Erotetic Reasoning Corpus. A data set for research on natural question processing

Paweł Łupkowski[1,2,*], *Mariusz Urbański*[1,2], *Andrzej Wiśniewski*[1], *Wojciech Błądek*[2], *Agata Juska*[2], *Anna Kostrzewa*[2], *Dominika Pankow*[2], *Katarzyna Paluszkiewicz*[1,2], *Oliwia Ignaszak*[2], *Joanna Urbańska*[1], *Natalia Żyluk*[1,2], *Andrzej Gajda*[1,2], *and Bartosz Marciniak*
[1] Institute of Psychology, Adam Mickiewicz University, Poznań
[2] Reasoning Research Group, Adam Mickiewicz University, Poznań

ABSTRACT

The aim of this paper is to present the Erotetic Reasoning Corpus (ERC) which constitutes a data set for research on natural question processing. We describe the theoretical background, linguistic data and tags used for the annotation process. We also discuss the potential areas in which the ERC can be exploited.

Keywords: questions, logic of questions, question processing, erotetic reasoning, corpus annotation

1 INTRODUCTION

The aim of this paper is to present a data set for research on natural question processing named the Erotetic Reasoning Corpus (hereafter ERC).[1] In discourse, interlocutors must deal with question processing in instances when questions are not followed by answers but by new questions or strategies of reducing said questions into auxiliary ques-

[*] P. Łupkowski, M. Urbański and A. Wiśniewski designed the ERC and data-collection process, super-annotated the corpus and wrote the paper. W. Błądek, A. Juska, A. Kostrzewa and D. Pankow annotated the ERC. K. Paluszkiewicz, O. Ignaszak, N. Żyluk and J. Urbańska contributed to the linguistic data collection. A. Gajda and B. Marciniak implemented parts of the ERC interface.

[1] The term 'erotetic' stems from Greek 'erotema' meaning 'question'. The logic of question is sometimes called erotetic logic. For an overview of logically oriented approaches to questions and questioning see, e.g., Harrah (2002), or Wiśniewski (2015).

tions[2]. Usually, such a situation takes place when an agent wants to solve a certain problem (expressed in the form of an initial question) but is not able to reach the solution using his/her own information resources. Thus, new data, collected via questioning are necessary. This phenomenon is studied within such theoretical frameworks as Inferential Erotetic Logic (see Wiśniewski 1995, 2013, Łupkowski 2016), inquisitive semantics (see Groenendijk and Roelofsen 2011), or KoS (see Ginzburg 2012, Łupkowski and Ginzburg 2013, 2016). Natural question processing also constitutes an interesting subject for empirical research. In order to facilitate research concerning question processing in natural language dialogues, we have decided to construct the ERC. The corpus consists of the linguistic data collected in our previous studies on the question processing phenomenon. The data are annotated with a tagset, making them easy to browse for reasoning structure, pragmatic features used, and the presence of normative erotetic concepts (see Section 2).

The paper is structured as follows. We start by presenting the basic concepts of natural question processing as modelled in Inferential Erotetic Logic. We use these concepts as a normative yardstick for our design choices for the ERC tag set. Afterwards, we describe the architecture of the ERC and the linguistic data used for the corpus. Then, we introduce the tagging schema designed and used for the ERC, describe the tagging process, and discuss selected issues concerning annotation reliability. We conclude with a summary of the current stage of the project and discussion of potential future developments and applications of the ERC.

2 MODELLING QUESTION PROCESSING IN INFERENTIAL EROTETIC LOGIC

In this section, we present the underlying erotetic logic concepts used for the ERC. Our logical framework of choice is that of the Inferential Erotetic Logic (IEL; see Wiśniewski 1995, 2013). This logic focuses on inferences whose premises and/or conclusions are questions (erotetic inferences). This choice was motivated by several factors. Here, we

[2]For more details see https://intquestpro.wordpress.com/.

only mention some of them – for a detailed discussion see Urbański *et al.* (2016a). Firstly, IEL is flexible: it is not tied up to any specific logic of declaratives. Secondly, the formal representation of questions employed in IEL is friendly to the user. In general, these representations fall under the schema $?\Theta$, where Θ is an object-language expression that is equiform to a metalanguage expression which denotes the set of direct answers to a question. For example, $?\{A_1, ..., A_n\}$ represents a question whose set of direct answers is the finite set of declarative formulas: $\{A_1, \ldots, A_n\}$.[3] Yet, questions are object-language expressions of a strictly defined form and have meanings on their own; the approach is still a non-reductionistic one (see Belnap 1986; Wiśniewski 1995, pp. 37–42). On the other hand, this approach inherits the advantages of the so-called *set-of-answers methodology* (Harrah 2002; see Peliš 2016, for a comprehensive introduction, and Wiśniewski 2013, pp. 16–17 for a discussion of the semi-reductionistic approach sketched above), whose idea stems from Hamblin's (1958, p. 162) postulate: "Knowing what counts as an answer is equivalent to knowing the question." Thirdly, IEL offers some straightforward tools for modelling erotetic inferences. What is especially important from our perspective is that IEL proposes some criteria for the validity of erotetic inferences. In the case of erotetic inferences which lead from an initial question and a (possibly empty) set of declarative premises to a question, the following criteria of validity are proposed:

1. *transmission of truth/soundness into soundness*: if the initial question is sound (i.e., there exists a true direct answer to this question) and all the declarative premises, if there are any, are true, then the question which is the conclusion must be sound;
2. *cognitive usefulness*: each direct answer to a question which is the conclusion is useful in answering the initial question by narrowing down the "space of possibilities" offered by the initial question (more precisely: for each direct answer B to the question which is the conclusion there exists a non-empty proper subset Y of the set of direct answers to the initial question such that Y must contain a true direct answer to the initial question if B is true and the declarative premises, if there are any, are true).

[3] Thus $A_1, ..., A_n$ are pairwise syntactically distinct formulas.

Valid erotetic inferences (of the above kind) can be defined as those in which *erotetic implication* (e-implication for short) holds between the initial question, the declarative premises, and the question which is the conclusion. As a matter of fact, the formal definition of e-implication offers precise explications for conditions of transmission of truth/soundness into soundness and of cognitive usefulness (Definition 1; see Wiśniewski 2013, p. 68). For the sake of simplicity, we consider here only questions with finite sets of direct answers, and assume that the underlying logic of declaratives is Classical Logic. Given this, erotetic implication can be defined as follows.

