Multimedia Corpora (Media encoding and annotation)

(Thomas Schmidt, Kjell Elenius, Paul Trilsbeek)

Draft submitted to CLARIN WG 5.7. as input to CLARIN deliverable D5.C-3 "Interoperability and Standards" [http://www.clarin.eu/system/files/clarin-deliverable-D5C3_v1_5-finaldraft.pdf]

Table of Contents

1 General distinctions / terminology

By a **Multimedia Corpus** we understand a systematic collection of language resources involving data in more than one medium. Typically, a multimedia corpus consists of a set of digital audio and/or video data and a corresponding set of textual data (the transcriptions and/or annotations of the audio or video data). Before we start discussing tools, models, formats and standards used for multimedia corpora in the following sections, we will devote a few paragraphs to clarifying our terminology.

1.1 Different types of multimedia corpora: spoken language vs. speech vs. phonetic vs. multimodal corpora vs. sign language corpora

Multimedia corpora are used in a variety of very different domains including, among others, speech technology, several subfields of linguistics (e.g. phonetics, sociolinguistics, conversation analysis), media sciences and sociology. Consequently, there is no consistent terminology, let alone a unique taxonomy to classify and characterise different types of multimedia corpora. Instead of attempting to define such a terminology here, we will characterise different types of corpora by describing some of their prototypical exponents. The categories and their characterising features are meant neither to be mutually exclusive (in

fact, many existing corpora are a mixture of the types), nor to necessarily cover the whole spectrum of multimedia corpora.

A **Spoken Language (or: Spontaneous Speech) Corpus** is a corpus constructed with the principal aim of investigating language as used in spontaneous spoken everyday interaction. A typical spoken language corpus contains recordings of authentic (as opposed to experimentally controlled or read) dialogues or multilogues (as opposed to monologues). It is transcribed orthographically, typically in a modified orthography which takes into account characteristic features of spontaneous speech (like elisions, assimilations or dialectal features), and it takes care to note carefully certain semi-lexical (or paraverbal) phenomena like filled pauses (hesitation markers) or laughing. A spoken language corpus may contain information about behaviour in other modalities (like mimics and gesture) in addition to the transcribed speech, but its main concern lies with linguistic behaviour. In that sense, corpora used in conversation or discourse analysis (e.g. the Santa Barbara Corpus of Spoken American English or the FOLK corpus) are prototypes of spoken language corpora. Meeting corpora like the ISL Meeting Corpus or the AMI Meeting Corpus can be regarded as another group of typical spoken language corpora. Less prototypical, but equally relevant members of this class are child language corpora (e.g. corpora in the CHILDES database), corpora of narrative (or: 'semi-structured') interviews (e.g. the SLX Corpus of Classic Sociolinguistic Interviews by William Labov), or corpora of task-oriented communication (like Map Tasks, e.g. the HCRC Map Task Corpus). Often, spoken language corpora are restricted to a specific interaction type (e.g. doctor-patient communication, classroom discourse or ad-hoc interpreting) or speaker type (e.g. adolescents or a certain social group).

A **Speech (Technology) Corpus** is a corpus constructed with the principal aim of building, training or evaluating speech technology applications like speech recognizers, dialogue systems or text-to-speech-systems. A speech corpus typically contains recordings of read or prompted speech – by one speaker in a studio for text-to-speech systems, or by more than hundreds of persons for speech recognition systems. The latter are often recorded in an office setting or recorded over mobile and/or fixed telephones. They are usually balanced over age and gender. Speech corpora are usually transcribed in standard orthography (possibly with labels for various acoustic noises and disturbances as well as truncated utterances), thus abstracting over idiolectal, dialectal or otherwise motivated variation in speech. Other modalities (like gestures or mimics) usually play no role in speech corpora (which are therefore usually audio corpora). In that sense, the SpeechDat project with up to 5000 speakers per language recorded over fixed telephone networks is a prototypical speech recognition database, as well as the corpora from the Speecon project, in which the recordings were made on location with high quality and also simple hands-free microphones. Also recordings of broadcast news are frequently used for speech recognition research and are relatively inexpensive to record. Besides containing spontaneous speech they include further challenges such as speaker change and non-speech events like music and jingles. The 1996 English Broadcast News Speech corpus is an example of these.

