the resulting corpora (Vetulani, 1989, 2000). One of the best described among them was the so-called St Claus experiment, which consisted in collection of a small but highly annotated corpus of information-acquisition-oriented question-answering dialogues (the St Claus corpus). This corpus contains 582 question-answer pairs collected at 30 sessions with human participants. The questions were collected at sessions involving two participants: the information seeker and the information provider. The information seeker was supposed to formulate written questions (at a sheet of paper) to the information provider about the content of a picture (about an intentionally banal subject: a scene with St. Claus, children, gifts, etc.) presented to the information provider. The information seekers were given a partial knowledge of the scene: the same picture with several blank areas. This very special setting and a particular mode of communication (paper-and-pencil) amounted with a number of observations, which, despite obvious limitations, are of interest especially at the early stage of QA system design (Vetulani, 1989). As the ENDIACC platform provides exactly what is necessary to support the St Claus (paper-and-pencil) experiment we decided to confront results obtained for the St Claus corpus, with the results of the new experiment (screen-keyboard) done under ENDIACC.

4.2. The new St Claus experiment

The restricted length of this paper is the reason why we limited ourselves to comment on one selected parameter among many others observed in the corpus generated by the new version of the St Claus experiment. We will consider the *length* of queries[2] addressed by the information seeker to the information provider. The length of an utterance stands in an obvious relationship with its complexity: short sentences are either simple or elliptical. The arithmetic mean of the sentence length and the distribution of values around the mean is important for the system designer for various engineering decisions as, e.g., the choice of parsing strategy. In order to capture this distribution we use the notion of the *length classes*, each of them consisting of the utterances of the same length. Analysis of the observed numerical data has confirmed the results of the corresponding observations made on the basis of the first, paper-and-pencil St Claus experiments.

In the corpus of 30 dialogues collected under ENDIACC we observed 355 queries (most of them being simple questions) containing 2250 words[3]. The average

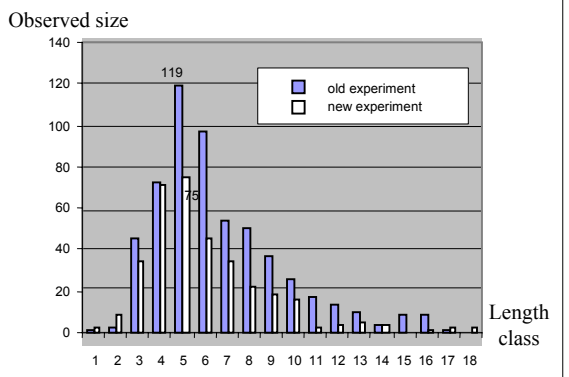
[2] We use the term *query* in a very general sense, meaning all utterances aiming at information acquisition. We omit those which are used only to structure the dialogue, as welcome or stop sentences

[3] By *word* we mean any string of letters separated by blanks (dots, commas,...) and without any blank (dot, comma....) inside.

length was 6.3. The average length calculated for non-elliptical queries (301) was a bit higher and equalled 6.7..Comparison with the average observed in the first St Claus corpus (6.6) has strengthened our hypothesis that, *as a rule*, questions asked by the information seeker tend to be short. We observed 54 elliptical queries in the corpus, i.e., ca. 15% of the total number of queries, distributed in 24 (out of 30) sessions. Ellipsis appears to be a common phenomenon in queries with however relatively small individual impact (only 7 participants used it more then twice). The average of elliptical queries is 4.5 which conforms to the common sense expectation that the omitted sentence element may be recovered using a short, 2-3 word expression.

At the Fig. 2 we present cardinalities of the length classes for the experiment described above (*new St Claus experiment - white blocs*) compared to the length classes observed on the first St Claus experiment[4] (*grey blocs*).

Fig. 2. Frequency classes in St Claus experiments



From the analysis of this diagram we infer that deviations from the observed mean are rather small: 146 (out of 355) queries have length 4 or 5, the 5-th frequency class being the largest one (75). When comparing the old and new St Claus corpora we see that in both cases distribution of length classes is similar and the shape of the frequency diagram resembles the Poisson probability law. We observe that switching from the paper-and-pencil question and answer exchange to the presently very widely practised chatting using keyboard (both techniques being supported with graphical information) does not change some important formal characteristics, as e.g., length distribution, relatively rare elliptical structures, rare relative clauses etc.

(In the first experiment exception was made for "St Claus" considered as one word, here we do not make such exceptions.)

[4] In the first St Claus experiment the observed dialogues were longer then in the new one (582 queries for 30 dialogue sessions vs. 355 now, for the same number of sessions). Therefore the grey blocs at Fig. 2 are respectively higher (max=119 vs. 75, in both cases the respective curves reach their maxima for the length class 5). Remark that the shapes of respective curves are similar.

5. Present and future work

The reason for selecting a very banal experiment theme in the initial investigations was to put the participants in a very familiar, simple situation where they share the common-sense-knowledge. But it is clear the environment ENDIACC may be applied in order to generate dialogues much more typical for the intended real-life man-machine interaction. An exemplary situation being a subject our current investigation is a dialogue between the service provider (like rent-a-flat or selling products) where a graphical information is a natural support of a commercial conversation as an additional information source. Corpus collection is in progress at the time of writing the present paper. In the future an ELC system (derived from POLINT, cf. (Vetulani, 2000)) will be connected to the ENDIACC environment as dialogue "participant".

5. Acknowledgements

The author wishes to thank all experiment participants for offering their time and his Master and PhD students who agreed to participate as experimenters for corpus collection.

References

Chapanis, A. (1973). The communication of factual information through various channels. In: *Information storage retrieval*, vol. 9. Pergamon Press, 215--331.

Maybury, Mark T. (2002): Report on the Workshop on Question Answering: Strategy and Resources, Tuesday May 28, 2002, Palacio de Congreso de Canarias.

Vetulani, Z. (1997). A system for Computer Understanding of Texts. In: Murawski, R. & Pogonowski, J. (Eds.), Euphony and Logos (Poznań Studies in the Philosophy of the Sciences and the Humanities, vol. 57) Rodopi, Amsterdam-Atlanta, 387--416.

Vetulani, Z. (1989). Linguistic problems in the theory of man-machine communication in natural language. A study of consultative question answering dialogues. Empirical approach. Brockmeyer, Bochum.

Vetulani, Z. (2000): Understanding Human Language by Computers: Projects in Artificial Intelligence and Language Technology, in: Yosiho Hamamatsu et al. (eds), Formal Methods and Intelligent Techniques in Control, Decision Making, Multimedia and Robotics. Proceedings of the 2nd International Conference, Polish-Japanese Institute of Information Technology, Warsaw, October 2000, 218--229.

Vetulani, Z., Marciniak, J. (2000). Corpus Based Methodology in the Study and Design of Systems with Emulated Linguistic Competence; in: Dimitris N. Christodoulakis (ed.), Natural Language Processing - NLP 2000, Lecture Notes in Artificial Intelligence, no 1835, Springer, pp. 346--357.

Walker, D. (red.) (1978). Understanding Spoken Language, Elsevier North-Holland, New York.