First Suggestions for an Emotion Annotation and Representation Language

Marc Schröder
DFKI GmbH
Saarbrücken, Germany
schroed@dfki.de

Hannes Pirker
OFAI
Vienna, Austria
hannes.pirker@ofai.at

Myriam Lamolle
LINC – University Paris 8 – IUT
de Montreuil, Paris, France
lamolle@iut.univ-paris8.fr

ABSTRACT
This paper suggests a syntax for an XML-based language for representing and annotating emotions in technological contexts. In contrast to existing markup languages, where emotion is often represented in an ad-hoc way as part of a specific language, we propose a language aiming to be usable in a wide range of use cases, including corpus annotation as well as systems capable of recognising or generating emotions. We describe the scientific basis of our choice of emotion representations and the use case analysis through which we have determined the required expressive power of the language. We illustrate core properties of the proposed language using examples from corpus annotation.

INTRODUCTION
Representing emotional states in technological environments is necessarily based on some representation format. Ideally, such an Emotion Annotation and Representation Language (EARL) should be standardised to allow for data exchange, re-use of resources, and to enable system components to work together smoothly.

As there is no agreed model of emotion, creating such a unified representation format is difficult. In addition, the requirements coming from different use cases vary considerably. In the Network of Excellence HUMAINE, we have nevertheless formulated a first suggestion, leaving much freedom to the user to "plug in" their preferred emotion representation. The possibility to map one representation to another will make the format usable in heterogeneous environments where no single emotion representation can be used.

DIFFERENT DESCRIPTIVE SCHEMES FOR EMOTIONS
A unified theory or model of emotional states currently does not exist [1]. Out of the range of existing types of descriptions, we focus on three that may be relevant when annotating corpora, or that may be used in different components of an emotion-oriented technological system.

Categorical representations are the simplest and most widespread, using a word to describe an emotional state. Such category sets have been proposed on different grounds, including evolutionarily basic emotion categories [2]; most frequent everyday emotions [3]; application-specific emotion sets [4]; or categories describing other affective states, such as moods or interpersonal stances [5].

Dimensional descriptions capture essential properties of emotional states, such as arousal (active/passive) and valence (negative/positive) [6]. Emotion dimensions can be used to describe general emotional tendencies, including low-intensity emotions.

Appraisal representations [7] characterise emotional states in terms of the detailed evaluations of eliciting conditions, such as their familiarity, intrinsic pleasantness, or relevance to one's goals. Such detail can be used to characterise the cause or object of an emotion as it arises from the context, or to predict emotions in AI systems [8,9].

USE CASES AND REQUIREMENTS FOR AN EMOTION ANNOTATION AND REPRESENTATION LANGUAGE
In order to ensure that the expressive power of the representation language will make it suitable for a broad range of future applications, the design process for EARL was initiated by performing a collection of use cases among members of HUMAINE. This list of use cases for emotional representations comprises i) manual annotation of emotional content of (multimodal) databases, ii) affect recognition systems and iii) affective generation systems such as speech synthesizers or embodied conversational agents (ECAs). On the basis of these use cases and the survey of theoretical models of emotions, a first list of requirements for EARL was compiled, which subsequently underwent discussion and refinement by a considerable number of HUMAINE participants.

Among the different use cases, the annotation of databases poses the most refined and extended list of requirements, which also covers the requirements raised in systems for recognition or generation.

In the simplest case, text is marked up with categorical labels only. More complex use cases comprise time-varying encoding of emotion dimensions [6], independent annotation of multiple modalities, or the specification of relations between emotions occurring simultaneously (e.g. blending, masking) [3].

EARL is thus requested to provide means for encoding the following types of information.

Emotion descriptor. No single set of labels can be prescribed, because there is no agreement – neither in theory nor in application systems – on the types of emotion de-

scriptors to use, and even less on the exact labels that should be used. EARL has to provide means for using different sets of categorical labels as well as emotion dimensions and appraisal-based descriptors of emotion.

Intensity of an emotion, to be expressed in terms of numeric values or discrete labels.

Regulation type, which encodes a person's attempt to regulate the expression of her emotions (e.g., simulate, hide, amplify).

Scope of an emotion label, which should be definable by linking it to a time span, a media object, a bit of text, a certain modality etc.