A **Phonetic (Speech) Corpus** is a corpus with the principal aim of carrying out research into the phonetics, phonology and prosody of language. A typical phonetic corpus contains recordings of elicited, i.e. read or prompted, speech, typically monologues The utterances are often syllables, words or sentences that are phonetically transcribed, sometimes augmented with prosodic labels. Typically, a phonetic corpus contains no information about other modalities. TIMIT, an Acoustic-Phonetic Continuous Speech Corpus of American English, is a prototypical phonetic corpus.

A **Multimodal Corpus** is a corpus with the principal aim of making the multimodal aspect of interaction accessible for analysis. In contrast to the other types, it does not prioritize any single modality, but rather treats speech, gestures, mimics, body posture etc. as equally important aspects. In contrast to a typical spoken language corpus, a multimodal corpus will use systematic coding schemes, rather than free descriptions, to describe non-verbal behaviour. In that sense, the SmartKom corpus is a prototypical multimodal corpus.

A **Sign Language Corpus** is a corpus in which the focus is not on the acoustic/articulatory modality of a spoken language, but on the visual/gestural modality of a signed language. Transcription of sign language corpora is typically done with the help of specialized transcription systems for sign languages. The Corpus NGT is a prototypical example of a sign language corpus.

The following table summarises distinguishing features of the five corpus types.

Corpus type	Recordings	Modalities	Research interest	Transcription	Prototype(s)
Spoken language	Multilogues, Audio or Video	Verbal more important than non-verbal	Talk in interaction	Modified Orthography	SBCSAE corpus FOLK corpus
Speech	Monologues, Audio	Verbal	Speech technology	Standard Orthography	SpeechDat corpora
Phonetic	Monologues, Audio	Verbal	Phonetic/Phonology	Phonetic	TIMIT
Multimodal	Video	Verbal and Non-verbal equally important	Multimodal behaviour	Modified or standard orthography	SmartKom corpus
Sign Language	Video	Signs	Sign language	Sign language transcription system	NGT corpus

It is important to note that many of the tools described in section 3 are suitable to be used with several or all of these corpus types, although they may have a more or less outspoken preference towards a single one of them (e.g. FOLKER for spoken language, Praat for phonetic, ANVIL for multimodal corpora). Still, standardisation efforts may have to take into account that different corpus types and the corresponding research communities have diverging needs and priorities so that evolving standards may have to be differentiated according to this (or a similar) typology.

1.2 Media encoding vs. Media annotation

By definition, a multimedia corpus contains at least two types of content: (audio or video) recordings and (text) annotations. Insofar as annotations are derived from and refer to the recordings, recordings are the primary and annotations the secondary data in multimedia corpora. Standardization is a relevant issue for both types of content. However, whereas the annotations are a type of content specific to the scientific community, audio or video recordings are used in a much wider range of contexts. Consequently, the processes for the definition and development of standards, as well as their actual state and their intended audience differ greatly for the two content types. We will treat standards and formats for the representation of audio and video data under the heading of "Media encoding" in section 2, and tools and formats for annotating audio and video data under the heading of "Media annotation" in section 3.

