Building a corpus of annotated dialogues: the ADAM experience

Roldano Cattoni, ITC-irst, Trento, Italy, <cattoni@itc.it>
Morena Danieli, Loquendo, Torino, Italy, <Morena.Danieli@loquendo.it>
Andrea Panizza, Università del Piemonte Orientale "A. Avogadro",Vercelli, Italy, <Andrea.Panizza@CSELT.IT>
Vanessa Sandrini, ITC-irst, Trento, Italy, <sandrini@itc.it>
Claudia Soria, ILC-CNR, Pisa, Italy, <soria@ilc.pi.cnr.it>

1. Introduction

ADAM[1] is a corpus of annotated spoken dialogues currently being developed as part of the Italian national project SI-TAL[2]. Each dialogue is annotated at five levels of linguistic information: prosody, morphosyntax, syntax, semantics and pragmatics. The five levels were chosen for both practical (their interest for real applications) and scientific reasons (the possibility to investigate inter-level phenomena). For each level a corresponding annotation scheme has been defined that provides annotation instructions, examples and criteria. The result of each annotation is an XML file that encodes the content of a dialogue with respect to a particular level according to the annotation scheme of that level. The aim of this paper is therefore to present the ADAM corpus and the experience gained in defining and building such multi-level corpus. Section 2 describes the ADAM spoken corpus that includes both human-human and human-machine dialogues in the semantic domain of tourism and railways transportation. Section 3 provides a detailed introduction to the transcription format and to the five annotation schemes, one for each level of linguistic information. Section 4 focuses on the architectural issues of the ADAM corpus: essential requirements that drove the design process – like corpus reusability – are presented and discussed.

2. Corpus description

The ADAM spoken corpus is a collection of 450 vocal dialogues: they are both human-human (200 dialogues) and human-machine (250 dialogues). All the dialogues are recordings and transcriptions of telephone conversations in the semantic domain of tourism and railway transportation. The format of the audio files is the standard format for telephone signal data recommended by the SPEECHDAT project directions.

The human-human dialogues are simulated telephone conversations between two experimental subjects, playing the roles of a travel agent and of a caller, respectively. They had to perform pre-defined scenarios, including the request for train or flight timetables, the request for hotel arrangements, the reservation of both transportation and hotel, and the communication of credit card information. Each of the two acoustic signals (agent and caller) has been captured by two microphones (one directional and one "close-talk"), recorded on a digital tape as signed linear PCM 16bit at 16kHz. The human-human dialogues are longer than the human-machine ones, since the amount of information exchanged is greater. The total amount of recorded speech is more than 7 hours, for a total number of 58377 words for the human-human dialogues, while the length of the human-machine dialogues amounts to around 1250 utterances.

The human-machine dialogues were collected on the field: they are interactions between the automatic telephone information service of the Italian railway company and callers, recorded during an experimental phase of that service[3]. The callers called the system during night hours (from nine p.m. to eight p.m.). The calls came from all Italy: each call is from a different caller and several varieties of standard Italian are represented in the speech data. The system was designed to provide callers with information about the Italian train timetable. The dialogue manager allowed the user to enter several task parameters in a single turn; in case of recognition errors, the dialog manager exploited a set of repair strategies for recovering from the misunderstanding and being able to access the timetable database with the correct task parameters. The transaction success rate of the automatic service was around 85%, so 15% out of the dialogues present miscommunication phenomena. The speech signal was recorded at 8kHz and stored according the PCM-Ulaw 8 bit protocol.

[1] The acronym ADAM stands for "Architecture for Dialogue Annotation on Multiple Levels", see Soria, Cattoni and Danieli (2000).
[2] The ADAM Corpus will be released by the end of 2001. A pilot set of 30 dialogues is currently available upon request.
[3] The automatic service is based on CSELT speech and dialogue technologies (see Baggia, Castagneri, and Danieli 2000 for some details on service architecture and system performance).

3. Transcription and annotation practices

Each dialogue in the ADAM corpus is represented by an orthographic transcription (physically an XML file), which in turn is linked to an audio file containing the corresponding recording. In addition, the transcription of each dialogue is associated to five XML annotation files, according to five different levels or *layers* of linguistic information, namely prosody, morphosyntax, syntax, semantics and pragmatics. The five levels of annotation were mainly chosen in consideration of their interest for practical applications of the annotated material. In spite of the number of levels considered, and their sometimes conflicting requirements, we tried to develop a coherent, unitary approach to design and application of annotation schemes. In particular, in developing the different annotation schemes for the five levels envisaged, attention was paid to be consistent with criteria of robustness, wide coverage and compliance with existing standards and previous efforts in annotation of spoken dialogues. As a general criterion, however, attention has been paid so as to design the various annotation schemes and transcription conventions to be general enough to accommodate as many different annotation practices as possible.

