Multimedia Annotation with Multilingual Input Methods and Search Support

Hennie Brugman, Harriet Spenke, Markus Kramer, Alexander Klassmann

Max-Planck-Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
hennie.brugman@mpi.nl

Abstract

A tool set to create complex multimedia/multimodal annotations and to exploit them is described. Due to its possibility to flexibly define tiers and associate languages/writing systems with it and to even mix characters from different writing systems it is a tool which is especially suitable for work in multilingual environments. Also the search interface supports the multilingual features allowing to search for complex patterns in the annotations.

1. Introduction

Currently, most corpus projects in language engineering (LE) and field-linguistics (FL) are based on multimedia recordings and annotate multimodal behavior in communicative situations. Especially in field linguistics and multimodality research it is understood that in many situations human communication is much better covered by video compared to sound only recordings. Modern technology supports these trends since also for video standardized compression algorithms such as MPEG 1/2/4 and documented media file formats such as MPG, AVI and Quicktime[1] are available.

In several institutes these recordings are now completely digitized forming a multimedia online archive where users have immediate access to raw data at any moment in time. Users need to be able to flexibly annotate multimedia signals which can cover several audio and video tracks (multi camera recordings for gesture analysis) and time series such as eye-tracking or gesture data recorded with the help of data gloves. In many circumstances in LE and FL the annotations have to be done in several languages and increasingly often not only the major spoken ones. In FL it is a requirement that the indigenous communities are supported by providing annotations in their writing system. Of course, providing a phonetic/phonemic layer is a necessity in both fields in many cases.

Finally, modern multimedia/multimodal environments have to provide mechanisms to exploit multidimensional data including not only media tracks, but also many tiers of annotations. In multimodal annotations easily more than 40 tiers are created to describe the movements for example of the different articulators involved. The exploitation in general includes three components: (1) a visualization component offering different views on the complex data; (2) a search component which allows to find specific patterns and which prints out the hits in typical ways such as a concordance style of output; (3) a statistical component which can cover a wide variety of functions.

Modern multimedia/multimodal tool sets such as EUDICO [1] allow users to flexibly work with such complex structured multimedia/multimodal documents. Our goal is to make all functionality briefly indicated here available. This paper describes the actual state of the EUDICO tool set and that of the planned versions.

2. Multilinguality & Multimodality

As already indicated the annotations often have to include descriptions in several languages or character sets. For example, a phonetic tier can be annotated using the IPA set. Many character sets (fonts) which include some flavor of the IPA set exist on various platforms. The move towards UNICODE tries to unify all these attempts, based on the IPA Kiel definitions [2]. However, not all IPA characters ever defined by the researchers are included. It is not yet clear how the appropriate glyphs have to be represented within UNICODE. This aspect will be discussed in more detail below.

Further, especially field linguists are confronted with the requirement to support the writing system of the country where the recordings were made. At the MPI this currently means that we have to support Chinese, Arabic, Hebrew and extended Cyrillic in addition to the ISO-Latin characters. In near future we have to support Hindi and Bengali and due to the involvement in the DOBES project about documenting endangered languages others will come. So the tool to be used must support input and rendering methods for these writing systems. It has to support ligatures and right-to-left writing as they are for example used in Arabic, or sequence changes dependent on context, as they occur for example in Bengali. Therefore, it must be possible to associate a tier with a character set attribute.

In addition, it is increasingly often observed that loan words are used in languages for which there is no standard orthography and which should be annotated in the original writing system. This is especially true in studies of, for example, minority or endangered languages. Such occurrences require the possibility to switch easily between character sets, i.e. to mix orthographic text with phonemic/phonetic encodings. While orthographic text is often realized with ISO-Latin characters, phonemic/ phonetic encodings are done in IPA. Similarly, it should be possible to mix, for example, French orthography with Arabic ligatures in cases where Arabic immigrants are being recorded in France. Also, a researcher may wish to encode one tier mainly in Chinese characters but also wishes to include phonemic characters. These scenarios could become standard when tools start to support this in flexible ways. This mixing of different character sets often requires efficiency of data entry, i.e. the user often has to

[1] Quicktime describes more than a file format. It is a full API describing how to work with multi-track media files.

be equipped with keyboard layouts which he has to be able to create himself in an easy way.

When encoding gestures or signs in sign language or tones for tone languages special character sets are needed. For gestures and signs the HAMNOSYS hand shape forms [3] are widely accepted. Some scientists would like to enter and render them. Similarly this is true for the easy annotation of tones, for example in languages like Vietnamese which have 8 different forms. Special glyphs would make it easy for users to understand the encodings. Currently often some standard characters like numbers are used to indicate tone.