1.3 Data models/file formats vs. Transcription systems/conventions

Annotating an audio or video file means systematically reducing the continuous information contained in it to discrete units suitable for analysis. In order for this to work, there have to be rules which tell an annotator which of the observed phenomena to describe (and which to ignore) and how to describe them. Rather than providing such concrete rules, however, most data models and file formats for multimedia corpora remain on a more abstract level. They only furnish a general structure in which annotations can be organised (e.g. as labels with start and end points, organised into tiers which are assigned to a speaker) without specifying or requiring a specific semantics for these annotations. These specific semantics are therefore typically defined not in a file format or data model specification, but in a transcription convention or transcription system. Taking the annotation graph framework as an example, one could say that the data model specifies that annotations are typed edges of a directed acyclic graph, and a transcription convention specifies the possible types and the rules for labelling the edges. Typically, file formats and transcription systems thus complement each other. Obviously, both types of specification are needed for multimedia corpora, and both can profit from standardisation. We will treat data models and file formats in sections 3.1 to 3.3 and transcription conventions/systems in section 3.4. Section 3.5 concerns itself with some widely used combinations of formats and conventions.

1.4 Transcription vs. Annotation / Coding vs. Metadata

So far, we have used the term **annotation** in its broadest sense, as, for example, defined by Bird/Liberman (2001: 25f):

> [We think] of 'annotation' as the provision of any symbolic description of particular portions of a pre-existing linguistic object

In that sense, any type of textual description of an aspect of an audio or video file can be called an annotation. However, there are also good reasons to distinguish at least two separate types of processes in the creation of multimedia corpora. MacWhinney (2000: 13) refers to them as **transcription** and **coding**.

> It is important to recognize the difference between transcription and coding. Transcription focuses on the production of a written record that can lead us to understand, albeit only vaguely, the flow of the original interaction. Transcription must be done directly off an audiotape or, preferably, a videotape. Coding, on the other hand, is the process of recognizing, analyzing, and taking note of phenomena in transcribed speech. Coding can often be done by referring only to a written transcript.

Clear as this distinction may seem in theory, it can be hard to draw in practice. Still, we think that it is important to be aware of the fact that media annotations (in the broad sense of the word) are often a result of two qualitatively different processes – transcription on the one hand, and annotation (in the narrower sense) or coding on the other hand. Since the latter process is less specific to multimedia corpora (for instance, the lemmatisation of an orthographic spoken language transcription can be done more or less with the same methods and formats as a lemmatisation of a written language corpus), we will focus on standards for the former process in section 3 of this chapter.

For similar reasons, we will not go into detail about **metadata** for multimedia corpora. Some of the formats covered here (e.g. EXMARaLDA) contain a section for metadata about interactions, speakers and recordings, while others (e.g. Praat) do not. Where it exists, this kind of information is clearly separated from the actual annotation data (i.e. data which refers directly to the event recorded rather than to the circumstances in which it occurred and was documented) so that we think it is safe to simply refer the reader to the relevant CLARIN documents on metadata standards.

2 Media encoding

2.1 Audio encoding

2.1.1. Uncompressed:

WAV, AIFF, AU or raw header-less PCM, NIST Sphere

2.1.2. Formats with lossless compression:

FLAC, Monkey's Audio (filename extension APE), WavPack (filename extension WV), Shorten, TTA, ATRAC Advanced Lossless, Apple Lossless, MPEG-4 SLS, MPEG-4 ALS, MPEG-4 DST, Windows Media Audio Lossless (WMA Lossless).

2.1.3. Compressed Formats:

MP3, OGG, Flac, WMA, Vorbis, Musepack, AAC, ATRAC

2.2 Video encoding

2.2.1. Container formats

2.2.2. Codecs

3 Media annotation

3.1 Tools and tool formats

3.1.1 ANVIL (Annotation of Video and Language Data)

Developer:	Michael Kipp, DFKI Saarbrücken, Germany
URL:	http://www.anvil-software.de/
File format documentation:	Example files on the website, file formats illustrated and explained in the user manual of the software

