

# Nexing Corpus: a corpus of verbal protocols on syllogistic reasoning

António Branco\*, José Leitão†, João Silva\*,  
Luís Gomes†

\* University of Lisbon, Faculty of Sciences  
Dep. Informatics, Campo Grande, 1700 Lisboa, Portugal  
{ahb, jsilva}@di.fc.ul.pt

† University of Coimbra, Faculty of Psychology  
R. Colégio Novo, 3000 Coimbra, Portugal  
jleitao@ci.uc.pt

## Abstract

In this paper, we describe the Nexing Corpus and report on the tools implemented and the tasks undertaken for its development. The Nexing Corpus includes (i) a collection of written transcriptions of verbal data elicited during a psycholinguistic experiment on syllogistic reasoning; and (ii) performance data concerning that experiment, such as latencies, confidence levels and accuracy of answers provided. The verbal productions recorded in the corpus are of a specific linguistic type that is seldom, if at all, represented in corpora. These data are relevant for the development of human language technologies aimed at modeling this type of linguistic behavior, which is not uncommon in evolved interactions of cooperative agents. This corpus with thinking aloud data on syllogistic reasoning is also an important source of material for cognitive science, in particular for research on the nature of human deductive reasoning.

## 1. Introduction

In this paper, we describe the Nexing Corpus and report on the tools implemented and the tasks undertaken for its development.

The Nexing Corpus includes (i) a collection of written transcriptions of verbal data elicited during psycholinguistic experiment on syllogistic reasoning; and (ii) performance data such as latencies, confidence levels and accuracy of the answers provided by the participants in the experiment.

The corpus was developed in the scope of the activities of the Nexing project. The main objective of the Nexing project is to contribute for improving the automated mapping between (orthographic) form and (linguistic) meaning, on the one hand, and between (linguistic) meaning and knowledge (representation), on the other hand, in what concerns natural language negation. It is a multi-disciplinary project fostering the convergence of methods, results and expertise from Informatics, Applied Logic, Cognitive Psychology and Formal Linguistics in the areas of the syntax, semantics, pragmatics and reasoning.<sup>1</sup>

In Section 2, we describe the data included in the Nexing Corpus. In Section 3, the format used, both for the XML conformant structure of the documents and for the written transcriptions of the verbal data is presented. In Section 4, we introduce the applications we implemented to facilitate the development of the corpus and to visualize it.

## 2. Data

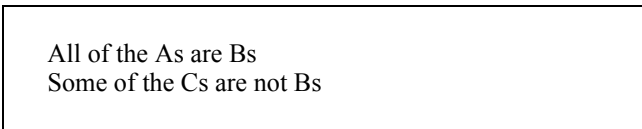
The source of the data collected in the corpus is the verbal productions gathered during a psycholinguistic experiment on syllogistic reasoning. These productions were audio recorded and latter transposed into written

format. At present, the Nexing Corpus is built on the basis of these written transcriptions.

### 2.1. Elicitation procedure

In the psycholinguistic experiment, each participant was presented with each of the 64 categorical syllogistic problems. The problems were presented in two sequences, in two separate occasions, to avoid work overload for the participants. The syllogistic problems were presented one at the time, on a computer screen. The rate of presentation of the problems was controlled by the participants in the experiment.

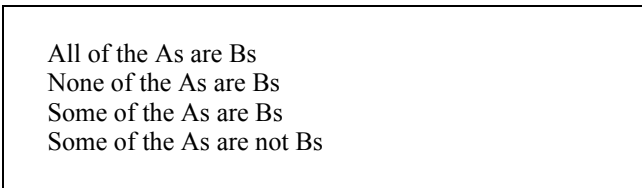
Each of the 64 problems is a pair of quantified assertions, the problem premises, for which the participant was asked to infer a necessarily true conclusion, if one existed, or to state the impossibility to infer such conclusion, if no valid conclusion is possible.



All of the As are Bs  
Some of the Cs are not Bs

Figure 1. An example of a syllogistic problem

Each one of the two premises in a problem is in one of four moods, corresponding to the four quantificational determiners *All*, *None*, *Some*, *Some...not* combined with terms *B*, and *A* or *C*.



All of the As are Bs  
None of the As are Bs  
Some of the As are Bs  
Some of the As are not Bs

Figure 2. The four moods of premises in figure A-B

<sup>1</sup> Nexing can be visited at <http://www.di.fc.ul.pt/~ahb/nexing.htm>.