Definition 1 (Erotetic implication)**.** *A question Q e-implies a question Q_1 on the basis of a set X of declaratives ($Im(Q, X, Q_1)$) iff:*

1. *for each direct answer A to the question Q: $X \cup \{A\}$ entails a disjunction of all the direct answers to the question Q_1, and*
2. *for each direct answer B to the question Q_1 there exists a non-empty proper subset Y of the set of direct answers to the question Q such that $X \cup \{B\}$ entails a disjunction of all the elements of Y.*

It is easily seen that clauses (1) and (2) of Definition 1 mirror the criteria of validity discussed above.

Applying erotetic implication for modelling certain real-life linguistic phenomena resulted in identifying two other versions of this kind of relation, weaker than the one just defined (which we shall further on call the canonical erotetic implication). These are the weak erotetic implication (Urbański *et al.* 2016a) and the falsificationist erotetic implication (Grobler 2012; Wiśniewski 2013), both of which modify the second condition of the original definition.

Definition 2 (Weak erotetic implication)**.** *A question Q weakly e-implies a question Q_1 on the basis of a set X of declaratives ($Im_w(Q, X, Q_1)$) iff:*

1. *for each direct answer A to the question Q: $X \cup \{A\}$ entails a disjunction of all the direct answers to the implied question Q_1, and*
2. *for some direct answer B to the question Q_1 there exists a non-empty proper subset Y of the set of direct answers to the question Q such that $X \cup \{B\}$ entails a disjunction of all the elements of Y.*

Definition 3 (Falsificationist erotetic implication)**.** *A question Q f-implies a question Q_1 on the basis of a set X of declaratives ($Im_f(Q, X, Q_1)$) iff:*

1. *for each direct answer A to the question Q: $X \cup \{A\}$ entails a disjunction of all the direct answers to the question Q_1, and*
2. *for some direct answer B to the question Q_1, $X \cup \{B\}$ eliminates at least one direct answer to Q.*

The concept of elimination used in Definition 3 is construed as follows: a formula A *eliminates* a formula B just in case B must be false if A is true, given the underlying semantics (for a precise definition see Wiśniewski 2013, p. 34).

The properties described in the second clauses of definitions 1, 2, and 3 will be referred to below as 'usefulness', 'w-usefulness', and 'f-usefulness', respectively.

Table 1 presents examples of erotetic implication of the three presented types.

Q, X, Q_1	e-implication
$?\{p, q \vee r\}, \emptyset, ?\{p, q, r\}$	Im
$?p, p \leftrightarrow q, ?q$	Im
$?\{\neg p, r, s\}, \emptyset, ?\{p, q, \neg q\}$	Im_f
$?\{p, q, v\}, s \rightarrow p, ?\{s, \neg s\}$	Im_w
$?\{\neg p, r, s\}, \neg p \vee r \vee s, ?\{p, q, \neg q\}$	Im_w, Im_f
$?\{p, q, w\}, p \vee q \rightarrow r, p \vee q \vee w, ?\{r, \neg r\}$	Im_w, Im_f
$?\{p, q, v\}, p \vee q, r \leftrightarrow q, ?\{r, \neg r\}$	Im, Im_w, Im_f

Table 1:
Examples of canonical (Im), weak (Im_w) and falsificationist (Im_f) erotetic implication

Notions introduced in this section will be reflected by the tagset used to annotate the ERC, described in detail in Section 4 of the present paper.

Using e-implication as a tool allows for modelling many aspects of natural question processing, i.e. a situation in which an initial question is internally processed by an agent, and where the outcome is either a new question concerning the subject matter or a strategy of reducing the initial question into auxiliary questions. In both cases, e-implication allows for the description and assessment of the inferences which lead from questions to questions.

The basic areas of applicability of the analysis of the described phenomena include: the search for information in distributed resources, question answering (in particular, cooperative answering), problem solving (in particular, problem solving by interrogation), proof theory and automated deduction (proof search, complexity issues).

3 LINGUISTIC DATA

The linguistic data used for the ERC were gathered for research on question processing. The outcomes of three research projects are employed here. These are: the Erotetic Reasoning Test, QuestGen and Mind Maze.

The Erotetic Reasoning Test (in Polish: Test Rozumowań Erotetycznych, TRE) is a tool used in the research described in detail in (Urbański *et al.* 2016a). The test contains 3 items (with an imposed time limit of 30 min). Each item consists of a detective-like story in which the initial problem and evidence gained are indicated. The task is to pick a question (one out of four), each answer to which will lead to some solution to the initial problem. The subjects are asked to justify their choices.

Let us present here an exemplary tasks from TRE (translated into English). The task is entitled "The Bomb":

> In the capital of a certain country someone planted a bomb in the palace of the king. The best royal engineer, who arrived immediately, established the following facts:
> 1. There are three wires in the bomb: green, red and orange;
> 2. To disarm the bomb either the green or the red wire must be cut. Cutting the wrong wire will cause an explosion;
> 3. If the bomb has been planted by Steve, cutting the green wire will disarm it;
> 4. If the bomb has been planted by John, cutting the red wire will disarm it. Moreover, no one but John would have used the red wire;
> 5. If the bomb has not been planted on an even day of the month, the culprit is Steve;
> 6. The bomb has been planted either by Steve, or by John, or by someone else.

> Each of the following questions below can be answered either 'yes' or 'no'. Mark the question to which the answer (regardless of it being 'yes' or 'no') will allow you to establish, in the shortest time possible, which wire should be cut in order to disarm the bomb:
>
> > Was the bomb planted on an even day of the month?
> >
> > Was the bomb planted by Steve?
> >
> > Was the bomb planted by John?
> >
> > Was the bomb planted by someone else than Steve or John?
>
> Justify your choice.

TRE-entries of the ERC have a well-established structure: there is a story, a question chosen by the subject and then a justification of the choice. An exemplary justification (translated into English) provided by a subject for the "Bomb" story is presented below (see Urbański *et al.* 2016a, p. 41).

> If we'll get an affirmative answer to this question, then we'll know that the green wire needs to be cut. If a negative one, then there will be only one possibility left – the red wire, and additionally we'll know that the culprit is John.