ANVIL was originally developed for multimodal corpora, but is now also used for other types of multimedia corpora. ANVIL defines two file formats, one for specification files and one for annotation files. A complete ANVIL data set therefore consists of two files (typically, one and the same specification file will be used with several annotation files in a corpus, though).
The specification file is an XML file telling the application about the annotation scheme, i.e. it defines tracks, attributes and values to be used for annotation. In a way, the specification file is thus a formal definition of the transcription system in the sense defined above.
The annotation file is an XML file storing the actual annotation. The annotation data consists of a number of annotation elements which point either into the media file via a start and an end offset or to other annotation elements and which contain one ore several feature value pairs with the actual annotation(s). Individual annotation elements are organised into a number of tracks. Tracks are assigned a name and one of a set of predefined types (primary, singleton, span).
ANVIL's data model can be viewed as a special type of an annotation graph. It is largely similar to the data models underlying ELAN, EXMARaLDA, FOLKER, Praat and TASX.

```
<annotation>
  <head>
    <specification src="lq-demo-spec.xml"/>
    <video src="lq1-2-reich.avi"/>
    <!-- [...] -->
  </head>
  <body>
    <track name="trl" type="primary">
      <el index="0" start="1.514341115" end="1.747961878">
        <attribute name="token">wir</attribute>
      </el>
      <el index="1" start="1.747961878" end="2.130250453">
        <attribute name="token">reden</attribute>
        <attribute name="emphasis">moderate</attribute>
      </el>
      <!-- [...] -->
    </track>
    <!-- (ctd.) -->
    <track name="rst" type="span" ref="trl">
      <el index="1" start="4" end="11">
        <attribute name="relation1">elaboration</attribute>
        <attribute name="direction1">backward</attribute>
      </el>
      <el index="2" start="12" end="13">
        <attribute name="relation2">evidence</attribute>
        <attribute name="relation1">attribution</attribute>
        <attribute name="direction2">backward</attribute>
        <attribute name="direction1">forward</attribute>
      </el>
      <!-- [...] -->
    </track>
  </body>
</annotation>
```

Figure 1: Excerpt of an ANVIL annotation file

3.1.2 CLAN (Computerized Language Analysis)/CHAT (Codes for the Human Analysis of Transcripts) / Talkbank XML

Developers:	Brian MacWhinney, Leonid Spektor, Franklin Chen, Carnegie Mellon University, Pittsburgh
URL:	http://childes.psy.cmu.edu/clan/
File format documentation:	CHAT file format documented in the user manual of the software, Talkbank XML format documented at http://talkbank.org/software/, XML Schema for Talbank available from http://talkbank.org/software/.

The tool CLAN and the CHAT format which it reads and writes were originally developed for transcribing and analyzing child language. CHAT files are plain text files (various encodings can be used, UTF-8 among them) in which special conventions (use of tabulators, colons, percentage signs, control codes, etc.) are used to mark up structural elements such as speakers, tier types, etc. Besides defining formal properties of files, CHAT also comprises instructions and conventions for transcription and coding – it is thus a file format as well as a transcription convention in the sense defined above.
The CLAN tool has functionality for checking the correctness of files with respect to the CHAT specification. This functionality is comparable to checking the well-formedness of an XML file and validating it against a DTD or schema. However, in contrast to XML technology, the functionality resides in software code alone, i.e. there is no explicit formal definition for correctness of and no explicit data model (comparable to a DOM for XML files) for CHAT files.
CHAT files which pass the correctness check can be transformed to the Talbank XML format using a piece of software called chat2xml (available from http://talkbank.org/software/chat2xml.html).
There is a variant of CHAT which is optimised for conversation analysis style transcripts (rather than child language transcripts). The CLAN tool has a special mode for operating on this variant.

```
@UTF8
@Begin
@Languages:	en
@Participants:	CHI Ross Target_Child, MAR Mark Brother,
MOT Mary Mother,
	FAT Brian Father
*CHI:	I decided to wear my Superman shirt .
%snd:"boys49a1"_0_6130
%mor:	pro|I part|decide-PERF prep|to n|wear
pro:poss:det|my n:prop|Superman
	n|shirt .
%xsyn:	1|2|SUBJ 2|0|ROOT 3|2|JCT 4|3|POBJ 5|7|MOD
6|7|MOD 7|4|OBJ 8|2|PUNCT
%com:	first use of "I decided ." In morning while dressing .
@New Episode
@Tape Location:	185
@Date:	20-MAR-1982
@Situation:	Marky asked where the Darth_Vader was .
*FAT:	you mean the Darth_Vader head ?
%snd:"boys49a1"_6130_14572
%mor:	pro|you v|mean det|the n:prop|Darth_Vader n|head ?
%xsyn:	1|2|SUBJ 2|0|ROOT 3|5|DET 4|5|MOD 5|2|OBJ
6|2|PUNCT
*CHI:	but you really call it the Darth_Vader collection caser .
%snd:"boys49a1"_14572_19030
%mor:	conj:coo|but pro|you adv:adj|real-LY v|call pro|it
det|the n:prop|Darth_Vader
	n|collection n:v|case-AGT .
%xsyn:	1|0|ROOT 2|4|SUBJ 3|4|JCT 4|1|COORD 5|4|OBJ
6|9|DET 7|9|MOD 8|9|ENUM
	9|4|JCT 10|1|PUNCT
```