The pattern of the terms in the premises is in one of four figures: A-B/B-C, B-A/C-B (the asymmetrical figures), and B-A/B-C, A-B/C-B (the symmetrical figures). The distinction symmetrical vs. asymmetrical refers to the position occupied by the middle term, B.

Asymmetrical figures:	
<i>Quant</i> of the As are Bs	<i>Quant</i> of the Bs are As
<i>Quant</i> of the Bs are Cs	<i>Quant</i> of the Cs are Bs
Symmetrical figures:	
<i>Quant</i> of the As are Bs	<i>Quant</i> of the Bs are As
<i>Quant</i> of the Cs are Bs	<i>Quant</i> of the Bs are Cs

Figure 3. The four figures of the premises, where *Quant* stands for one of the four quantificational determiners

The four possible moods in each of the two premises (4 x 2) multiplied by the two possible orders of the premises (2 x (4 x 2)), multiplied by the four possible figures (4 x (2 x (4 x 2))), yield the total of 64 problems used in the experiment.

When a participant declared that he had reached a solution, he should opt for either stating a conclusion relating the A and C terms in the premises, or declaring the impossibility of deriving a valid conclusion. He had then to stop the timing device and communicate his conclusion to the experimenter, who would type that conclusion in, completing the syllogistic pattern displayed in the computer screen.

All of the Bs are As
<u>None of the Bs are Cs</u>
Some of the As are not Cs

Figure 4. An example of a syllogism, including its valid conclusion

The experiment was conducted between March 1999 and June 1999. It involved 28 participants, 11 male and 17 female, who were undergraduate students at the University of Coimbra, majoring in either Psychology or Educational Sciences. All of the participants were speakers of the standard variant of Portuguese.

## 2.2. Verbal data

The 64 syllogistic problems were presented randomly to the participants. For each problem, participants were asked to orally report on the reasoning they were going through while searching for valid conclusions. At this stage, the verbalization was elicited by the standard thinking aloud instructions. The researcher's interference was thus limited to encourage the participant to keep on talking if he was to remain silent for more than five seconds.

After having dealt with the 64 problems, each participant was interviewed by the researcher leading the

experiment, and asked to explain thoroughly his reasoning for a selection of ten of the problems previously solved. Both the thinking aloud task and the reflexive conversation were analogically recorded using a table microphone.

## 2.3. Performance data

Besides the written transcription of the verbal productions, the corpus includes also the participants' appraisal of their confidence in the solution arrived at for each problem, rated in a six point scale. After having stated a solution to a particular problem, the participant was presented with a screen detailing the six possible confidence levels, of which he had to select the one that best described his confidence in the solution stated.

1. Completely Sure
2. Very sure
3. Sure
4. Unsure
5. Very Unsure
6. Completely Unsure

Figure 5. The six confidence levels

When the participant felt he was ready to state his final conclusion for each problem, he would press a designated button in a button box. This would prompt the computer to register the time spent in solving that particular problem, and the screen would then present the participant with the list of the six confidence levels.

The characterization of each problem, the correct answer, the answer provided by the participant, the data on the elapsed times, and the characterization of the participant are all recorded in the corpus.

## 3. Format

The corpus is encoded as a series of plain text files in the ISO-8859-1 character set (Latin-1) with an XML conformant structure. This allows to comply with an emerging formatting standard for data storing and makes it possible that using and handling the corpus be platform-independent as well as readable and editable by any text editor.

### 3.1. Annotation Conventions

The annotation conventions we adopted are adapted from the EAGLES recommendations. In the one hand we did not need all the generalization provided by guidelines such as those proposed in (Ide and Priest-Dorman, 2000). On the other hand, we needed both to record a very particular type of linguistic data — the transcription of the thinking aloud productions of participants while solving syllogistic problems —, as well as performance data related to the completion of that task. There was thus the need to integrate information very specific to this type of data, both linguistic and non-linguistic, in view of its utilization both from a language engineering and a cognitive science perspective.