3.1 Transcription

The representation of dialogues in terms of orthographic transcription implies making several choices about which aspects and phenomena of dialogues to represent. In accordance with the EAGLES recommendations for the representation of spoken language (see Gibbon 1999), the following information is made explicit:

- turns and speakers: according to common practice, a turn is identified with a stretch of talk produced by a single speaker. Each turn is numbered and the corresponding speaker is identified with a conventional label
- words: the transcription conventions adopted represent each recognisable word in orthographic form; non-recognisable words and word fragments are represented through approximated orthographic rendering of the perceived sound
- pauses are represented by means of specific symbols; length of pauses is also specified
- hesitation signals or fillers are given a specific status and classified as a separate category
- non-linguistic phenomena such as coughs, laughs or noises from surrounding context are classified and signalled via dedicated symbols.

For an illustration, see the following example:

> t_002B: {a} buongiorno mi scusi mi chiamo annamaria degasperi [breath] senta io avrei bisogno di prenotare un treno [...] [breath] che parte da roma lunedi` magari verso le otto di mattina non piu` tardi [...] [breath] {e} [puff] per verona pero` per cortesia mi basterebbe un posto solo [ puff ] c' e` qualcosa [puff]

As a general criterion, we follow the practice of recording only actually pronounced words, without making assumptions about the nature and type of words in cases of word truncation and disfluency phenomena. The same holds for non-standard forms such as dialectal expressions or mispronounced words.

In these cases, the actually pronounced word or word fragment is represented. Optionally, the annotator can further specify whether the form is to be considered as a non standard form or an interrupted form. It is also possible, if needed, to specify the target form possibly intended by the speaker. This range of information can be expressed via a set of dedicated attributes that come as optional extensions of the basic annotation weaponry.

3.2 Prosodic annotation

In the ADAM corpus, the Prosodic Annotation (PA) is used to represent the prosodic structure of utterances at the suprasegmental level. This structure is represented by providing a subset of the ToBI scheme (Tones and Break Index, see Silverman at al. 1992).

ToBI is a widely used system since its five different levels of break indexes are able to capture a variety of intonation phenomena. The ToBI subset used in the ADAM project concerns the annotation of prosodic phrasing.

The break indexes used are five, from 0 to 4. They are applied on the basis the following criteria:

- *index 0*: used in clitic groups, and every time that a word is atonic and is leaned on following words (for example: *I **[0]** am; a **[0]** car*);

- *index 1*: used every time in which there is not a word separation (word boundary stronger than clitic but without intonational cut, for example: *I **[0]** saw **[1]** him **[1]** yesterday*, where between the words "saw" and "him", and between "him" and "yesterday", there are neither separation nor pauses);
- *index 2*: used to mark anomalous intermediate boundaries (mostly when a break occurs in an unexpected place, for example *I **[0]** bought **[2]** a **[0]** new **[1]** ca,.* where between the words "bought" and "a", there is an unexpected pause);
- *index 3*: used to mark intermediate intonation boundaries.

In the ADAM corpus, utterances of spontaneous speech makes it difficult to distinguish those cases where it is better to use index 3 from those where index 4 would be more appropriate. Therefore, index 4 is used according to the following criteria:

- *index 4*: used to mark complete disjunctions (an obvious case is at the end of turns) and used to mark intermediate disjunctions too, if and only if they are conclusive, for example in the cases of clear logic separations and intonation patterns changing (*Ok, I booked you a seat. Do you need more information?...)*

In this example, the two parts of the utterance are clearly separated and there is a change in the intonation pattern (from assertive to interrogative).

In addition to these five break indexes, we can use the following symbols for PA:

- symbol [*p*]: combined with the break index symbol (ex. *2p, 3p*) and used every time the boundary has a feature of hesitation;
- symbol [-]: used to signal an uncertainty in attributing levels;
- symbol [*u*]: used to signal uncertainty combined with hesitation.