3. ELAN: EUDICO Linguistic Annotator

3.1. Tier Setup

ELAN is the annotation tool within the EUDICO annotation framework. It works for speech signals as well as for video signals, i.e. all segmentation and play features are symmetric. An enhanced version to visualize for example eye tracking data and gesture data is expected in 2002.

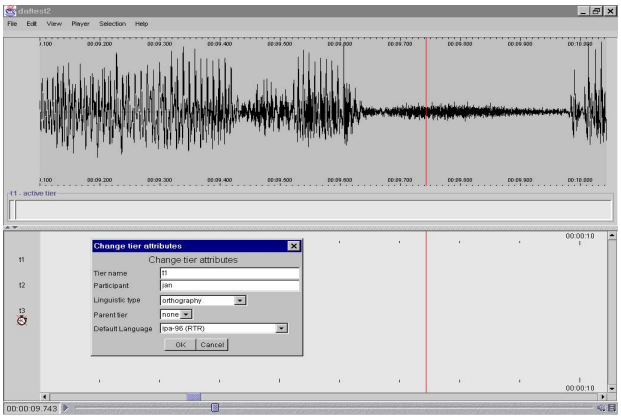

Figure 1 shows a screen shot which indicates the current way to define tier types.

Users can define their own set of annotation tiers for specific purposes, i.e. its type, its tier metadata and its dependencies with respect to other tiers. The type of an annotation tier can include a number of constraints on content, time alignment and structural embedding of the annotations on the tier. This is already partly available in the current version (see figure 1).

Browsing existing predefined tier setups will be a useful addition in the near future. Being able to create, browse and re-use tier setups is essential in situations where multimodal behavior is annotated, since annotating gesture, signs and facial expressions, for example, can easily amount to more than 40 layers as already indicated. If communicative situations are recorded, this is multiplied by the number of interacting partners. This requires the need to reuse tier type definitions and complete tier setups from earlier studies or from other researchers. Since many different research interests will have to be satisfied, it is also necessary to offer ergonomic methods to modify or extend such setups.

3.2. Selection Modes

ELAN has many useful options to easily define the time periods an annotation in one tier should be associated with. It is possible to define a time selection by mouse and keyboard and to easily use shortcuts to modify the left and right boundaries stepwise. The size of the step is dependent on the preferred mode: in sound mode the step size is chosen to be smaller than in video mode where it only makes sense to step frame by frame.

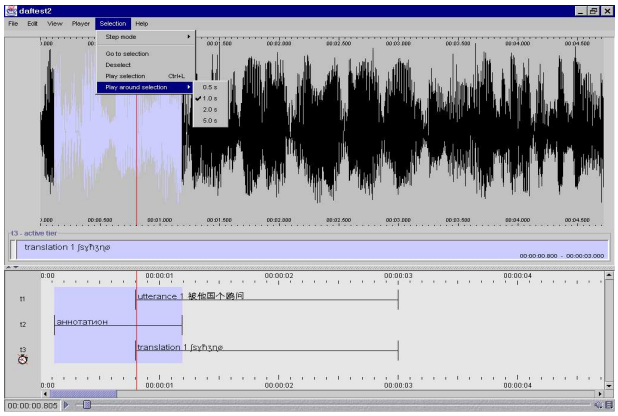

Figure 2 indicates how a selection can be specified when only sound is available. The screen shot shows three different information types: (1) On top the speech wave with a selected area. (2) At the bottom the time aligned tier viewer. (3) In the middle an extended viewer which shows annotations on a given tier for their duration (like a subtitle). It is especially used where the time viewer does not offer enough space to visualize all characters (especially since the user can select the character size to be used for the rendering). The pull-down menus show one of the ELAN features to facilitate operation. The user can define the size of the left and right context of the selection when playing.

An idea of the selection feature is given in figure 2. The selection can also be done in the video window. While dragging the mouse the video pane will show the content of the actual frame so that the user can easily see where he is. Due to the many possible tiers (ELAN has no limitation) pixel management is an important issue. All visualization components for the annotations allow the user to select the tiers he wants to see and to bring them with easy mouse drag and drop actions in a preferred order on screen.