The details and structure of the annotations used are as follows:

- **<protocol>**: each physical document/file contains a protocol (i.e. the complete experiment for one participant), and it is delimited by this tag. A protocol has a unique identifier (**pId**) and its structure consists of a header (**<header>**), one or more speakers (**<speaker>**) and the transcriptions (**<body>**).
- **<header>**: the header of a protocol is enclosed by this tag, and it includes information on the corpus (**<corpus>**) and the transcription guidelines observed (**<guidelines>**).
- **<corpus>**: this tag delimits information on the ownership of the document (**<ownership>**), the language of the linguistic data in the document (**<language>**), details on the encoding process (**<corpusEncoding>**) and details on the data gathering and transcription process (**<dataGatheringAndTranscription>**).
- **<ownership>**: information on the ownership of the corpus information is delimited by this tag, which contains the identification of (**<ownerId>**) and information (**<info>**) about the owner of the document.
- **<language>**: this tag marks the verbal data language.
- **<corpusEncoding>**: this contains information on **<responsible>**, **<assistant>**, **<date>** — the date when the encoding format was defined —, and **<transcription\_conventions>** — a description of the conventions for transcription from oral into written form that were observed.
- **<dataGatheringAndTranscription>**: this contains information on the **<responsible>**, the **<assistant>**, the address where to obtain a copy of the corpus, stated in **<info>**, the date when the data was registered, in **<gatherDate>**, the date when the data was transcribed, in **<transcDate>**.
- **<guidelines>**: this tag delimits information on the guidelines followed for the transcription of the verbal data:
  - **<gid>**: an identifier of the guidelines, in our case an adapted version of (Leech *et al.*, 1998), as described in Section 3.2 below;
  - **<author>**: the authors of the guidelines: Geoffrey Leech, Martin Weisser, Andrew Wilson and Martine Grice;
  - **<date>**: the date when the guidelines were created: October 18, 1998.
- **<speaker>**: this tag delimits information on the speaker whose verbal production is being encoded; it includes:
  - **<spId>**: the speaker's ID, defined in order to preserve speaker's privacy;
  - **<spTranscId>**: the speaker's ID for the transcription, rendered as a single letter, so that the transcription is not encumbered by the use of the full speaker's ID;
  - **<spRole>**: the speaker's role, which, for these corpora, can only have the value of subject or interviewer;
  - The speaker's age (**<spAge>**), sex (**<spSex>**), dialect (**<spDialect>**) and profession (**<spSocial>**).
- **<body>**: this tag delimits information from the different stages of the protocol. It includes several syllogistic problems and their attempted solutions (**<syllProbl>**), several justifications for the solutions found for the problems (**<justification>**), and an interview (**<interview>**).
- **<syllProbl>**: this tag marks one instance of a syllogistic problem. Each instance has a unique ID (**<syllProblId>**) and the following structure:
  - **<session>**: the problems were solved in two separate sessions (due to their large number); this tag marks in which session the problem was solved;
  - **<repetitionOf>**: sometimes a syllogistic problem had to be repeated: in such case this tag marks the ID of the original problem;
  - **<problem>**: this tag encloses all the information about the problem being solved: it defines the type of syllogism (**<syllType>**), the quantifiers for the first (**<quantP1>**) and second (**<quantP2>**) premise, the solution to the syllogism (**<solution>**) and if the converse answer is allowed (**<converseAllowed>**);
  - **<duration>**: the time, in milliseconds, taken by the participant to solve the problem;
  - **<spCertainty>**: the participant's certainty in his answer;
  - **<transc>**: this tag delimits the written transcription of the thinking aloud produced by the participant while solving the problem;
  - **<spAnswer>**: this tag encodes the subject's proposed solution: its structure includes the form of the answer (**<form>**), the quantifier used (**<quantifier>**) and if the proposed solution was correct (**<correctness>**).
- **<justification>**: this tag delimits the justification produced by the participant. It has a unique ID (**<justId>**) and the following structure:
  - **<justifiedSyllproblId>**: the ID of the syllogistic problem being justified;
  - **<transc>**: the written transcription of the justification.

- **<interview>**: this tag delimits the interview of the participant by the researcher responsible for the experiment. It has the only tag **<transc>**, embracing the actual transcription of the interview.

### 3.2. Transcription Conventions

The conventions for written transcription of oral data are also based on the recommendations from the EAGLES project (Leech *et al.*, 1998).

It is worth noting that each protocol is divided into three different parts: the solving of the syllogistic problems, the justifications, and finally the interview. The first involves only one speaker but the last two have a dialogue between the subject and the interviewer, and therefore we had to include tags to indicate who is speaking to whom and to account for phenomena such as overlapping speech.