QuestGen is an online game the aim of which is to engage players in generating a large collection of questions for a certain piece of story written in a natural language (as such it might be perceived as an example of a game with a purpose – see Von Ahn and Dabbish 2008). The idea of the game was presented in (Łupkowski 2011), while its implementation is described in (Łupkowski and Wietrzycka 2015) and (Łupkowski and Ignaszak 2017). In the game, two randomly chosen players are engaged in solving a detective puzzle. One of them plays as the Detective, the other as the Informer. The Detective's objective is to solve the presented puzzle by questioning the Informer. Each story in the game has two versions (one for the Detective and one for the Informer), containing all the additional data necessary to solve the puzzle. The Detective is allowed to use only yes/no questions and cannot ask straightforwardly for the solution. The Detective may ask as

many questions as s/he wants/needs (as long as they are simple yes-no questions). The Informer is obliged to answer the Detective's questions in accordance with the information presented in the Informer's part of the story. Each story is played within a time limit. The game is played in cooperative mode, i.e. the Detective and the Informer play together constrained by the time limit and obtain points for each puzzle solved.

As an example of the task from the QuestGen game, we present the Detective's part of a story entitled "Arsen L.":

> Imagine that you are a detective who is following the well-known international villain Arsen L. You are trying to establish if Arsen L. went to Paris, London, Kiev, or Moscow. You look through your notes and this is the information you have managed to gather so far:
> 1. Arsen L. left for Paris or London if and only if he departed in the morning;
> 2. Arsen L. left for Kiev or Moscow if and only if he departed in the evening;
> 3. If Arsen L. took a train, then he did not leave for London or Moscow;
> 4. If Arsen L. left for Paris or Kiev, then he took a train.
>
> So, where did Arsen L. go?
> Before you answer this question you may ask several auxiliary questions of the railway station employee. Remember: your time is limited. Ask only yes/no questions. It is pointless to ask the employee directly about where Arsen L. went because he does not have a clue.

Solutions gathered within the QuestGen project have a well-established structure, very much like the ones from TRE. A QG-entry of the ERC consists of the story which is followed by the main question (expressing the problem to be solved by the player). Afterwards, we observe the sequence of the Detective's questions and the Informer's answers which is ended by the proposed solution to the main question and the feedback given by the Informer. This gives us more interaction than in the TRE case. We observe short dialogues between players. An

example (translated into English) of such a dialogue for the "Bomb" story is presented below (see Łupkowski and Ignaszak 2017, p. 239):

DETECTIVE: Is it the case that Anthony has something to do with de bomb?
INFORMER: No.
DETECTIVE: So it is the case that Roger is guilty?!
INFORMER: Yes.
DETECTIVE: Orange, isn't it?
INFORMER: Yes.
DETECTIVE: Orange.

Mind Maze (in Polish "Takie życie") is a card game published by Igrology. In the game, one of the players plays the role of the game master (GM) and the other one tries to solve a puzzle presented by the game master. the GM tells a short story (inspired by true events) and the objective of the player is to figure out how the story happened by asking questions to the GM. Only yes/no questions are allowed here (with two additional admissible answers: "It is not important/relevant" and "It is not known"). *Mind Maze* was used as the core element for the semi-structured study of question processing (see Urbański and Żyluk 2016 and Urbański *et al.* 2016a). The researcher played the role of the GM and subjects were players. Game sessions were recorded and then transcribed.

To give an example (translated into English) of the types of problems to solve in the Mind Maze game, let us consider the one entitled "The Traveller":

> A man without a single visa visited eight different countries in a single day. None of the authorities of these countries tried to remove him. What was his profession and how did he manage to do this?

Solutions gathered in the described study are the most complex ones in the ERC data set. They have no clear structure as they are more or less free dialogue leading to the solution of the initial problem. The shortest conversation included in the ERC has 760 words, while

Table 2:
Characteristics of the linguistic data set of ERC

Source	Files	Words
TRE	270	81.169
QG	116	21.944
TZ	16	30.619
Sum	402	133.732

the longest one is 3.367.[4] An example *Mind Maze* interaction between the player and the game master (translated into English) is presented below:

PLAYER: Is this building a cultural one?
GM: Cultural one... in what sense it is a cultural building?
PLAYER: Related to culture, history, art? Related to culture?
GM: But, how would you define this „related"?
PLAYER: Related... it is used for cultural purposes, developement related issues, for people. To some extent educational ones?

To differentiate the aforementioned sources, we will refer to them as the ERC sub-corpora, the TRE, QG, and TZ, respectively. The whole ERC consists of 402 files (solutions). Table 2 presents a summary of the gathered data. Note that all of the data are in Polish; however, the tagset used for the annotation allows for the data to be analyzed by English-speaking researchers.

4 TAGGING

The tagging schema for the ERC consists of three layers:

1. The structural layer – representing the structure of the tasks used for the studies described in Section 3. Here, we distinguish between elements such as: instructions, justifications, different types of questions, and declaratives.
2. The inferential layer – which allows for normative elements described in Section 2 to be identified.
3. The pragmatic layer – representing various events that may occur in the dialogue, like e.g. long pauses. It also contains tags that

[4] For comparison, the longest files for the TRE and the QG have 387 and 230 words respectively.

enable the expression of certain events related to the types of tasks used (like e.g. when a forbidden question – that is, question of the form which is not allowed in a certain entry – is used).

Let us now present, and explain in detail, the tags used in the ERC. Each task in the ERC is tagged with the KORPUS tag which has two obligatory attributes:

- first one specifying the sub-corpus of ERC (namely whether the task comes from Erotetic Reasoning Test: TRE, QuestGen: QG or Mind Maze: TZ),
- second one specifying the name of the task and the number of the subject/player who solved it.

4.1 *Structural layer*

The structural layer of annotation consists of the following tags: INSTRUCTION; JUSTIFICATION; DECLARATIVE; QUESTION.

- The INSTRUCTION: the tag indicates instruction for a given task.
- The JUSTIFICATION: a justification given by a subject is indicated with this tag.
- The DECLARATIVE: tag marking declaratives.
- The QUESTION: tag for indicating questions.