3.3 Morpho-syntactic annotation

The ADAM proposal for morphosyntactic and syntactic annotation is a two-layer annotation structure, containing respectively information on word category and morphosyntactic features (*pos* tagging), and non recursive phrasal nuclei (called *chunks*). The morphosyntactic annotation level encodes the following information: a) identification of morphological words and linking to their corresponding orthographic counterparts; b) annotation of their pos-category; c) annotation of morphosyntactic features (such as number, gender, person, tense, etc.); d) annotation of their corresponding lemma. The particular tag set, though adapted to representation of Italian, is compliant with EAGLES recommendations (Gibbon 1999).

An example is given below:

```
<mw id="mw_004" lemma="BUONGIORNO" pos="I" mfeats="X">buongiorno</mw>
<mw id="mw_005" lemma="ESSERE" pos="V" mfeats="S1IP">sono</mw>
<mw id="mw_006" lemma="ANNAMARIA" pos="SP" mfeats="FS">annamaria</mw>
<mw id="mw_007" lemma="DEGASPERI" pos="SP" mfeats="NN">degasperi</mw>
<mw id="mw_008" lemma="VOLERE" pos="V" mfeats="S1DP">vorrei</mw>
<mw id="mw_009" lemma="PRENOTARE" pos="V" mfeats="F">prenotare</mw>
<mw id="mw_010" lemma="UN" pos="RI" mfeats="MS">un</mw>
<mw id="mw_011" lemma="VIAGGIO" pos="S" mfeats="MS">viaggio</mw>
```

In addition, the tag set is structured into a *core scheme*, supplying basic means for annotating morphological information, and a periphery tag set, which serves the purpose of making provision for further linguistic annotation to be added to obligatory information.

This is the case, for instance, of a set of optional tags devised in order to annotate the so-called "discourse marker" class of words, i.e. a range of words belonging to several traditional grammatical categories and characterising themselves as a compact class as to their discursive or dialogic function. In this case, additional features are provided as an orthogonal dimension to recommended classification, so as to make it possible to express the fact that a given word is functioning as a discourse marker:

```
<mw id="mw_104" lemma="LE" pos="PQ" mfeats="FS3">le</mw>
<mw id="mw_105" lemma="ANDARE" pos="V" mfeats="S3IP">va</mw>
<mw id="mw_106" lemma="BENE" pos="B" mfeats="X">bene</mw>
<mw id="mw_107" lemma="SI'" pos="B" mfeats="X" dfeats="PF">si'</mw>
<mw id="mw_108" lemma="QUINDI" pos="C" mfeats="X" dfeats="MD">quindi</mw>
<mw id="mw_109" lemma="OTTO" pos="N" mfeats="NN">otto</mw>
<mw id="mw_110" lemma="E" pos="CC" mfeats="X">e</mw>
<mw id="mw_111" lemma="UN" pos="RI" mfeats="MS">un</mw>
```

Robustness and coverage were a crucial aspect in the development of the two schemes, in particular for what concerns i) syntactic constructions specific of spoken dialogues (ellipses, anacolutha, non verbal predicative sentences etc.), and ii) disfluencies (repetitions, false starts, trailing off etc.). The syntactic annotation level is built on top of the previous one and consists in identification of non-recursive phrasal nuclei (called *chunks*) and annotation of their category[4], as well as of their internal structure. The preference given to shallow parsing over, e.g., phrase structure trees is chiefly motivated by the locality of the analysis offered by this approach, a useful feature if one wants to prevent a local parsing failure from backfiring and causing the entire parse of an utterance to fail. This is particularly desirable when dealing with particularly noisy and fragmented input such as spoken dialogue transcripts. For an illustration of syntactic annotation, see the examples below:

```
<pn id="pn_004" type="INT" href="dial_002_mor.xml#id(mw_004)">
    buongiorno
  <h id="h_004" href="dial_002_mor.xml#id(mw_004)"/> </pn>
<pn id="pn_005" type="FV" href="dial_002_mor.xml#id(mw_005)">
    sono
  <h id="h_005" href="dial_002_mor.xml#id(mw_005)"/> </pn>
<pn id="pn_006" type="N" href="dial_002_mor.xml#id(mw_006)">
    annamaria
  <h id="h_006" href="dial_002_mor.xml#id(mw_006)"/> </pn>
<pn id="pn_007" type="FV" href="dial_002_mor.xml#id(mw_007)">
    degasperi
  <h id="h_007" href="dial_002_mor.xml#id(mw_007)"/> </pn>
<pn id="pn_008" type="FV"
    href="dial_002_mor.xml#id(mw_008)..id(mw_009)">vorrei prenotare
  <d id="d_001" type="modal" href="dial_002_mor.xml#id(mw_008)"/>
  <h id="h_008" href="dial_002_mor.xml#id(mw_009)"/> </pn>
<pn id="pn_009" type="N"
    href="dial_002_mor.xml#id(mw_010)..id(mw_011)">un viaggio
  <h id="h_009" href="dial_002_mor.xml#id(mw_011)"/> </pn>
```