The following is an overview of the transcription conventions observed:

- Words are transcribed using their standard form, i.e. they are written in the same way as they appear in the dictionary, regardless of the way they were pronounced.
- Numbers, dates, time, currency, and so on are transcribed in full:
  - e.g.: two thousand and one
- For acronyms, letters that are pronounced individually are separated by spaces:
  - e.g.: X M L
- Full stops, question and exclamation marks are used conventionally, while commas and other punctuation marks, such as parenthesis, etc., are not used.
- Asterisks are added to the end of incomplete words:
  - e.g.: *incomplete\**
- Unintelligible speech is marked with %\*%.
- If the transcriber is not sure of what was said but tries to guess it, he should signal the uncertainty by enclosing his guess with %:
  - e.g.: *extensible %markup% language*
- Pauses are marked with {pause:n} where n indicates the length of the pause and varies from 1, for a short pause, up to 4:
  - e.g.: to be {pause:2} or not to be
- In order to signal quasi-lexical vocalisations, the transcribers are restricted to choosing from a set of standardized forms (+um, +uh, +uhh, +uh-huh, +ohh and +ah) to avoid the proliferation of variants.
- Contractions are transcribed in their full form, but the corresponding tag includes both the contracted and the full form.
  - e.g.: {contraction: gonna, going to} going to
- Truncations, where a letter is repeated at the beginning of an expression, are marked with the {truncation} tag:
  - e.g.: {truncation:g} goal
- Repetitions, where it is an expression that is repeated, are signalled with {repetition}:
  - e.g.: {repetition:he} he is
- False starts, where an expression is interrupted and then corrected, are tagged with {falseStart}:
  - e.g.: {falseStart:he} she
- Non-verbal sounds are transcribed by means of the tag {nonVerbal:x}, where x is a place older for a standard set of codes (breath, laugh, cough, clearThroat and yawn), including one (noise) reserved to work as a catch-all for sounds not representable by the other options:
  - e.g.: {nonverbal:cough}
- Each protocol is divided into three distinct phases: the solving of the syllogistic problems, the justifications and finally the interview. The first involves only one speaker but the last two have a dialogue between the subject and the interviewer, and therefore we must include extra information to indicate who is actually speaking and to account for dialogue specific phenomena such as overlapping of speech:
  - To identify the speaker we precede his/her turn with {speaker:x}, where x is an identifier given to the speaker on the header of the protocol. The subject's identifiers may range from A to C, and the letter Z is reserved to identify the interviewer.
  - Sections of overlapped speech are marked with {beginOv:n} for the beginning and {endOv:n} for the end, with n being the identifier of the overlapped section.
    - e.g.:
 

```
{speaker:A} it works {beginOv:3}
like this {endOv:3}
{speaker:B} {beingOv:3} no it
does {endOv:3} not
```