Combination of multiple emotions appearing simultaneously. Both the co-occurrence of emotions as well as the type of *relation* between these emotions (e.g. dominant vs. secondary emotion, masking, blending) should be specified.

Labeller confidence expresses the labeller's degree of confidence with the emotion label provided.

In addition to these information types included in the list of requirements, a number of additional items were discussed. Roughly these can be grouped into information about the person (i.e. demographic data but also personality traits), the social environment (e.g., social register, intended audience), communicative goals, and physical environment (e.g. constraints on movements due to physical restrictions). Though the general usefulness of many of these information types is undisputed, they are intentionally not part of the currently proposed EARL specification. If needed, they have to be specified in domain-specific coding schemes that embed EARL. It was decided to draw the line rather strictly and concentrate on the encoding of emotions in the first place, in order to ensure a small but workable representation core to start with. The main rational to justify this restrictive approach was to first provide a simple language for encoding emotional states proper, and to leave out the factors that may have led to the actual expression of this state. Thus, EARL only encodes the fact that a person is, e.g., trying to hide certain feelings, but not the fact that this is due to a specific reason such as social context. Clearly, more discussion is needed to refine the limits of what should be part of EARL.

PROPOSED REALISATION IN XML

We propose an extendable, XML-based language to annotate and represent emotions, which can easily be integrated into other markup languages, which allows for the mapping between different emotion representations, and which can easily be adapted to specific applications.

Our proposal shares certain properties with existing languages such as APML [10], RRL [8], and EmoTV coding scheme [3], but was re-designed from scratch to account for the requirements compiled from theory and use cases. We used XML Schema Definition (XSD) to specify the EARL grammar, which allows us to define abstract datatypes and extend or restrict these to specify a particular set of emotion categories, dimensions or appraisals.

The following sections will present some core features of the proposed language, using illustrations from various types of data annotation.

Simple emotions

In EARL, emotion tags can be simple or complex. A simple `<emotion>` uses attributes to specify the category, dimensions and/or appraisals of one emotional state. Emotion tags can enclose text, link to other XML nodes, or specify a time span using start and end times to define their scope.

One design principle for EARL was that simple cases should look simple. For example, annotating text with a simple "pleasure" emotion results in a simple structure:

```
<emotion category="pleasure">Hello!</emotion>
```

Annotating the facial expression in a picture file `face12.jpg` with the category "pleasure" is simply:

```
<emotion xlink:href="face12.jpg" category="pleasure"/>
```

This "stand-off" annotation, using a reference attribute, can be used to refer to external files or to XML nodes in the same or a different annotation document in order to define the scope of the represented emotion.

In uni-modal or multi-modal clips, such as speech or video recordings, a start and end time can be used to determine the scope:

```
<emotion start="0.4" end="1.3" category="pleasure"/>
```

Besides categories, it is also possible to describe a simple emotion using emotion dimensions or appraisals:

```
<emotion xlink:href="face12.jpg" arousal="-0.2"
  valence="0.5" power="0.2"/>

<emotion xlink:href="face12.jpg" suddenness="-0.8"
  intrinsic_pleasantness="0.7" goal_conduciveness="0.3"
  relevance_self_concerns="0.7"/>
```

EARL is designed to give users full control over the sets of categories, dimensions and/or appraisals to be used in a specific application or annotation context (see below).

Information can be added to describe various additional properties of the emotion: an emotion *intensity*; a *confidence* value, which can be used to reflect the (human or machine) labeller's confidence in the emotion annotation; a *regulation* type, to indicate an attempt to suppress, amplify, or simulate the expression of an emotion; and a *modality*, if the annotation is to be restricted to one modality.

For example, an annotation of a face showing simulated pleasure of high intensity:

```
<emotion xlink:href="face12.jpg" category="pleasure"
  regulation="simulate" intensity="0.9"/>
```

In order to clarify that it is the face modality in which a pleasure emotion is detected with moderate confidence, we can write:

```
<emotion xlink:href="face12.jpg" category="pleasure"
  modality="face" confidence="0.5"/>
```

In combination, these attributes allow for a detailed description of individual emotions that do not vary in time.

Complex emotions

A `<complex-emotion>` describes one state composed of several aspects, for example because two emotions co-occur, or because of a regulation attempt, where one emotion is masked by the simulation of another one.