3.4 Semantic annotation

The framework developed for ADAM allows the annotation of the semantic information through *concepts*. A concept is a typed structure that, using an ontology (e.g. a set of symbols that encode the a-priori information), represents the semantic information in a synthetic and non-ambiguous form. The result of the conceptual annotation of a dialogue is therefore an XML file in which a (possibly empty) collection of concepts is associated to each turn.

For the design of the annotation scheme of the conceptual level we have identified the following requirements: (1) *soundness*: the scheme should refer to well studied and formally sound representational approaches; (2) *expressiveness*: the scheme should allow the representation of the content of complex dialogues; (3) *minimality*: each turn should be annotated in a unique way (this requirement is rather strong since it is difficult to identify the "best" abstract level for the semantic content; therefore the requirement actually means that the annotation scheme should provide practical rules and criteria for this problem); (4) *simplicity*: the syntactic complexity of the language describing the concepts is to be minimised; (5) *locality*: each turn is independent of the previous turns and, in general, of the dialogue history; (6) *portability*: the annotation scheme should be domain-independent.

The ADAM annotation scheme takes inspiration from the so called "Frame-based Description Languages" (Cattoni and Franconi 1990) a framework developed in the field of the Knowledge Representation. In our scheme a concept is encoded like a "frame", a typed structure with "slots". Slots represent the properties of the concept and its relations with other concepts. Slots are encoded with the couple <slot-name, slot-value>: the former contains the name of a property, the latter either a simple value or a reference to another concept. This recursion allows the encoding of complex and structured semantics information. There are different types of concepts according to the content to be represented (e.g. "time", "trip", "room").

An example is needed at this point: given the simple sentence "The train leaves from Rome", the corresponding semantic annotation is:

```
<concept id="c_001" ctype="trip">
<slot sname="transportation-type" svalue="train"/>
<slot sname="origin" svalue="rome"/>
</concept>
```

[4] Syntactic annotation in ADAM is done automatically with manual check.

In this case only a concept (of type “trip”) is used, with two slots (properties): “transportation-type” with value “train” and “origin” with value “rome”.

Complex concepts can be encoded using reference to simpler ones. For example the sentence “The train leaves from Rome at eight on Saturday fifteen” is annotated with :

```
<concept id=“c_001” ctype=“trip”>
<slot sname=“transportation-type” svalue=“train”/>
<slot sname=“origin” svalue=“rome”/>
<slot sname=“departure-time” svalue=“*c_002”/>
</concept>
<concept id=“c_002” ctype=“time”>
<slot sname=“hour” svalue=“8:00”/>
<slot sname=“week-day” svalue=“saturday”/>
<slot sname=“month-day” svalue=“15”/>
</concept>
```

The simpler concept of type “time” (with identifier “c_002”) encodes a specific point in time. The more complex concept of type “trip” refers to such time to identify the value of the property “departure-time” - to do this the star ‘*’ character followed by the identifier of the referenced concept is used.

As far as the ontology is concerned, three categories of symbols may be distinguished: (1) symbols that identify the type of concept (the value of the “ctype” attribute of <concept>), (2) symbols that identify the name of concepts’ property (the value of the “sname” attribute of <slot>), (3) symbols that identify the value of concepts’ property (the value of the “svalue” attribute of <slot>). The user is free to adopt his/her conventions to encode the three categories of symbols of the ontology; nevertheless a good reference are the rules and symbols adopted by the C-STAR Consortium (Waibel 1996) for the inter-lingua: they have been developed on the basis of the experience on six different (Asiatic and European) languages and this appears to guarantee a good inter-lingua portability. For the semantic annotation of the ADAM corpus we actually adopted the C-STAR conventions: i) all symbols are English words, ii) complex symbols of categories (1) and (2) are obtained by means of the dash ‘-’ character (e.g. *week-day*, *interval-time*), iii) complex symbols of categories (3) are obtained by means of the underscore ‘_’ character (e.g. *new_york*).