Figure 2: Excerpt of a CHAT text file

3.1.3 ELAN (EUDICO Linguistic Annotator) / EAF (ELAN Annotation Format)

Developer:	Han Sloetjes, MPI for Psycholinguistics, Nijmegen
URL:	http://www.lat-mpi.eu/tools/elan/
File format documentation:	Example set on the tool's website, XML schema inside the source distribution (available from the tool's website), format and data model explained in various unpublished (?) documents and slides.

ELAN is a versatile annotation tool and one of the major components of the LAT (Language Archiving Technology) suite of software tools from the MPI in Nijmegen. ELAN has been extensively used for the documentation of endangered languages, for sign language transcription and for the study of multimodality, but its area of application probably goes beyond these three corpus types.
ELAN reads and writes the EAF format, an XML format based on an annotation graph inspired data model, which has many similarities with the data models underlying ANVIL, EXMARaLDA, FOLKER, Praat and TASX.
Annotations are organised into (possibly interdependent) tiers of different types. Controlled vocabularies can be defined and also stored inside an EAF file. The tool and its format provide mechanisms for making use of categories inside the ISO-CAT registry and for relating annotations to IMDI metadata.

```
<ANNOTATION_DOCUMENT AUTHOR="" DATE="2006-06-13T15:09:43+01:00" FORMAT="2.3" VERSION="2.3">
  <HEADER MEDIA_FILE="" TIME_UNITS="milliseconds">
    <MEDIA_DESCRIPTOR MEDIA_URL="file:///D:/Data/elan/elan-example1.mpg" MIME_TYPE="video/mpeg"/>
    <MEDIA_DESCRIPTOR EXTRACTED_FROM="elan-example1.mpg" MEDIA_URL="file:///D:/Data/elan/elan-example1.wav"
          MIME_TYPE="audio/x-wav"/>
  </HEADER>
  <TIME_ORDER>
    <TIME_SLOT TIME_SLOT_ID="ts1" TIME_VALUE="0"/>
    <TIME_SLOT TIME_SLOT_ID="ts2" TIME_VALUE="280"/>
    <TIME_SLOT TIME_SLOT_ID="ts3" TIME_VALUE="440"/>
    <!-- [...] -->
  </TIME_ORDER>
  <TIER DEFAULT_LOCALE="en" LINGUISTIC_TYPE_REF="utterance" PARTICIPANT="" TIER_ID="K-Spch">
    <ANNOTATION>
      <ALIGNABLE_ANNOTATION ANNOTATION_ID="a1" TIME_SLOT_REF1="ts2" TIME_SLOT_REF2="ts5">
        <ANNOTATION_VALUE>so from here.</ANNOTATION_VALUE>
      </ALIGNABLE_ANNOTATION>
    </ANNOTATION>
    <ANNOTATION>
      <ALIGNABLE_ANNOTATION ANNOTATION_ID="a2" TIME_SLOT_REF1="ts22" TIME_SLOT_REF2="ts24">
        <ANNOTATION_VALUE>yeah</ANNOTATION_VALUE>
      </ALIGNABLE_ANNOTATION>
    </ANNOTATION>
    <!-- [...] -->
  </TIER>
  <TIER DEFAULT_LOCALE="en" LINGUISTIC_TYPE_REF="part of speech" PARENT_REF="W-Words"
    PARTICIPANT="" TIER_ID="W-POS">
    <ANNOTATION>
      <REF_ANNOTATION ANNOTATION_ID="a120" ANNOTATION_REF="a23">
        <ANNOTATION_VALUE/>
      </REF_ANNOTATION>
    </ANNOTATION>
    <ANNOTATION>
      <REF_ANNOTATION ANNOTATION_ID="a121" ANNOTATION_REF="a24">
        <ANNOTATION_VALUE>pro</ANNOTATION_VALUE>
      </REF_ANNOTATION>
    </ANNOTATION>
  </TIER>
  <LINGUISTIC_TYPE GRAPHIC_REFERENCES="false" LINGUISTIC_TYPE_ID="utterance" TIME_ALIGNABLE="true"/>
  <LINGUISTIC_TYPE CONSTRAINTS="Time_Subdivision" GRAPHIC_REFERENCES="false"
    LINGUISTIC_TYPE_ID="words" TIME_ALIGNABLE="true"/>
  <LINGUISTIC_TYPE CONSTRAINTS="Symbolic_Association" GRAPHIC_REFERENCES="false"
    LINGUISTIC_TYPE_ID="phonetic_transcription" TIME_ALIGNABLE="false"/>
  <!-- [...] -->
  <CONTROLLED_VOCABULARY CV_ID="POS" DESCRIPTION="Part of Speech">
    <CV_ENTRY DESCRIPTION="noun">n</CV_ENTRY>
    <CV_ENTRY DESCRIPTION="verb">v</CV_ENTRY>
    <CV_ENTRY DESCRIPTION="interjection">int</CV_ENTRY>
    <!-- [...] -->
  </CONTROLLED_VOCABULARY>
</ANNOTATION_DOCUMENT>
```