For example, to express that an expression could be either pleasure or friendliness, one could annotate:

```
<complex-emotion xlink:href="face12.jpg">
  <emotion category="pleasure" confidence="0.5"/>
  <emotion category="friendliness" confidence="0.5"/>
</complex-emotion>
```

The co-occurrence of a major emotion of "pleasure" with a minor emotion of "worry" can be represented as follows.

```
<complex-emotion xlink:href="face12.jpg">
  <emotion category="pleasure" intensity="0.7"/>
  <emotion category="worry" intensity="0.5"/>
</complex-emotion>
```

Simulated pleasure masking suppressed annoyance would be represented:

```
<complex-emotion xlink:href="face12.jpg">
  <emotion category="pleasure" regulation="simulate"/>
  <emotion category="annoyance" regulation="suppress"/>
</complex-emotion>
```

If different emotions are to be annotated for different modalities in a multi-modal clip, there are two choices. On the one hand, they can be described as different aspects of one complex emotion, and thus share the same scope, i.e. the same start and end time:

```
<complex-emotion start="0.4" end="1.3">
  <emotion category="pleasure" modality="face"/>
  <emotion category="worry" modality="voice"/>
</complex-emotion>
```

Alternatively, the expressions in the different modalities can be described as separate events, each with their own temporal scope:

```
<emotion start="0" end="1.9" category="pleasure"
  modality="face"/>
<emotion start="0.4" end="1.3" category="worry"
  modality="voice"/>
```

It is an open question which of these alternatives is most useful in practice.

Annotating time-varying signals

Two modes are previewed for describing emotions that vary over time. They correspond to types of annotation tools used for labelling emotional database. The Anvil [11] approach consists in assigning a (possibly complex) label to a time span in which a property is conceptualised as constant. This can be described with the start and end attributes presented above.

The Feeltrace [6] approach consists in tracing a small number of dimensions continuously over time. In EARL, we propose to specify such time-varying attributes using embedded `<samples>` tags.

For example, a curve annotated with Feeltrace describing a shift from a neutral state to an active negative state would be realised using two `<samples>` elements, one for each dimension:

```
<emotion start="2" end="2.7">
  <samples value="arousal" rate="10">
    0 .1 .25 .4 .55 .6 .65 .66
  </samples>
  <samples value="valence" rate="10">
    0 -.1 -.2 -.25 -.3 -.4 -.4 -.45
  </samples>
</emotion>
```

The output of more recent descendents of Feeltrace, which can be used to annotate various regulations or appraisals, can be represented in the same way. A sudden drop in the appraisal "consonant with expectation" can be described:

```
<emotion start="2" end="2.7">
  <samples value="consonant_with_expectation" rate="10">
    .9 .9 .7 .4 .1 -.3 -.7 -.75
  </samples>
</emotion>
```

This relatively simple set of XML elements addresses many of the collected requirements.

A FAMILY OF EARL DIALECTS: XML SCHEMA DESIGN

Our suggested solution to the dilemma that no agreed emotion representation exists is to clearly separate the definition of an EARL document's structure from the concrete emotion labels allowed, in a modular design. Each concrete EARL dialect is defined by combining a base XML schema, which defines the structure, and three XML schema "plugins", containing the definitions for the sets of emotion categories, dimensions and appraisal tags, respectively. Different alternatives for each of these plugins exist, defining different sets of category labels, dimensions and appraisals.

For example, to allow emotions to be described by a core set of 27 categories describing everyday emotions in combination with two emotion dimensions, the EARL dialect would combine the base schema with the corresponding plugins for the 27 categories and the two dimensions, and the "empty set" plugin for appraisals. Another EARL dialect, describing emotions in terms of four application-specific categories, would combine the base schema with an application-specific category plugin and two "empty set" plugins for dimensions and appraisals.

Even though EARL will provide users with the freedom to define their own emotion descriptor plugins, a default set of categories, dimensions and appraisals will be proposed, which can be used if there are no strong reasons for doing otherwise.

MAPPING EMOTION REPRESENTATIONS

The reason why EARL previews the use of different emotion representations is that no preferred representation has yet emerged for all types of use. Instead, the most profitable representation to use depends on the application. Still, it may be necessary to convert between different emotion representations, e.g. to enable components in a multi-modal generation system to work together even though they use different emotion representations [8].