It is important to emphasise here that the scheme is domain-independent so that the annotation is portable. In fact even if the domain changes or the ontology is enriched with new symbols, the annotation scheme and the corresponding representation in XML doesn’t change. For example, let us change the domain from the Transport Information context to that of Hotel Reservation: given the sentence “I would like a single room in Venezia for Saturday fifteen”, the annotation scheme doesn’t change even if the types of concepts, the name and values of their properties change:

```
<concept id=“c_001” ctype=“room”>
<slot sname=“quantity” svalue=“1”/>
<slot sname=“type” svalue=“single”/>
</concept>
<concept id=“c_002” ctype=“location”>
<slot sname=“value” svalue=“venice”/>
</concept>
<concept id=“c_003” ctype=“time”>
<slot sname=“week-day” svalue=“saturday”/>
<slot sname=“month-day” svalue=“15”/>
</concept>
```

Although most of the concepts encode strictly domain-dependent information, some domain-independent (or cross-domain) concepts do exist, like the temporal expressions. An user is clearly free to annotate temporal expressions as he/she likes; nevertheless for ADAM we have defined a set of predefined concepts to represent temporal expressions, taking inspiration from the Verbmobil TEL (Temporal Expression Language) (Reithinger 1999).

3.5 Pragmatic annotation

In several recent works on dialogue, there is an underlying assumption about the explicative power of the dialogue act notion for characterising discourse attitudes in human-human and human-machine dialogue (for example, Isard and Carletta 1995 and Di Eugenio et al. 1998), and some authors argue for the use of dialogue act tagging schemes for dialogue system evaluation. The ADAM meta-scheme[5] for pragmatic annotation is based on that widely held assumption, i.e. the pragmatic dimension

[5] For an explanation of the concept of “meta-scheme” see Section 4.

of the dialogues is characterised in terms of the "linguistic acts and the contexts in which they are performed" (Stalnaker 1970). The goal of the pragmatic annotation level is the attribution of one (or more than one) dialogue act tags to each utterance of the dialogue.

The scheme is a modified version of the tagging schemes DASML (Core and Allen 1997) and SWITCHBOARD-DAMSL (Jurafsky et al. 1997). In particular, it shares with DAMSL the features used to capture the communicative dimension of a dialogue turn. The ADAM annotation meta-scheme allows a three-layer pragmatic annotation practice: at the first layer, each dialogue turn is characterised with respect to its communicative level; at the second layer, the annotation captures the illocutionary dimension of the utterance(s) included in the turn; at the third layer, the discourse relationships among different utterances are characterised.

3.5.1 The communicative dimension

Table 1 reports the tags, and some examples provided by the ADAM pragmatic scheme for annotating the communicative status of the dialogue turns. The annotation of this dimension is largely inspired by Core and Allen (1997). The four tags may be used by the annotators to characterise the following aspects:

1. TASK: the turn provides a contribution to the fulfilling of the goal of the conversation
2. TASK-MANAGEMENT: the turn addresses some specific features of the problem-solving process related to the task;
3. COMMUNICATION-MANAGEMENT: the turn addresses phenomena connected with the maintenance of the communication channel.
4. OTHER-LEVEL: the communication content of the turn cannot be characterised with any of the previous three tags (for example, jokes, word-plays, *cliché*, etc...)

TAG	EXAMPLES
TASK	"Do you want to reserve on that flight?"
TASK-MANAGEMENT	"I'm taking note of your requiremnts for the flight reservation. "
COMMUNICATION-MANAGEMENT	"Please, hold on!"
OTHER-LEVEL	"Better late than never"

Table 1

3.5.2 The dialog act dimension

The dialog act labels provided by the ADAM pragmatic annotation scheme are reported in Table 2. The pragmatic annotation meta-scheme allows to characterise each utterance of the dialog with one or more dialog act tags on the basis of the role(s) of the utterance in the discourse. At the beginning of the annotation practice, we tried to apply to the annotation of this linguistic level the same minimality principle stated for the conceptual level, i.e. tagging each utterance one and only one dialogue act label. However, we soon realised that this simplification was not applicable at the pragmatic level, even in a task-oriented domain and for "simple" (question answering) dialogues. For example, indirect speech acts were hard to capture. At present, the pragmatic meta-scheme allows to attribute a primary label (characterising the direct speech act) and one, or more, secondary ones (for coding the indirect act). The secondary label is optional[6].

The dialogue act tags cover a large set of illocutionary functions in a task-oriented domain, and we believe that the scheme may be re-used, at some extent, for different domains. However, as for the other levels of the ADAM corpus, the problem of re-usability has been approached in terms of the pragmatic meta-scheme and its formal realisation (see below, section 4). Under this aspect, the distinction between the communicative dimension of a turn and the illocutionary content(s) of its utterance(s) is likely to be re-used in several annotation tasks.