Tier specific dependencies allow to operate efficiently, since either the boundaries of the parent annotation will be taken over or will function as constraints. Also an annotation on a certain tier can be identified as the one which will be associated with an annotation on another tier, i.e. annotations cannot only be linked to time periods but also to already existing annotations. While these possibilities are already existing in the current version, the visualization of hierarchical structures within annotations is something to be fully supported in the coming June release.

3.3. Multilingual Support

As indicated above, ELAN supports various character sets already now, i.e. it has appropriate input and

rendering methods. It already offers methods for Chinese, Cyrillic, Arabic, Hebrew and IPA. Its kernel has the capability to render, for example, Arabic ligatures or Bengali sequence shifts. It makes, amongst others, use of the GUK library [4]. The user can specify the preferred character set in the tier type specification, but it also allows the user to mix characters from different writing systems on one tier.

In figure 3 it is indicated how for example IPA characters can be entered. The keyboard layout can be displayed and the user can either use the physical keyboard or the one shown on the display to enter characters. A separate editing widget is opened to show what is being typed. When finishing this specific annotation all widgets showing the texts are immediately updated, i.e. also the time line view at the bottom will include the corresponding texts in a synchronized way.

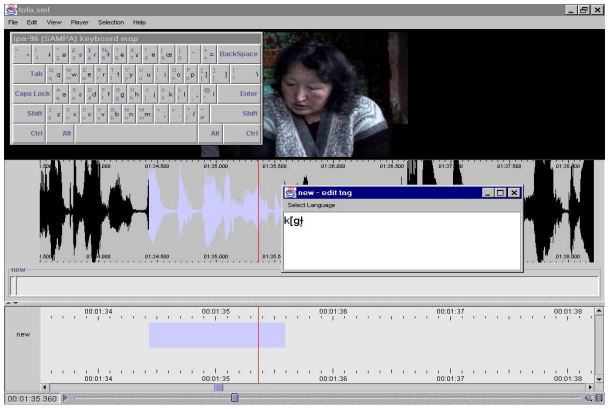

Figure 3 shows how IPA characters can be entered and how the edited string is displayed until it is accepted as being finished. Of course, any annotation can be opened again by double clicking on the annotation in for example the time line view.

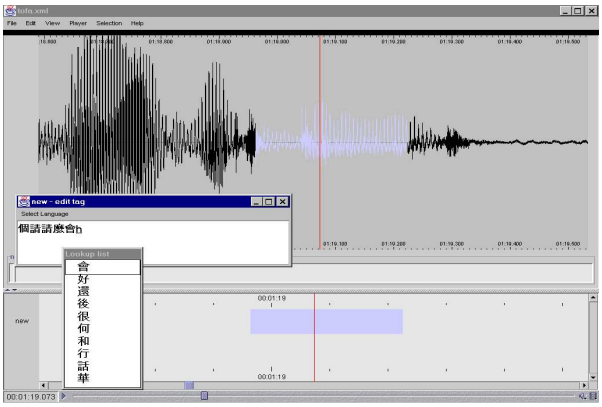

Figure 4 shows the way that Chinese characters can be entered. This works by first typing in PinYin characters and then selecting one from a list of given options.

Currently, the Chinese input method offers the Pinyin style of input, i.e. Roman character sequences lead to a list of possible Chinese characters from which a selection can be made with the mouse or arrow keys (see fig. 4). At the moment we lack support for advanced techniques which do a frequency ranking of such characters or which minimize the number of options by including lexical processing.

More work on multilingual input is planned to make EUDICO even more professional and user-friendly. We will extend the list of supported languages one by one. The used software library contains already a large variety of writing systems. Support for integration of character sets for tones and hand shapes requires some further research, for example on how to extend UNICODE.

It should be mentioned here that efficient encoding requires the availability of handy shortcuts. ELAN already has defined many of them for various operations. In the June release additional optimized modes of operation will be added for better support of specific annotation tasks.

4. XML and UNICODE Support

All EUDICO components support UNICODE, i.e. given the appropriate fonts it can render various character sets in a large number of relevant writing systems. However, it was indicated that UNICODE does not contain all glyphs yet which are necessary to efficiently support the annotations of a number of researchers. It has to be seen how UNICODE itself will develop. Besides what already was said about support for tones and hand shapes, it is known that still many less frequently used Chinese characters are not supported and that, for example, ancient Navajo glyphs are not yet included. It can be expected that UNICODE will be extended to support a much greater set of glyphs. It turns out that the user definable regions in UNICODE cannot be used, because a few big companies have clear intentions to use the number range for their purposes. Any other usage therefore would become problematic. Recently, an extension of the UTF-8 mechanism was discussed which would allow to represent many more characters.