Figure 3: Excerpt of an ELAN annotation file

3.1.4 EXMARaLDA (Extensible Markup Language for Discourse Annotation)

Developers:	Thomas Schmidt, Kai Wörner, SFB Multilingualism, Hamburg
URL:	http://www.exmaralda.org
File format documentation:	Example corpus on the tool's website, DTDs for file formats at http://www.exmaralda.org/downloads.html#dtd, data model and format motivated and explained in [Schmidt 2005a] and [Schmidt2005b].

EXMARaLDA's core area of application are different types of spoken language corpora (for conversation and discourse analysis, for language acquisition research, for dialectology), but the system is also used for phonetic and multimodal corpora (and for the annotation of written language). EXMARaLDA defines three inter-related file formats – Basic-Transcriptions, Segmented-Transcriptions and List-Transcriptions. Only the first of these two are relevant for interoperability issues. A Basic-Transcription is an annotation graph with a single, fully ordered timeline and a partition of annotation labels into a set of tiers (aka the "Single timeline multiple tiers" data model: STMT). It is suitable to represent the temporal structure of transcribed events, as well as their assignment to speakers and to different levels of description (e.g. verbal vs. non-verbal). A Segmented-Transcription is an annotation graph with a potentially bifurcating time-line in which the temporal order of some nodes may remain unspecified. It is derived automatically from a Basic-Transcription and adds to it an explicit representation of the linguistic structure of annotations, i.e. it segments temporally motivated annotation labels into units like utterances, words, pauses etc. EXMARaLDA's data model can be viewed as a special type of an annotation graph. It is largely similar to the data models underlying ANVIL, ELAN, FOLKER, Praat and TASX.

```
<basic-transcription>
  <head>
    <meta-information>
      <transcription-name>PearStory</transcription-name>
      <referenced-file url="PearStory.mp3"/>
      <!-- [...] -->
    </meta-information>
    <speakertable>
      <speaker id="SPK0">
        <abbreviation>X</abbreviation>
        <sex value="f"/>
        <!-- [...] -->
      </speaker>
      <speaker id="SPK1">
        <abbreviation>Y</abbreviation>
        <!-- [...] -->
      </speaker>
    </speakertable>
  </head>
  <basic-body>
    <common-timeline>
      <tli id="T0" time="0.0" type="intp"/>
      <tli id="T1" time="1.9000"/>
      <!-- [...] -->
      <tli id="T5" time="5.0930"/>
      <tli id="T6" time="9.2000"/>
    </common-timeline>
    <tier id="TIE0" speaker="SPK0" category="sup" type="a" display-name="">
      <event start="T1" end="T3">louder </event>
      <!-- [...] -->
    </tier>
    <tier id="TIE1" speaker="SPK0" category="v" type="t" display-name="X [v]">
      <event start="T0" end="T1">So it starts out with: A </event>
      <event start="T1" end="T2">roo</event>
      <event start="T2" end="T3">ster crows. </event>
      <event start="T3" end="T4">((1,4s)) </event>
      <event start="T4" end="T5">((breathes in)) </event>
      <event start="T5" end="T6">And theen you see ehm a maan in maybe </event>
      <!-- [...] -->
    </tier>
    <tier id="TIE2" speaker="SPK0" category="nv" type="d" display-name="X [nv]">