[6] Under this respect the pragmatic annotation scheme presented here and applied in the ADAM corpus differs from the one described in Soria, Cattoni and Danieli (2000).

LABEL	EXAMPLE
Statement	*I'm leaving today*
Request	*I'd need a double room*
Accept	*The flight leaving at ten is nice for me*
Accept-Part	*Yes, but I'd need an extra bed for my child*
Open-Option	*Do you want me to reserve the return flight?*
Action-Directive	*Please, reserve two seats on the BA3476*
Repeat-Rephrase	*Oh, you said BA3476, the one leaving at 10 pm*
Collaborative-Completion	*…and I want to leave from NY next Sunday*
Conventional-Opening	*Hello, this is the Tourist Information Desk*
Conventional-Closing	*Good-bye*
Backchannel/Acknowledge	*Yes, of course*
Backchannel/Question	*Is that ok?*
Summarize/Reformulation	*So, you want to leave around 8 p.m.*
Or-Question	*Do you prefer a room with view on the garden or on the street?*
Apology	*Excuse me*
Thanking	*Thank you for calling*
Offer-Commit	*I've to check if there is a reduced fare available*
Yes/No Question	*Do you want to reserve the return flight on Thursday?*
Open-Question	*Which company do you prefer to travel?*
Reject	*No, I don't like to travel with this air company*
Yes-Answer	*Yes*
No-Answer	*No*
Response-Acknowledgement	*I agree*
Dispreferred-Answers	*No, I'd prefer to have a smoking room*
Opinion	*I believe this is the best solution*
Appreciation	*I enjoyed very much to work with you*
Abandoned/Uninterpretable	*I thin…*
Suggestion	*Perhaps we could try with another travel agent*
Signal-Non-Understanding	*Pardon?*
Signal-Understanding	*I see*
3rd-Party-Conversation	*Fido, stop barking, I can't hear a word!*
Other	*You know, I'd need to take a week off*

Table 2: Dialogue-act tag set

4. Architectural framework

The ADAM approach is mainly driven by the need of making the corpus widely reusable across different research and application purposes. In short, the concept of *corpus reusability* could be rephrased as the sum of the two concepts of "cross disciplinary acceptability" and "wide circulability". The two concepts refer, respectively, to the fact that a corpus a) either express a consensual or standard view about the type of information encoded (the content or *semantics* of annotation) and b) express a consensual view for what concerns the way information is physically encoded (the encoding *syntax* or markup language). The latter goal seems to have been reached through adoption of XML as a de-facto standard. On the other hand, the former point is more tricky and represents the motivation for the many well-known standardisation efforts. Adoption of common or standard annotation schemes, however, seems to be at least only partially practicable, for the very simple reason that establishing what the needs of future users might be is hardly feasible. An alternative to the search for standards is thus desirable.

A way out from the standardisation stricture is represented by concentrating our efforts in corpus design on maximisation of *corpus flexibility*. We claim that the degree of flexibility of a corpus depends on the extent to which the annotation is easily and quickly modifiable at a moderately low cost by subsequent users of the corpus.

To clarify a little, we can think of at least two possible scenarios where a user might need to customise the annotation provided with a corpus. First, it might be the case that a user wishes to reuse a corpus which is annotated for several types of linguistic information, but lacks of a particular annotation type; the potential user could nevertheless be interested in the existing annotations, and would like to supplement them with a new one. On the other hand, it might be the case that a user is interested in some annotation only (e.g., *pos*-tagging or syntactic structure) and s/he might want to leave aside other annotation types. Reusability of an annotated corpus can thus be thought of as a function of the extent to which new levels of linguistic information can be added, or uninteresting ones can be removed. This is what we call the *vertical* dimension of customisation in annotated corpora. Second, for each level of linguistic analysis, an annotated corpus is likely to be reused depending on the extent to which existing annotation can be changed, so as to accommodate different annotation practices. It is often the case that a corpus which is annotated with a given annotation scheme "hard-

wires" the annotation so as it is impossible to replace the annotation without reverting to the raw text and rebuilding the annotation from scratch, which is enormously expensive. This is what we call the *horizontal* dimension of customisation of an annotated corpus.

The extent to which an annotated corpus can be flexible enough to be compliant with these two requirements clearly depends on the particular choices made at the design level about the organisation and structuring of annotation.