The DECLARATIVE and QUESTION tags enable certain attributes to specify further details. These attributes are presented in Figure 1 and 2. Pointing out one of the attributes marked with a solid line is obligatory. The ones marked with a dashed line are non-obligatory.

The QUESTION tag is associated with the following attributes:

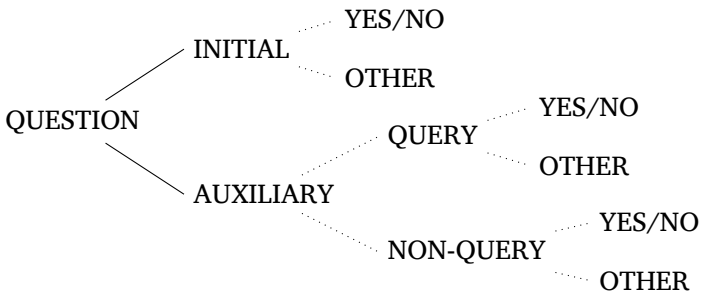


Figure 1: The QUESTION tag and its attributes

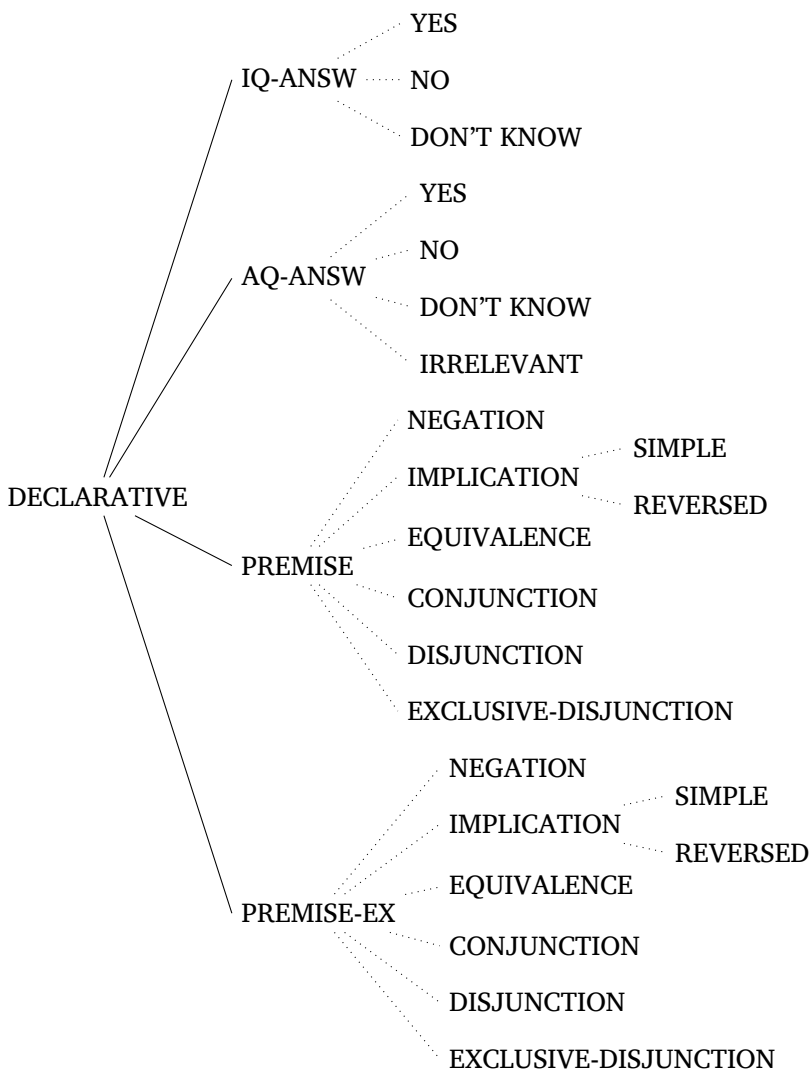


Figure 2: The DECLARATIVE tag and its attributes

1. INITIAL: points out the initial question. Additional attributes allow for specifying whether the initial question is of the yes/no or other type.
2. AUXILIARY: marks questions recognized as auxiliary ones. Attributes associated with the tag indicate whether the auxiliary question is a query and point to its type (yes/no or other type of question).

The DECLARATIVE tag is associated with the following attributes:

- IQ-ANSW: indicates an answer to the initial question. The type of answer given might be specified by: YES, NO, DON'T KNOW.
- AQ-ANSW: indicates an answer to the auxiliary question. Similarly to the IQ-ANSW case, the type of answer given might be further specified by: YES, NO, DON'T KNOW, IRRELEVANT.
- PREMISE: used for premises (declarative ones). Additional attributes may be used to specify a logical structure of the recognized premise. For the premises with the implication as the main connective a more detailed characteristics may be provided with the tags: SIMPLE or REVERSED.
- PREMISE-EX: used for a declarative premise which allows for exceptions. To exemplify such a premise, consider the following (from "The Party" task of TRE): "The King of Hearts stays till the end of only those parties at which the March Hare doesn't tell jokes (although even then the King sometimes leaves earlier)."

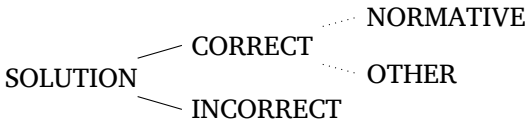
Additional attributes for these tags are the same as those for the PREMISE tag.

4.2 *Inferential layer*

The inferential layer consists of nine tags: SOLUTION; TRANSMISSION; USEFULNESS; W-USEFULNESS; F-USEFULNESS; E-OTHER; ENTAILMENT; D-OTHER; IMP-ERROR. This layer plays an important role in the ERC making our data set unique. The tags used here stem from the IEL's ideas and concepts presented in Section 2. This layer makes it possible to track and study how these concepts are applied and used in the context of reasonings enforced by the tasks used for our subcorpora.

SOLUTION: this tag indicates the solution given by a subject. Additional attributes allow for specifying whether the solution is correct (note that each task in the ERC has a predefined normative solution) and how this solution has been reached (i.e. whether it is in line with the assumed normative way of obtaining the solution – e.g. erotetic search scenario in the case of QG tasks). Attributes of the SOLUTION tag are presented in Figure 3.