Of course, fonts need to be available to support the rendering of these characters and suitable ideas for input methods have to be worked out.

EUDICO's standard output format is the XML-based generic EAF (EUDICO Annotation Format) (see appendix). It allows to make contents persistent which can be created according to the Abstract Corpus Model which is the nucleus of EUDICO [5]. Here our intentions are fairly comparable with what is currently worked out especially at NIST - called the ATLAS Interchange Format (AIF) [6]. Since AIF could not yet handle all necessary requirements (AIF did not yet support a tier concept) a EUDICO Interchange Format was defined (for the EAF schema see the appendix). However, we would like to join the AIF train to achieve a high degree of interoperability world-wide as soon as possible. The main structural components are of EAF are: (1) Explicitly ordered time slot values referring to potentially concrete time values; (2) information about the tier types and (3) as many Tiers with AlignableAnnotations or ReferenceAnnotations as necessary. While the first refer to time slots, the latter refer to existing annotations.

Due to the construction of EUDICO with its Abstract Corpus Model in the center it is straightforward to support other formats such as CHAT [7], SHOEBOX [8] including its hierarchical encodings, and relational databases such as used by MediaTagger (an older multimedia annotation tool developed at the MPI) [9].

Especially in the DOBES project the developers were confronted with a variety of other formats. Very popular amongst linguists turned out to be the use of MS WORD or EXCEL due to their early support for different character sets. Another issue is the usage of other software such as PRAAT (mainly for speech analysis, but also used for transcriptions) and Transcriber (a very popular and efficient program for sound annotation). While PRAAT has its own format and its own style to handle for example IPA characters, Transcriber produces XML format. However, its schema definition offers some peculiarities. To handle these various formats converters were built to map the annotations to EAF where possible. The MS WORD converter offers a small language with help of which the user can specify the document's structure, so that a proper XML file can be generated.

It should be added briefly, that EUDICO's media components are based on Java Media Framework which makes use of native media players on the different platforms. With the exception of MACs, where Quicktime currently does not properly support MPEG1/2 movies, MPEG1/2 is seen as the choice for media streaming dependent on the available network bandwidth. On MACs codecs such as Sorensen or Cinepak can be used, but this requires to store another converted copy of the video.

5. Corpus Exploitation

The EUDICO tool set also includes exploitation features. Basically, we can distinguish between visualization, printing and searching functionality. Currently, there is only a trivial statistical component.

5.1. Visualization

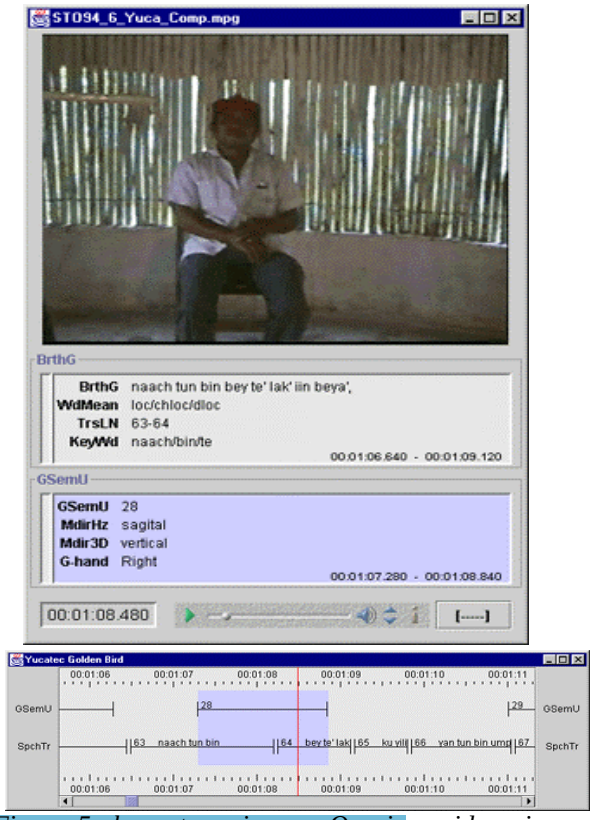

Figure 5 shows two viewers. One is a video viewer with fields for dynamically generated subtitles. As mentioned the user can choose what he wants to see in the subtitle fields and in which order. The other viewer contains the already introduced time line view. Due to the narrow boundaries of a time selection, often not the whole text can be visualized. A simple mouse click allows the user to see the whole string.