If, for instance, all types of annotation are flattened onto a single representation level, it is clear that the customising operations above become hardly feasible. In ADAM we aim at maximising corpus flexibility by appealing to the two related notions of *modularity of annotation* and use of *annotation meta-schemes*.

The notion of modularity of annotation refers to data architecture (see MATE, Dybkjaer et al. 1998). In an annotated corpus, several different types of annotation or linguistic information may be present in relation to the same input data. These types of information can be thought of as independent, yet related, levels or *dimensions* of linguistic description. We thus can think of a level of prosodic analysis, another of pos-tagging, another of semantic analysis, etc. By *annotation modularity* we mean that the different layers of annotation are to be kept independent one of another. In the ADAM Corpus synchronisation among the different analyses and between these and the speech signal is ensured by the different annotations (stored as separate files) making reference to the same input file. This file, containing the transcription of the dialogue, is in turn linked to the audio file in PCM (a-low or u-low) format. Support for this structure is provided by the use of XML as mark-up language. By adopting this structure, annotation layers are linguistically heterogeneous and mutually orthogonal, so that changing one of them affects others only to a limited extent; layers are nevertheless indirectly related through a) their hinging on a common reference file (the "raw" text represented by the transcription file); b) the indirect correlation of the linguistic information they convey. This vertical modularity of the ADAM approach has interesting consequences for the purposes of reusability.

A potential user of the ADAM Corpus is left free to select, among the proposed levels of annotation, those which best reflect his/her theoretical and practical interests. (S)he can also feel the need for adding a new layer of information, not contemplated in today's ADAM realisation. By the way, level modularity is also of theoretical interest, since most annotation schemes we know differ mainly in the way pieces of linguistic information categorised, rather than in the intrinsic nature of these levels.

Moreover, level modularity seems to have a useful impact on our theoretical understanding of the linguistic phenomena at stake, since it is capable of expressing correlation relationships between layers, and ultimately between dimensions of linguistic analysis.

Horizontal customisation in annotated corpora can be enhanced by implementing the concept of *annotation meta-schemes*. According to our view, an annotation meta-scheme is a general descriptive framework in which different annotation schemes can be accommodated. In many cases the same unit of linguistic information can be annotated in different, arguably mutually incompatible ways, which are nonetheless all compatible with the recommended vertical modularity described above: so it is better to provide the potential user with the possibility of adopting any arbitrary annotation scheme without being forced to re-build the annotation from scratch or to forcefully comply with some other annotation scheme, no matter how standardised. To do so, it is necessary to have a representation format for the annotation that is general enough for competing schemes to be mutually substitutable. In other words, it is necessary to make the representation of annotation schemes as much scheme-independent as possible. It should be noted how the ADAM different annotation schemes do not, in fact, merely amount to another set of ready-made annotation schemes, but actually are represented in their XML annotation format in such a way that, for each annotation level, those features that are common to several competing schemes become slots or descriptive element tags to be associated with linguistic elements; the values of these attributes can be any arbitrary set of tags. Let us consider, for instance, the case of pragmatic annotation. The main difference between annotation schemes for this level of analysis lies in the particular types of dialogue act chosen rather than in the notion of dialogue act itself, which appears to be uncontroversial. If, however, we adopt a scheme where the basic descriptive element of any arbitrarily long set of words is the general tag <dialogue act>, further described by an attribute "type", different schemes can be applied to the same corpus without totally discarding the existing annotation: a substitution in the set of values will be enough. Conversion from one annotation scheme into another is easily done through XSLT transformations. It is our belief that enforcing this practice in the design of annotation schemes will bring us to more effective corpora exchange and reuse[7].

[7] In addition, the meta-scheme can be seen as a tool for effective comparison of alternative annotation schemes.

Finally, it must be noted that actual corpus reusability crucially depends also on the physical format or mark-up language used for corpus encoding[8]. As already stated throughout the paper, the mark-up language used for the encoding of the ADAM Corpus is XML. XML proved to be the ideal candidate for a number of reasons, all related to corpus reusability. First, it is an emerging and widespread standard, which ensures a good degree of corpus reusability in the times to come. Second, because of its platform-independence it enhances the potential for wide circulation of the annotated material, together with a considerable flexibility of use. More crucially, however, XML proved essential for implementation of the architectural choices described above. Annotation modularity is supported via extensive use of Xlink elements (DeRose et al. 2000). Each XML element in the annotation files is actually an hypertextual link which refers to an element (or set of elements) in the transcription file. All annotations for each dialogue are thus connected to the same input reference source (the transcription), thus ensuring synchronisation of the different annotations and still preserving their independence. On the other hand, the concept of annotation meta-scheme is easily implemented in XML, thanks to translation of the different annotation schemes content-independent. In other words, a general preference was given towards representing the different annotation tags as values of generic, scheme-independent attributes of XML elements. In this way the different annotation schemes (represented as different DTDs) are represented in a generic enough way, so that a future user of the corpus will only need to change the values of the different attributes for the entire annotation scheme to be changed. We believe that this approach represents a further value of the ADAM Corpus.