Figure 3:
The SOLUTION tag and its attributes



TRANSMISSION: this tag is used for such justifications that cover the first condition of the definition of erotetic implication, i.e. transmission of truth/soundness (including the canonical one as well as the weak one and the falsificationist one – see Definitions 2 and 3 in Section 2).

USEFULNESS: this tag is used for such justifications that cover the second condition of the definition of (canonical) erotetic implication, i.e. cognitive usefulness.

W-USEFULNESS: this tag is used for such justifications that cover the second condition of the definition of the weak erotetic implication.

F-USEFULNESS: this tag is used for such justifications that cover the second condition of the definition of the falsificationist erotetic implication

E-OTHER: marks such justifications that are not modelled by Inferential Erotetic Logic.

ENTAILMENT: this tag is used for such justifications that correctly refer to logical entailment.

D-OTHER: this tag is used for such justifications that incorrectly refer to logical entailment or to a different type of relation between declaratives.

IMP-ERROR: denotes justifications in which a subject interpreted the material implication in the incorrect way (according to Classical Logic).

4.3 *Pragmatic layer*

The pragmatic level consists of the five tags. It should be noted that certain pragmatic layer tags are used only within selected sub-corpora as described below.

Q-FORBIDDEN: allows one to point out when a forbidden question appears in the solution of tasks in the QG and TZ subcorpora. This refers to the rules provided for a given task. For example, this tag is used in the case of a QuestGen task when the Detective will ask directly

about the solution. In the Mind Maze tasks, this tag appears when a player uses a question other than that of a yes/no type.

WRONGINFO: this tag is used in the QG sub-corpus. It denotes a situation wherein the Informer provides a wrong piece of information to the Detective in the game. "Wrong", in this case, means different than the one given in the Informer's part of the story. This tag will also be used in situations in which the Detective asks a question marked as Q-FORBIDDEN and the Informer answers with something different than the desired "I don't know" answer.

KEY-INFO: is used for the TZ sub-corpus. It indicates additional information provided by the game master (the information provided is not an answer to a question in the game).

TOPIC: is also a tag used in the TZ sub-corpus for marking topics (as defined by van Kuppevelt (1995)) as they appear in a dialogue.

LONG-PAUSE: the tag is used in the QG and TZ sub-corpora for indicating long pauses in the game.

An example annotated ERC file is presented in Figure 4. The figure presents the file from the TRE sub-corpus of the ERC, the task name is "Bomb" and the file number is 31 – this is visible in the first line containing the tag <KORPUS A1 = "TRE" A2 = "Bomba31">. The structure of the file is clearly visible owing to the structural layer of the tags used. We can identify the instruction part as well as the premises and the initial question, solution, and justification provided by the subject in this case. Tags used to annotate premises provide information about their structure (visible as the A2 attribute), e.g. in the last premise, an exclusive disjunction is used. The initial question is identified by a <QUESTION> tag with the A1 attribute stating "INITIAL". The A2 attribute informs us that this is not a simple yes/no question. Let us now take a closer look at the solution, which is indicated by the following tag: <SOLUTION A1 = "CORRECT" A2 = "NORMATIVE">. Attributes of this tag inform us that the solution provided by the subject is the correct one, what is more, it is also normative. This leads us to the justification part of this file. There we find two tags: <TRANSMISSION /> and <USEFULNESS />, which provide information about the normativity of the provided correct solution – this

```
<KORPUS A1="TRE" A2="Bomba31">

<INSTRUCTION>
 Wprowadzenie: W stolicy pewnego kraju ktoś podłożył bombę w pałacu króla. Najlepszy saper
     królewski, który przybył na miejsce, ustalił sobie jedynie znanymi sposobami kilka faktów:

(a) <DECLARATIVE A1="PREMISE" A2="CONJUNCTION" A4="1">W bombie znajdują się trzy kabelki:
    zielony, czerwony i pomarańczowy.</DECLARATIVE>
(b) <DECLARATIVE A1="PREMISE" A2="EXCLUSIVE-DISJUNCTION" A4="2">Żeby unieszkodliwić bombę trzeba
    przeciąć albo zielony, albo czerwony kabelek. Przecięcie niewłaściwego kabelka spowoduje
    wybuch.</DECLARATIVE>
(c) <DECLARATIVE A1="PREMISE" A2="IMPLICATION" A3="SIMPLE" A4="3">Jeżeli bombę podłożył Stefan,
    to unieszkodliwia ją przecięcie zielonego kabelka.</DECLARATIVE>
(d) <DECLARATIVE A1="PREMISE" A2="EQUIVALENCE" A4="4">Jeżeli bombę podłożył Ignacy, to
    unieszkodliwia ją przecięcie czerwonego kabelka. Co więcej, nikt inny do tego celu nie
    wykorzystałby czerwonego kabelka.</DECLARATIVE>
(e) <DECLARATIVE A1="PREMISE" A2="IMPLICATION" A3="SIMPLE" A4="5">Jeśli bomby nie podłożono w
    dzień parzysty, to zrobił to Stefan.</DECLARATIVE>
(f) <DECLARATIVE A1="PREMISE" A2="EXCLUSIVE-DISJUNCTION" A4="6">Bombę podłożył albo Stefan, albo
    Ignacy, albo jeszcze ktoś inny.</DECLARATIVE>

Instrukcja: Na każde z poniższych pytań można uzyskać jedną z dwóch odpowiedzi:
'tak' albo 'nie'. Zaznacz symbolem 'x' tylko jedno pytanie, na które dowolna
odpowiedź (niezależnie od tego, czy będzie to 'tak' czy 'nie') pozwoli jak najszybciej
ustalić, <QUESTION A1="INITIAL" A2="OTHER">który kabelek należy przeciąć, żeby unieszkodliwić
    bombę.</QUESTION>

[ ] <QUESTION A1="AUXILIARY" A2="QUERY" A3="YES/NO" A4="1">Czy bombę podłożono w dzień parzysty?
    </QUESTION>
[ ] <QUESTION A1="AUXILIARY" A2="QUERY" A3="YES/NO" A4="2">Czy bombę podłożył Stefan?</QUESTION>
[x] <SOLUTION A1="CORRECT" A2="NORMATIVE"><QUESTION A1="AUXILIARY" A2="QUERY" A3="YES/NO"
    A4="3">Czy bombę podłożył Ignacy?</QUESTION></SOLUTION>
[ ] <QUESTION A1="AUXILIARY" A2="QUERY" A3="YES/NO" A4="4">Czy bombę podłożył ktoś inny niż
    Stefan lub Ignacy?</QUESTION>

Uzasadnij, dlaczego wybrałaś/wybrałeś to właśnie pytanie.
</INSTRUCTION>

<JUSTIFICATION>
<TRANSMISSION />
<USEFULNESS />
Bomba Stefana może mieć zielony Bomba Ignacego – kabel czerwony i zielony Bomba kogoś innego –
    kabel zielony Jeśli dowiemy się, że to Ignacy to trzeba będzie przeciąć czerwony, gdy okaże
    się, że to nie on, to w każdym innym przypadku będzie to kabel zielony niezależnie czy to
    Stefan cze ktoś inny podłożył bombę.
</JUSTIFICATION>
</KORPUS>
```