For that reason, EARL will be complemented with a mechanism for mapping between emotion representations. From a scientific point of view, it will not always be possible to define such mappings. For example, the mapping between categories and dimensions will only work in one direction. Emotion categories, understood as short labels for complex states, can be located on emotion dimensions representing core properties; but a position in emotion dimension space is ambiguous with respect to many of the specific properties of emotion categories, and can thus only be mapped to generic super-categories. Guidelines for defining scientifically meaningful mappings will be provided.

OUTLOOK

We have presented the expressive power of the EARL specification as it is currently conceived. Some specifications are still suboptimal, such as the representation of the start and end times, or the fact that regulation types cannot be associated a numerical degree (e.g., degree of simulation). Other aspects may be missing but will be required by users, such as the annotation of the object of an emotion or the situational context. The current design choices can be questioned, e.g. more clarity could be gained by replacing the current flat list of attributes for categories, dimensions and appraisals with a substructure of elements. On the other hand, this would increase the annotation overhead, especially for simple annotations, which in practice may be the most frequently used. An iterative procedure of comment and improvement is needed before this language is likely to stabilise into a form suitable for a broad range of applications.

The suggestions outlined in this paper have been elaborated in a detailed specification, currently submitted for comment within HUMAINE. Release of a first public draft is previewed for June 2006. We are investigating opportunities for promoting the standardisation of the EARL as a recommended representation format for emotional states in technological applications.

ACKNOWLEDGMENTS

We gratefully acknowledge the numerous constructive comments we received from HUMAINE participants. Without them, this work would not have been possible.

This research was supported by the EU Network of Excellence HUMAINE (IST 507422) and by the Austrian Funds for Research and Technology Promotion for Industry (FFF 808818/2970 KA/SA). OFAI is supported by the Austrian Federal Ministry for Education, Science and Culture and by the Austrian Federal Ministry for Transport, Innovation and Technology.

This publication reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

REFERENCES

1. Scherer, K. et al., 2005. Proposal for exemplars and work towards them: Theory of emotions. HUMAINE deliverable D3e, http://emotion-research.net/deliverables
2. Ekman, P. (1999). Basic emotions. In Tim Dalgleish and Mick J. Power (Ed.), *Handbook of Cognition & Emotion* (pp. 301–320). New York: John Wiley.
3. Douglas-Cowie, E., L. Devillers, J-C. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox (2005). Multimodal Databases of Everyday Emotion: Facing up to Complexity. In *Proc. InterSpeech*, Lisbon, September 2005.
4. Steidl, S., Levit, M., Batliner, A., Nöth, E., & Niemann, H. (2005). "Of all things the measure is man" - automatic classification of emotions and inter-labeler consistency. *ICASSP 2005, International Conference on Acoustics, Speech, and Signal Processing, March 19-23, 2005, Philadelphia, U.S.A., Proceedings* (pp. 317--320).
5. Scherer, K.R. (2000). Psychological models of emotion. In J. C. Borod (Ed.), *The Neuropsychology of Emotion* (pp. 137–162). New York: Oxford University Press.
6. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). 'FEELTRACE': An instrument for recording perceived emotion in real time, *ISCA Workshop on Speech and Emotion, Northern Ireland* , p. 19-24.
7. Ellsworth, P.C., & Scherer, K. (2003). Appraisal processes in emotion. In Davidson R.J. et al. (Ed.), *Handbook of Affective Sciences* (pp. 572-595). Oxford New York: Oxford University Press.
8. Krenn, B., Pirker, H., Grice, M., Piwek, P., Deemter, K.v., Schröder, M., Klesen, M., & Gstrein, E. (2002). Generation of multimodal dialogue for net environments. *Proceedings of Konvens*. Saarbrücken, Germany.
9. Aylett, R.S. (2004) Agents and affect: why embodied agents need affective systems Invited paper, 3rd Hellenic Conference on AI, Samos, May 2004 Springer Verlag LNAI 3025 pp496-504
10. de Carolis, B., C. Pelachaud, I. Poggi, M. Steedman (2004).APML, a Mark-up Language for Believable Behavior Generation, in H. Prendinger, Ed, *Life-like Characters. Tools, Affective Functions and Applications*, Springer.
11. Kipp, M. (2004). Gesture Generation by Imitation - From Human Behavior to Computer Character Animation. Boca Raton, Florida: Dissertation.com.