      <event start="T0" end="T1">rHA on rKN, lHA on lSH </event>
      <event start="T1" end="T3">rHA up and to the right </event>
      <!-- [...] -->
    </tier>
  </basic-body>
</basic-transcription>
```

Figure 4: Excerpt of an EXMARaLDA Basic-Transcription

3.1.5 FOLKER (FOLK Editor)

Developers:	Thomas Schmidt, Wilfried Schütte, Martin Hartung
URL:	http://agd.ids-mannheim.de/html/folker.shtml
File format documentation:	Example files, XML Schema and (German) documentation of the data model and format on the tool's website

FOLKER was developed for the construction of the FOLK corpus, a spoken language corpus, predominantly addressing researchers in conversation analysis. Being built in parts on EXMARaLDA technology, FOLKER uses a data model based on STMT. However, the FOLKER XML file format stores transcriptions in a format in which the tier/annotation hierarchy is transformed into an ordered list of speaker contributions, thus bringing the format closer to structures typically used for written language. Optionally, transcriptions can be parsed according to the GAT transcription conventions. In addition to speaker contributions and temporally anchored segments, the file format will then also contain explicit markup for specific discourse entities like words, pauses, breathing etc.

```
<?xml version="1.0" encoding="UTF-8"?>
<folker-transcription>
  <head/>
  <speakers>
    <speaker speaker-id="CLA">
      <name>Clara</name>
    </speaker>
    <speaker speaker-id="JES">
      <!-- [...] -->
    </speaker>
  </speakers>
  <recording path="block.wav"/>
  <timeline>
    <timepoint timepoint-id="TLI_0" absolute-time="0.0"/>
    <timepoint timepoint-id="TLI_1" absolute-time="2.44443"/>
    <timepoint timepoint-id="TLI_2" absolute-time="3.17776"/>
    <timepoint timepoint-id="TLI_3" absolute-time="3.50253"/>
    <timepoint timepoint-id="TLI_4" absolute-time="3.75553"/>
    <timepoint timepoint-id="TLI_5" absolute-time="4.18886"/>
    <timepoint timepoint-id="TLI_6" absolute-time="5.35552"/>
    <timepoint timepoint-id="TLI_7" absolute-time="5.75552"/>
    <timepoint timepoint-id="TLI_8" absolute-time="6.277745"/>
    <timepoint timepoint-id="TLI_9" absolute-time="6.799965"/>
    <!-- [...] -->
  </timeline>
```

```
  <contribution speaker-reference="JES" start-reference="TLI_0"
    end-reference="TLI_4" parse-level="2">
    <uncertain>
      <w>it</w>
    </uncertain>
    <w>doesn</w>
    <w transition="assimilated">t</w>
    <w>matter</w>
    <w>he</w>
    <w transition="assimilated">s</w>
    <w>the</w>
    <w>boyfriend</w>
    <w>of</w>
    <w>one</w>
    <w>of</w>
    <w>your</w>
    <w>friends</w>
    <time timepoint-reference="TLI_1"/>
    <w>and</w>
    <w>as</w>
    <w>long</w>
    <w>as</w>
    <w>they</w>
    <w transition="assimilated">re</w>
    <time timepoint-reference="TLI_2"/>
    <w>atta<time timepoint-reference="TLI_3"/>ched</w>
  </contribution>
  <contribution speaker-reference="CLA" start-reference="TLI_2"
    end-reference="TLI_3"  parse-level="2">
    <w>what</w>
  </contribution>
  <contribution start-reference="TLI_4" end-reference="TLI_5"
     parse-level="2">
    <pause duration="0.43"/>
  </contribution>
  <!-- [.. ]-->
</folker-transcription>
```