Figure 4: An exemplary annotated ERC file

warrants the conclusion that the solution provided can be modelled in terms of canonical erotetic implication (see Definition 1).

4.4 *Descriptive statistics of the annotation*

Let us now take a closer look at the descriptive statistics of the ERC annotation.

We will start with the *structural layer* of the annotation. The number of INSTRUCTION tags is the same as the number of ERC files, as each task comes with its own instruction. We have 402 INSTRUCTION

tags (270 for TRE, 116 for QG and 16 for TZ). As for the JUSTIFICATION tag, it is present only in the TRE sub-corpus and the number of these tags is equivalent to the number of TRE files in the ERC, i.e. 270. The reason for this is that each TRE solution consists of an auxiliary question indicated a subject and a justification provided for this choice (as described in Section 3). The ERC has 2.234 QUESTION tags, 1.350 in TRE sub-corpus, 375 in the QG and 527 in the TZ. Details are presented in Table 3. As for DECLARATIVE tags, there are 2.855 (TRE: 1.530, QG: 777, TZ: 548) – details are presented in Table 4.

	TRE	QG	TZ	Sum
QUESTION	1.335	357	527	2.234
INITIAL	270	116	16	402
INITIAL YES/NO	0	19	0	19
INITIAL OTHER	270	97	16	383
AUXILIARY	1.080	241	511	1.832
QUERY	1.080	238	452	1.770
QUERY YES/NO	1.080	238	442	1.760
QUERY OTHER	0	0	10	10
NON-QUERY	0	3	59	62
NON-QUERY YES/NO	0	3	13	16
NON-QUERY OTHER	0	0	46	46

Table 3: Descriptive statistics for the QUESTION tag

For the *inferential layer* we will first discuss the SOLUTION tag. The detailed numbers for this tag are presented in Table 5. The total number of occurances of the SOLUTION tag for the TZ sub-corpus is larger than the number of files. This is because the solution is divided into two parts for each file, corresponding to the dialogue structure. It should be noted that the vast majority of solutions for the ERC tasks were correct ones. (For the TZ sub-corpus NORMATIVE and OTHER attributes were not used).

For the TRE sub-corpus, additional inferential tags were also used. This is due to the structure of the solutions provided by the subjects, i.e. answers to initial questions and their corresponding justifications. There are 205 TRANSMISSION and 160 USEFULNESS tags used. For 149 cases the TRANSMISSION and USEFULNESS tags are both present, which constitutes the number of correct and normative solutions for the sub-corpus.

Table 4: Descriptive statistics for the DECLARATIVE tag

	TRE	QG	TZ	Sum
DECLARATIVE	1.530	777	548	2.855
IQ-ANSWER	0	109	11	120
YES	0	5	0	5
NO	0	10	0	10
DON'T KNOW	0	1	0	1
AQ-ANSWER	0	241	500	741
YES	0	109	191	300
NO	0	120	216	336
DON'T KNOW	0	12	21	33
IRRELEVANT	0	0	25	25
PREMISE	1.350	427	36	1.813
IMPLICATION	720	271	0	991
EQUIVALENCE	180	96	0	276
CONJUNCTION	90	0	0	90
EXCLUSIVE-DISJ	270	20	0	290

Table 5: Descriptive statistics for the SOLUTION tag

	TRE	QG	TZ	Sum
SOLUTION	268	109	17	394
CORRECT	190	91	17	298
CORRECT NORMATIVE	149	44	–	192
CORRECT OTHER	41	47	–	88
INCORRECT	78	18	0	94

Let us now discuss the *pragmatic layer* of annotation. As can be expected, there are no pragmatic tags in the ERC sub-corpus, due to the nature of the task involved. The numbers for this layer will get bigger for sub-corpora with more interaction involved. And we have 8 Q-FORBIDDEN and 29 WRONGINFO tags for the QG sub-corpus. As it was described above, the WRONGINFO tag is specific to the QG sub-corpus. The reason why this is the case for these tasks is that a randomly chosen player has to play the role of the informer in the game. S/he has to process additional information related to the puzzle and provide answers to the Detective within the specified time limit. As a result, we sometimes observe that the Informer provides wrong information. It is important to mark these utterances in the ERC, as this makes solving the puzzle harder or sometimes impossible

for the Detective. In the TZ sub-corpus, we observe more pragmatic tags, as here we are dealing with (almost) free dialogue. There are 16 Q-FORBIDDEN, 61 KEY-INFO, 438 TOPIC and 100 LONG-PAUSE tags used for these tasks. In the TZ context, especially, KEY-INFO and TOPIC are interesting as they were designed especially for this sub-corpus. TOPIC allows one to track how new topics related to the solution of a given story are introduced and resolved. As for the KEY-INFO tag, it is crucial for understanding how the solution to the initial question is reached as this tag indicates situations in which a game-master provides addition information, which facilitates the solving process.

To sum up, we observe 24 Q-FORBIDDEN, 29 WRONGINFO, 61 KEY-INFO, 438 TOPIC, and 100 LONG-PAUSE pragmatic layer tags in the ERC data. As we have mentioned, due to the nature of the tasks, these tags are present only in the QG and TZ sub-corpora of the ERC.

4.5 *Annotation and its reliability*

The tagging process was performed by 5 volunteers with solid background in erotetic logic. Each file was tagged by one annotator. What is more, each annotator tagged files only from one sub-corpus of the ERC. Thanks to this, s/he dealt with a consistent file structure and consistent subset of the tagset.