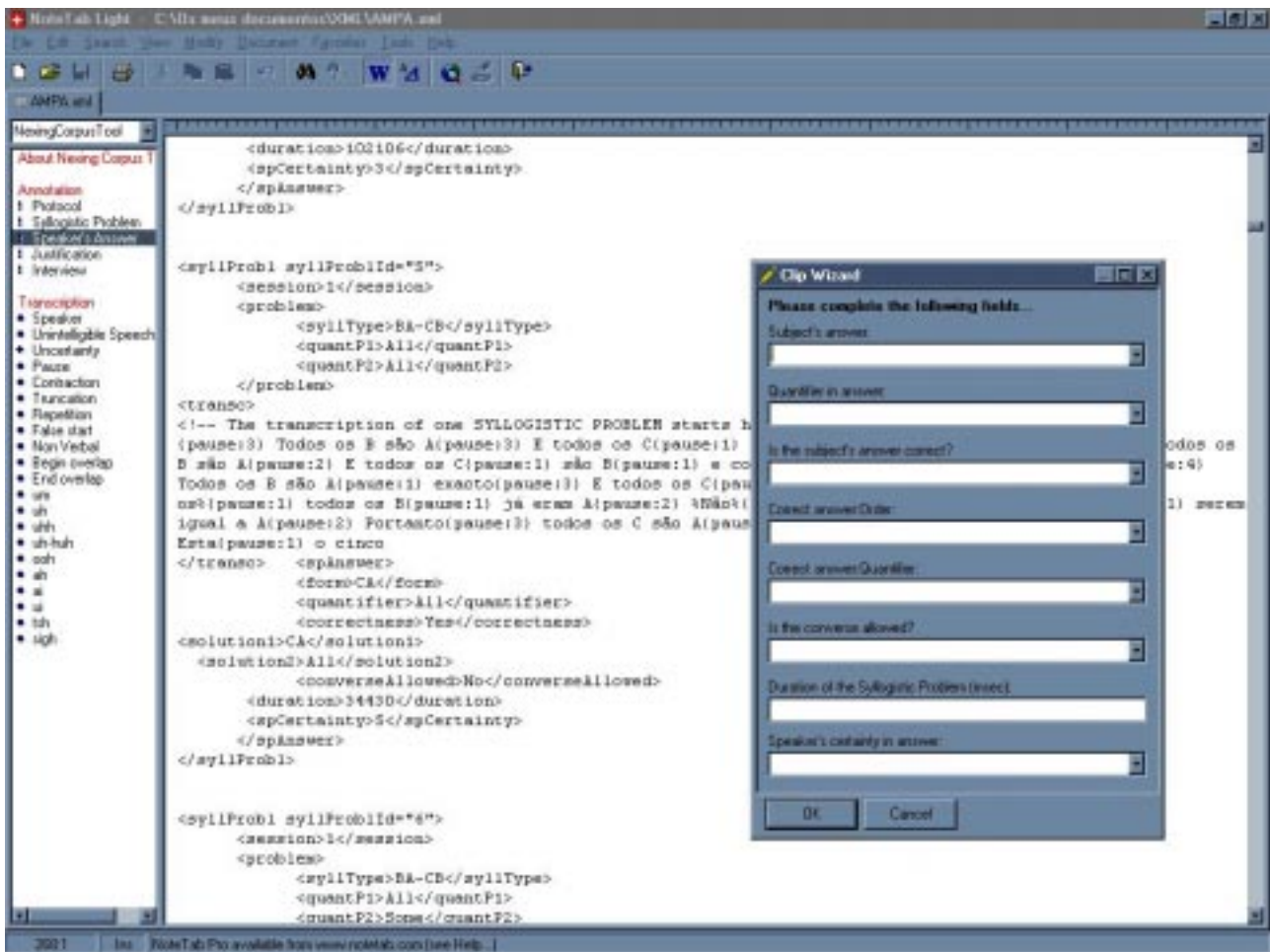


Figure 6. Annotation tool.

## 4. Tools

In order to facilitate the development and the usage of the Nexing Corpus, we developed two applications, the annotation tool and the visualization tool.

### 4.1. Annotation tool

In spite of our option for encoding the corpus in an XML conformant format, we felt that for the specific task of transcription and the concrete development of the corpus, an XML-dedicated, commercially available editor would not be necessary. The only advantage that such dedicated editor would bring over other choices would be the syntax-highlight that some XML editors permit.

Any plain text editor can handle XML files, which provided us with many possibilities to choose from. The transcription of the corpus involves assigning tags, which may turn out to be a tedious and error-prone task. This process was automated with the help of the functionalities provided by the editor NoteTab Light.<sup>2</sup>

This is a freeware text editor that easily allowed us to implement the needed automation without an overhead due to an excess of functionality. With this editor, we created what is called a Clipbook Library, which is

basically a set of macros to help editing the document. When invoked, these macros allow controlling the placement of embracing tags, their correct spelling and their completeness.

### 4.2. Visualization tool

While XML is an encoding format with the advantage of being handled by different platforms or text editors, an XML compliant text file tends, however, not to be suitable for an easy and accessible human manipulation. The tags that permit a standardly structured organization of the documents also obscure the data and the relevant content of the documents.

In order to allow for a productive use of the corpus, there was thus the need to hide some of the tags and use them to show different kinds of information in a different, user friendlier display. A straightforward answer to this need was the utilization of XSL style sheets, setting the details on how the information conveyed by each tag should be visually rendered to the final user. When placed in the relevant directory, specified in the corresponding XML files, the XSL style sheet file is used by an XSL compliant viewer to render the content of the original XML document in the required displaying format. The viewer we used for this purpose was the browser Internet Explorer 6.

<sup>2</sup> <http://www.notetab.com>.