The EUDICO tool set offers a number of stereotypic views to inspect the media and annotations. All views are optional, to allow user-based pixel management, and tightly synchronized, i.e. any selection in one widget will lead to the appropriate actions in the others. Again speech and sound are handled fully symmetrically. In this paper we will not mention all features in detail. The following figures are meant to give an impression.

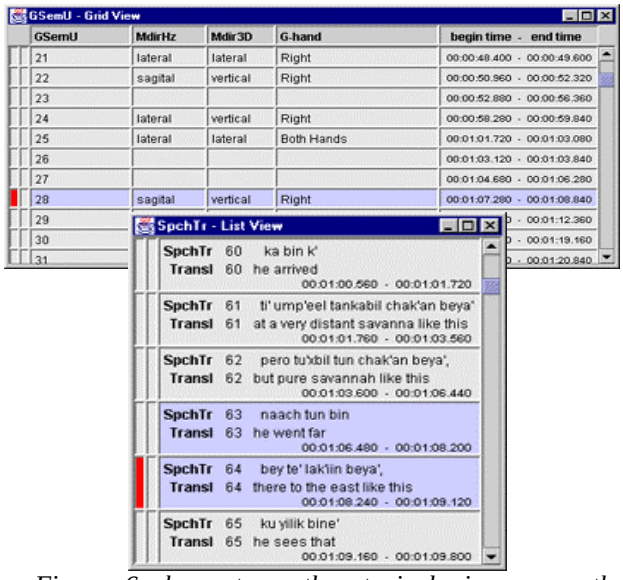

Figure 6 shows two other typical viewers on the annotations. There are a few others such as a pure text viewer which shows ongoing text as it is written in a text document. It is up to the user to select one or several. The marked areas indicate the synchronicity between the viewers such that navigation is easy for the user.

In figures 5 and 6 the visualization features of the current EUDICO version are indicated. Coming versions will include viewers for hierarchies such as syntax trees and interlinear text.

5.2. Multilingual Search Tool

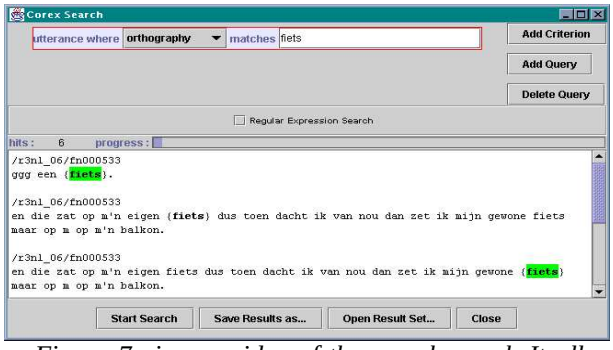

Figure 7 gives an idea of the search panel. It allows the user to specify patterns and associate them with annotation tiers. The search interface allows the user also to define distances between those patterns. The current hit list is a listing of the found annotations in context. By

clicking on them one can immediately jump back to the viewers which are synchronized with respect to the time instance of the hit.

Providing input and rendering methods for arbitrary character sets is not sufficient. Research requires the exploitation of stored annotations in various ways. Providing a user interface to enter queries which operate on the annotation documents is one of the first analysis tools to be provided. EUDICO offers this possibility in the form of a search interface that also allows the mixing of arbitrary characters for which input methods are available. The UNICODE representation in principle makes it very simple to search for the pattern. The combination of specifying regular expressions including this type of character mixtures per tier and entering distances, not only in terms of times but also in terms of text units, makes the EUDICO Search Tool a very powerful analysis tool for complex multimodal, multilingual annotations.

Figure 8 indicates such a jump back to one of the viewers, so that the hit can directly be analyzed in its multimedia and tier context.

The current version does not offer a concordance type of listing of the hits. This is planned for this year.

6. Availability and Support

The EUDICO tool set is available for free for academic usage and can be downloaded from the MPI tools site with the help of simple Webstart-based techniques [10] (www.mpi.nl/tools). Licenses for commercial usage and support can be given on request. It is also possible on special agreements to add special functionality. As example we can mention the intention to add a synchronized viewer for hand shapes in 3D based on the coordinates from data gloves. Requests in this direction have to be sent to the email address: software@mpi.nl. The EUDICO tool set is the major annotation and exploitation tool for a number of large projects and, at the MPI, for all multimedia/multimodal work. Therefore, we expect that EUDICO will not only be supported for many years, but that further development will take place in the coming years.