5. Conclusions and future work

In this paper we have described the methodological assumptions, the annotation practice and general architectural framework underlying the ADAM Corpus, which is a corpus of annotated spoken dialogues currently being developed as part of the Italian national project SI-TAL.

As far as annotation is concerned, our next step is to proceed to a validation phase, where the annotations performed by several annotators will be evaluated according to the metrics of (Isard and Carletta 1995). In addition to provide a concrete annotation experience, we have introduced what we believe to be an essential aspect to bear in mind in corpus design, namely the requirement of reusability. We have claimed that, for effective circulation and re-use of corpora, it is essential to make provision for as many practices of dialogue annotation as possible, as well as approaches to annotation at different levels, instead of providing fixed levels and schemes of analysis, no matter how standardised. Corpora will have a chance to be reused as far as it will be easy and relatively inexpensive to adapt them to different needs and application purposes. Use of XML as mark-up language is a further step toward this end.

6. References

Baggia P, Castagneri G, Danieli M 2000 Field trials of the Italian ARISE train timetable system. *Speech Communication* 31: 355-367.

Cattoni R, Franconi E 1990 Walking through the semantics of frame-based description languages: a case study. In *Proceedings of the Fifth International Symposium ISMIS '90*, Knoxville, TN, pp 234-241.

Core M, Allen J 1997 Coding dialogues with the DAMSL annotation scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Actions in Humans and Machines*, Cambridge, MA, pp 28-35.

DeRose S, Maler E, Orchard D, Trafford B 2000 XML linking language (Xlink). W3C Working Draft, 21 February 2000. http://www.w3.org/TR/xlink/.

Di Eugenio B, Jordan P, Moore J D, Thomason, R 1998 An empirical investigation of proposals in collaborative dialogues. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, Montréal, Canada, pp 325-329.

Dybkjaer L, Bernsen N O, Dybkjaer H, McKelvie D, Mengel A 1998 *The MATE markup framework*. MATE Deliverable 1.2. http://mate.nis.sdu.dk.

Gibbon D (ed) 1999 *Handbook of standards and resources for spoken language systems*. First supplement, EAGLES LE3-4244, Spoken Language Working Group.

[8] For a similar view see Ide and Brew (2000).

Ide N, Brew C 2000 Requirements, tools, and architectures for annotated corpora. In *Proceedings of the Workshop on Data Architecture and Software Support for Large Corpora, LREC 2000*, Athens, Greece, pp 1-5.

Isard A, Carletta J 1995 Replicability of transaction and action coding in the Map Task corpus. In Walker M, Moore J D (eds), *AAAI Spring Symposium: Empirical Methods in Discourse Interpretation and Generation*. Stanford, CA, pp 60-66.

Jurafsky D, Shriberg E, Briasca D 1997 *Switchboard DAMSL labeling project coder's manual*. Technical Report 97-02, University of Colorado, Institute of Cognitive Science, Boulder,

Reithinger N 1999 Robust information extraction in a speech translation system. In *Proceedings of Eurospeech '99*, Budapest, pp 2427-2430.

Silverman K, Beckman M, Pitrelli J, Ostendorf M, Wightam C, Price P, Pierrehumbert J, Hirschberg J 1992 TOBI: A standard for labeling English prosody. In *Proceedings of ICSLP 1992*, pp 867-870.

Soria C, Cattoni R, Danieli M 2000 ADAM: An architecture for xml-based dialogue annotation on multiple levels In Dybkjaer L, Hasida K, Traum D. (eds), *Proceedings of the First SIGDial Workshop on Discourse and Dialogue*, Association for Computational Linguistics, Hong Kong, pp 9-18.

Stalnaker, R C 1970 Pragmatics. *Synthese* 22, pp 272-289.

Waibel A 1996 Interactive translation of conversational speech. *Computer* 29(7): 41-48.