Annotation quality was ensured via a variety of measures. First of all, the structural tags layer is very intuitive and standardised for the TRE and QG sub-corpora (see description in Section 3). For these files, an experienced super-annotator (with expert knowledge in IEL) prepared and controlled the annotation schemas used. Each controversial case was discussed by the annotators.

Secondly, the output consists of XML files, thus RELAX NG XML schema was defined with the purpose of facilitating the annotation process. The schema specifies a pattern for the structure and the content of XML files and prevents incorrect use of tags by annotators. All of the ERC files were validated by the annotators themselves and afterwards by a super-annotator. The validation was performed in two steps: first general XML validity was checked and in the second step ERC XML schema were used to control the use of the ERC tagset. Structural validity was also checked within the ERC tools described below.

Thirdly, all of the ERC files were thoroughly controlled by the super-annotator. Every issue has been discussed between the annotators; and this is how final tagging was established.

In order to check the reliability of the annotation process, inter- and intra-annotator tests were performed.

For the inter-annotator test, a sample of 100 randomly chosen text units (retrieved from all three sub-corpora of ERC) was used. The units were chosen in such a way that they could be annotated with at least one ERC tag. The structure of the sample was the same as the whole ERC, i.e. 67% of units were retrieved from the TRE sub-corpus; 29% from the QG and 16% from the TZ. All of the units were supplemented with a necessary context.

The guideline for annotators contained explanations of all the ERC tags and examples of annotated text units. The control sample was annotated by two annotators (two logicians, one of whom had a solid background in the logic of question).

The reliability of the annotation was evaluated using κ (Carletta 1996), established by using the R statistical software (R Core Team 2013; version 3.3.1) with the *irr* package (Gamer *et al.* 2012). The interpretation of the kappa values is based on that of Viera and Garrett (2005).

The Fleiss κ for all three annotators was 0.8 (i.e. substantial) with 75% agreement over 100 cases. The agreements between the main annotation and others were high, as presented below:

- main and first annotator: $\kappa = 0.85$, with 86% agreement (almost perfect agreement);
- main and second annotator: $\kappa = 0.78$, with 80% agreement (substantial).

As can be expected, when it comes to a detailed analysis of the annotation, the most unproblematic cases were the ones annotated with tags from the structural and pragmatic layers of the ERC tagset. Annotation with the use of the inferential layer was more problematic. Cases where we observe disagreement between annotators concern the use of <TRANSMISSION /> and <USEFULNESS /> tags for the TRE sub-corpus samples. The reason for this may be that the use of these tags involves the interpretation of the justification provided by a subject in the light of an answer given for a particular task. As it

was explained above, we have paid special attention to this layer of annotation of the ERC. All of the tags used were checked by the super-annotator and each controversial case was discussed by the main ERC annotators.

We have also performed intra-annotator agreement rating test. For this test, another control sample of 100 examples was randomly chosen from the data (with the same structure as the sample for the inter-annotation study). In this case, two ERC annotators were employed to annotate the sample. The agreement between the main annotation and the two annotators was almost perfect – Fleiss $\kappa = 0.86$ with 82% agreement over 100 cases. The detailed results for annotators are presented below:

- main and first annotator: $\kappa = 0.87$, with 88% agreement;
- main and second annotator: $\kappa = 0.85$, with 86% agreement;
- first and second annotator: $\kappa = 0.86$, with 87% agreement.

5 ERC ON-LINE

The corpus is available via its web-site[5]. ERC is distributed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Several tools that allow one to to work with the corpus are provided on the ERC web-site.[6] The central tool is *ERC Search & Browse Tool*. This application allows one to display and browse ERC files, both with and without tags. It also allows one to search through corpus files. Keyword and tag search options have likewise been made available to users. In order to use a certain fragment of the ERC in one's paper, presentation, or poster one may take advantage of the ERC XML/LaTeX Parser (Gajda and Łupkowski 2016). The parser transforms original XML-annotated ERC files into appropriate LaTeX files. The parser is responsible for formatting and displaying the data form the corpus – it will be especially useful for preparing papers and presentations based on the ERC data. Hence the choice of using LaTeX as the output format for our tool. Obtained files may be simply pasted into an article,

[5] See https://ercorpus.wordpress.com/

[6] See https://ercorpus.wordpress.com/tools/.

presentation, or poster.[7] The last tool provided is *ERC XML Schema*. The ERC XML Schema describes the structure of corpus XML files. It allows for quick syntactic validation of corpus files and is very useful in the annotation process.

6 SUMMARY

In this paper, we have presented the Erotetic Reasoning Corpus. So far, the ERC data have been mainly analysed in the light of the normative yardstick provided by IEL. Urbański *et al.* (2016a) present research on correlations between the level of fluid intelligence and fluencies in two kinds of deductions: simple (syllogistic reasoning) and difficult ones (erotetic reasoning). The tool used to investigate erotetic reasoning is the Erotetic Reasoning Test. The paper presents the detailed analysis of the justifications provided by subjects. Urbański *et al.* (2016b) contains analyses of solutions to Mind Maze games. Łupkowski and Ignaszak (2017) model and discuss selected solutions of QuestGen tasks with focusing on normative vs. non-normative solutions.

In our opinion, however the ERC's potential scope of use is broad and reaches far beyond studies of the normative logical concepts vs. instances of real erotetic reasoning. The ERC consists of a significant amount of natural language data (see Table 2). The potential applications may cover the following example areas of interests:

- linguistic studies of the way questions are formulated in different contexts;
- research on dialogue management (this applies in particular to the TZ sub-corpus of the TRE, which consists of long natural language dialogues);
- problem solving studies concerning strategies of handling question decomposition, especially those with imposed time limits (such as the tasks in the QG sub-corpus of the ERC);
- studies focusing on the way a question should be asked (or an initial problem/task should be formulated) in order to make the solution easier to reach.

[7] For an overview of LaTeX in academic use see e.g. (de Souza e Silva Filho and Pinheiro 2010), (Flom 2005), (Hofert and Kohm 2010), (Łupkowski 2015), (Łupkowski and Urbański 2013).

ACKNOWLEDGEMENTS

Work on the Erotetic Reasoning Corpus was supported by the National Science Centre, Poland (DEC-2013/10/E/HS1/00172 and DEC-2012/04/A/HS1/00715).