7. Summary

The EUDICO tool set can be used to create multimedia/multimodal annotations and to exploit them. It is a modern tool in so far that it treats audio and video symmetrically, supports synchronized viewers, operates in a distributed environment where the different files can be anywhere on the net, supports multilingual input and search, is based on UNICODE and creates XML-based output.

While the current version has already many interesting features, the coming versions will add more efficiency for the annotator, support for print out, support for hierarchical type of annotations and adding input methods for other languages and writing systems. The further development of UNICODE, however, is relevant for adding other glyphs which are not yet part of the standard.

Further details about the EUDICO Tool Set can be seen on the web-page [21].

8. References

[1] EUDICO: www.mpi.nl/tools
[2] IPA Kiel: www2.arts.gla.ac.uk/IPA/ipa.html
[3] Hamnosys: www.sign-lang.uni-hamburg.de/Projects/HamNoSys.html
[4] GUK: www.dcs.shef.ac.uk/research/groups/nlp/gate
[5] H. Brugman, P. Wittenburg (2001). The application of annotation models for the construction of databases and tools. In Proceedings of the Workshop on Linguistic Databases. Philadelphia
[6] Atlas Interchange Format: www.nist.gov/speech/atlas
[7] B. McWhinney (1999) The CHILDES Project: Tools for analyzing talk. Second ed. Hillsdale, NJ: Lawrence Erlbaum
[8] Shoebox: www.sil.org/computing/shoebox
[9] MediaTagger: www.mpi.nl/world/tg/lapp/mt/mt.html
[10] Webstart: java.sun.com/products/javawebstart

9. Appendix

This appendix contains the DTD for the EUDICO Annotation Format (EAF).

```
<!-- edited with XML Spy v4.1 U (http://www.xmlspy.com) by Hennie
   Brugman (Technical Group) -->
<!--
     Eudico Annotation Format DTD
     version 0.1
     July 5, 2001
-->
<!ELEMENT ANNOTATION_DOCUMENT (HEADER,
   TIME_ORDER, TIER*, LINGUISTIC_TYPE*, LOCALE*)>
<!ATTLIST ANNOTATION_DOCUMENT
     DATE CDATA #REQUIRED
     AUTHOR CDATA #REQUIRED
     VERSION CDATA #REQUIRED
     FORMAT CDATA #FIXED "1.0"
>
<!ELEMENT HEADER EMPTY>
<!ATTLIST HEADER
     MEDIA_FILE CDATA #REQUIRED
     TIME_UNITS (NTSC-frames | PAL-frames | milliseconds)
   "milliseconds"
>
<!ELEMENT TIME_ORDER (TIME_SLOT*)>
<!ELEMENT TIME_SLOT EMPTY>
<!ATTLIST TIME_SLOT
     TIME_SLOT_ID ID #REQUIRED
     TIME_VALUE CDATA #IMPLIED
>
<!ELEMENT TIER (ANNOTATION*)>
```

```
<!ATTLIST TIER
      TIER_ID ID #REQUIRED
      PARTICIPANT CDATA #IMPLIED
      LINGUISTIC_TYPE_REF IDREF #REQUIRED
      DEFAULT_LOCALE IDREF #IMPLIED
      PARENT_REF IDREF #IMPLIED
>
<!ELEMENT ANNOTATION (ALIGNABLE_ANNOTATION |
   REF_ANNOTATION)>
<!ELEMENT ALIGNABLE_ANNOTATION
   (ANNOTATION_VALUE)>
<!ATTLIST ALIGNABLE_ANNOTATION
      ANNOTATION_ID ID #REQUIRED
      TIME_SLOT_REF1 IDREF #REQUIRED
      TIME_SLOT_REF2 IDREF #REQUIRED
>
<!ELEMENT REF_ANNOTATION (ANNOTATION_VALUE)>
<!ATTLIST REF_ANNOTATION
      ANNOTATION_ID ID #REQUIRED
      ANNOTATION_REF IDREF #REQUIRED
      PREVIOUS_ANNOTATION IDREF #IMPLIED
>
<!ELEMENT ANNOTATION_VALUE (#PCDATA)>
<!ELEMENT LINGUISTIC_TYPE EMPTY>
<!ATTLIST LINGUISTIC_TYPE
      LINGUISTIC_TYPE_ID ID #REQUIRED
>
<!ELEMENT LOCALE EMPTY>
<!ATTLIST LOCALE
      LANGUAGE_CODE ID #REQUIRED
      COUNTRY_CODE CDATA #IMPLIED
      VARIANT CDATA #IMPLIED
>
```