Given the current utilization of the corpus in some research tasks in our project, we defined an XSL style sheet that highlights mostly the performance data. An example of the formatting effect of that style sheet can be

seen in the screen shot presented in Fig. 7 below. The XML code corresponding to the first of the two syllogistic problems shown in Fig. 7 can be found in Fig. 8.

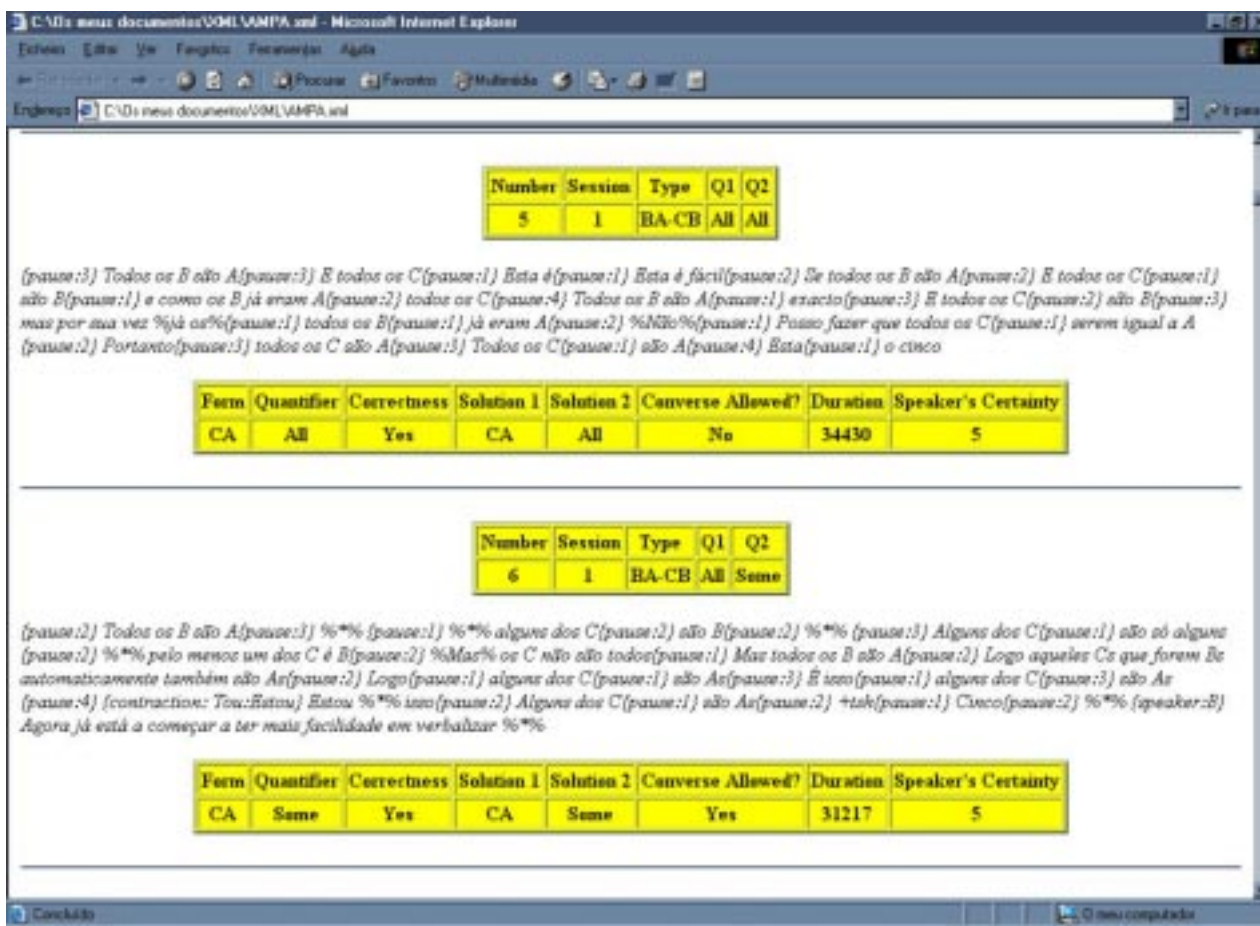


Figure 7. Visualization tool.

```
<syllProbl syllProblId="5">
  <session>1</session>
  <problem>
    <syllType>BA-CB</syllType>
    <quantP1>All</quantP1>
    <quantP2>All</quantP2>
  </problem>
  <transc>
  <!-- The transcription of one SYLLOGISTIC PROBLEM starts here -->
  {pause:3} Todos os B são A{pause:3} E todos os C{pause:1} Esta é{pause:1} Esta é
  fácil{pause:2} Se todos os B são A{pause:2} E todos os C{pause:1} são B{pause:1} e como os B
  já eram A{pause:2} todos os C{pause:4} Todos os B são A{pause:1} exacto{pause:3} E todos os
  C{pause:2} são B{pause:3} mas por sua vez %já os{pause:1} todos os B{pause:1} já eram
  A{pause:2} %Não%{pause:1} Posso fazer que todos os C{pause:1} serem igual a A{pause:2}
  Portanto{pause:3} todos os C são A{pause:3} Todos os C{pause:1} são A{pause:4} Esta{pause:1} o
  cinco
  </transc>
  <spAnswer>
  <form>CA</form>
  <quantifier>All</quantifier>
  <correctness>Yes</correctness>
  <solution1>CA</solution1>
  <solution2>All</solution2>
  <converseAllowed>No</converseAllowed>
  <duration>34430</duration>
  <spCertainty>5</spCertainty>
  </spAnswer>
</syllProbl>
```

Figure 8. Corpus excerpt.

## 5. Conclusions

At present the Nexing Corpus is made of 28 files, with around 15 000 tokens each, covering the transcription of around 30 hours of audio recordings. It is available for free download at <http://www.di.fc.ul.pt/~ahb/nexing.htm>.

This corpus can be explored for general language engineering purposes as well as to support further psycholinguistic research on reasoning.

The Nexing Corpus contains verbal productions of a specific linguistic type that is seldom, if at all, represented in corpora. Moreover, the linguistic data it contains correspond to verbalizations produced in a specific but important type of verbal behavior. They correspond to a reflexive discourse that is confluent with the mental activity of problem solving. Accordingly, these data may be highly relevant for the development of human language technologies aimed at handling this type of linguistic behavior, which is not uncommon in evolved interactions of cooperative agents (Cassell *et al.*, 2000).