REFERENCES

Nuel Belnap (1986), Approaches to the semantics of questions in natural language: part 1, in *From models to modules*, pp. 257–284, Ablex Publishing Corp.

Jean Carletta (1996), Assessing Agreement on Classification Tasks: The Kappa Statistic, *Computational Linguistics*, 22(2):249–254.

Paulo Rogério de Souza e Silva Filho and Rian Gabriel Santos Pinheiro (2010), Design and Preparation of Effective Scientific Posters using LaTeX, *The PracTEX Journal*, 2010(2), http://tug.org/pracjourn/2010-2/rogerio.html.

Peter Flom (2005), LaTeX for academics and researchers who (think they) don't need it, *The PracTEX Journal*, 2005(4), http://tug.org/pracjourn/2005-4/flom/flom.pdf.

Andrzej Gajda and Paweł Łupkowski (2016), Using LaTeX as an element of the Erotetic Reasoning Corpus interface, in Tomasz Przechlewski, Karl Berry, and Jerzy Ludwichowski, editors, *BachoTeX 2016: Convergence*, pp. 47–52, Polish TeX Users Group GUST, Bachotek.

M. Gamer, J. Lemon, and I.F.P. Singh (2012), irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84, http://CRAN.R-project.org/package=irr.

Jonathan Ginzburg (2012), *The Interactive Stance: Meaning for Conversation*, Oxford University Press, Oxford.

Adam Grobler (2012), Fifth part of the definition of knowledge, *Philosophica*, 86:33–50.

Jeroen Groenendijk and Floris Roelofsen (2011), Compliance, in Alain Lecomte and Samuel Tronçon, editors, *Ludics, Dialogue and Interaction*, pp. 161–173, Springer-Verlag, Berlin Heidelberg.

C. L. Hamblin (1958), Questions, *The Australasian Journal of Philosophy*, 36:159–168.

David Harrah (2002), The Logic of Questions, in D. M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic, Second Edition*, pp. 1–60, Kluwer, Dordrecht/Boston/London.

Marius Hofert and Markus Kohm (2010), Scientific Presentations with LaTeX, *The PracTEX Journal*, 2010(2), http://tug.org/pracjourn/2010-2/hofert.html.

Paweł Łupkowski (2011), Human computation—how people solve difficult AI problems (having fun doing it), *Homo Ludens*, 3(1):81–94, ISSN 2080–4555.

Paweł Łupkowski (2015), Making your researcher's life easier. How to prepare transparent and dynamic research reports with LaTeX, in Tomasz Przechlewski, Karl Berry, Bogusław Jackowski, and Jerzy Ludwichowski, editors, *BachoTeX 2015: various faces of typography*, pp. 42–48, Polish TeX Users Group GUST, Bachotek.

Paweł Łupkowski (2016), *Logic of Questions in the Wild. Inferential Erotetic Logic in Information Seeking Dialogue Modelling*, College Publications, London.

Paweł Łupkowski and Jonathan Ginzburg (2013), A corpus-based taxonomy of question responses, in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, pp. 354–361, Association for Computational Linguistics, Potsdam, Germany, http://www.aclweb.org/anthology/W13-0209.

Paweł Łupkowski and Jonathan Ginzburg (2016), Query Responses, *Journal of Language Modelling*, 4(2):245–293.

Paweł Łupkowski and Olivia Ignaszak (2017), Inferential Erotetic Logic in Modelling of Cooperative Problem Solving Involving Questions in the QuestGen Game, *Organon F*, 24(2):214–244, http://www.klemens.sav.sk/fiusav/doc/organon/2017/2/214-244.pdf.

Paweł Łupkowski and Mariusz Urbański (2013), Preparing for scientific conferences with LaTeX: A short practical how-to, *TUGboat*, 34(2):184–189.

Paweł Łupkowski and Patrycja Wietrzycka (2015), Gamification for Question Processing Research—the QuestGen Game, *Homo Ludens*, 7(1):161–171.

Michal Peliš (2016), *Inferences with Ignorance: Logics of Questions (Inferential Erotetic Logic & Erotetic Epistemic Logic)*, Acta Universtitatis Carolinae – Philosophica et Historica, Karolinum, Praha.

R Core Team (2013), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/, acess 20.03.2017.

Mariusz Urbański, Katarzyna Paluszkiewicz, and Joanna Urbańska (2016a), Erotetic Problem Solving: From Real Data to Formal Models. An Analysis of Solutions to Erotetic Reasoning Test Task, in Fabio Paglieri, Laura Bonetti, and Silvia Fellett, editors, *The Psychology of Argument: Cognitive Approaches to Argumentation and Persuasion*, pp. 33–46, College Publications, London.

Mariusz Urbański and Natalia Żyluk (2016), Sets of situations, topics, and question relevance, Technical report, AMU Institute of Psychology.

Mariusz URBAŃSKI, Natalia ŻYLUK, Katarzyna PALUSZKIEWICZ, and Joanna URBAŃSKA (2016b), A Formal Model of Erotetic Reasoning in Solving Some what Ill-Defined Problems, in D. MOHAMMED and M. LEWIŃSKI, editors, *Argumentation and Reasoned Action Proceedings of the 1st European Conference on Argumentation*, pp. 973–983, College Publications, London.

Jan VAN KUPPEVELT (1995), Discourse structure, topicality and questioning, *Journal of Linguistics*, 31:109–147.

Anthony J. VIERA and Joanne M. GARRETT (2005), Understanding Interobserver Agreement: The Kappa Statistic, *Family Medicine*, 37(5):360–363.

Luis VON AHN and Laura DABBISH (2008), Designing games with a purpose, *Communications of the ACM*, 51(8):58–67.

Andrzej WIŚNIEWSKI (1995), *The Posing of Questions: Logical Foundations of Erotetic Inferences*, Kluwer AP, Dordrecht, Boston, London.

Andrzej WIŚNIEWSKI (2013), *Questions, Inferences and Scenarios*, College Publications, London.

Andrzej WIŚNIEWSKI (2015), Semantics of Questions, in S. LAPPIN and Ch. FOX, editors, *The Handbook of Contemporary Semantic Theory, 2nd Edition*, pp. 273–313, Wiley-Blackwell, Oxford.