In what concerns further research in the cognitive science of reasoning, this corpus with thinking aloud data on syllogistic reasoning supplies an important source of material for research on the nature of human deductive processing. Although thinking aloud data is rarely collected and made available as a result of experiments on human reasoning, its relevance for the understanding of the mental processes involved has become increasingly evident. First, thinking aloud data highlights individual differences in reasoning strategies, thus preventing undue aggregate treatment of this sort of asynchronous data, and provides a venue for the specific study of these strategies (Ford, 1995). Second, the fact that each individual participant contributes to the corpus with an unusually extensive problem solving activity makes it possible to study the emergence and modifications of the actual

individual reasoning strategies, which take shape in the course of the participant's involvement in the task. Furthermore, the evaluation of detailed computational models of this type of reasoning process finds in this kind of data a new source for validation studies, since these models imply predictions of verbalizations confluent with the mental process being modeled (Stenning and Yule, 1997).

## 6. References

- Cassel, J., J. Sullivan, S. Prevost and E. Churchill (eds.), 2000. *Embodied Conversational Agents*, Cambridge MA: MIT Press.
- Ford, M., 1995. Two modes of mental representation and problem solving in syllogistic reasoning. *Cognitive Science*, 23, 247-303.
- Ide, N. and G. Priest-Dorman, 2000. *Corpus Encoding Standard*, Expert Advisory Group on Language Engineering Standards, <http://www.cs.vassar.edu/CES/>.
- Leech, G., M. Weisser, A. Wilson and M. Grice, 1998. *Survey and Guidelines for the Representation and Annotation of Dialogue*, Expert Advisory Group on Language Engineering Standards, LE-EAGLES-WP4-4 Integrated Resources Working Group, <http://www.ling.lancs.ac.uk/eagles/delivera/wp4final.htm>
- Stenning, K., P. Yule, 1997. Image and language in human reasoning: a syllogistic illustration. *Cognitive Psychology*, 34, 109-159.

## 7. Acknowledgments

The research reported in this paper was supported by FCT-Fundação para a Ciência e Tecnologia, under the contract FCT/SAPIENS99/34076/99 for the NEXING project ([www.di.fc.ul.pt/~ahb/nexing.htm](http://www.di.fc.ul.pt/~ahb/nexing.htm)).