Figure 5: Excerpt of a parsed FOLKER transcription file

3.1.6 Praat / TextGrid

Developers:	Paul Boersma/David Weenink
URL:	http://www.fon.hum.uva.nl/praat/
File format documentation:	(Sparse) description of the file format inside the tool's help database

Praat is a very widely used piece of software for doing audio annotation and phonetic analysis and thus for creating phonetic corpora. Praat reads and writes several audio formats and several text formats (all based on the same principle) for storing annotation data, acoustic measurements (pitch, intensity), etc. The file format relevant for this section is that of a Text Grid. The textGrid-file format is a plain text format. Different encodings, UTF-8 and UTF-16 among them, can be used. Annotations are organized into tiers and refer to the recording via timestamps. The data model is thus largely similar to the data models underlying ANVIL, ELAN, EXMARaLDA, FOLKER, and TASX.

```
File type = "ooTextFile"
Object class = "TextGrid"

xmin = 0.0
xmax = 47.01143378716898
tiers? <exists>
size = 6
item []:
   item [1]:
      class = "IntervalTier"
      name = ""
      xmin = 0.0
      xmax = 47.01143378716898
      intervals: size = 9
      intervals [1]:
         xmin = 0.0
         xmax = 1.900016816780248
         text = ""
      intervals [2]:
         xmin = 1.900016816780248
         xmax = 3.2510766568811755
         text = "louder "
      intervals [3]:
         xmin = 3.2510766568811755
         xmax = 31.04000569670313
         text = ""
      intervals [4]:
         xmin = 31.04000569670313
         xmax = 31.500004203597936
         text = "louder "
   item [2]:
      class = "IntervalTier"
      name = "X [v]"
      xmin = 0.0
      xmax = 47.01143378716898
      intervals: size = 53
      intervals [1]:
         xmin = 0.0
         xmax = 1.900016816780248
         text = "So it starts out with: A "
      intervals [2]:
         xmin = 1.900016816780248
         xmax = 2.0931989342595405
         text = "roo"
      intervals [3]:
         xmin = 2.0931989342595405
         xmax = 3.2510766568811755
         text = "ster crows. "
      intervals [4]:
         xmin = 3.2510766568811755
         xmax = 4.646368334964649
         text = "((1,4s)) "
      intervals [5]:
         xmin = 4.646368334964649
         xmax = 5.09300632412194
         text = "((breathes in)) "
      intervals [6]:
         xmin = 5.09300632412194
         xmax = 9.200016816639748
         text = "And theen you see ehm a maan in maybe "
      intervals [7]:
         xmin = 9.200016816639748
         xmax = 10.072686591293524
         text = "his fifties. "
```

Figure 6: Excerpt of a Praat TextGrid

3.1.7 Transcriber

Developers:	Karim Boudahmane, Mathieu Manta, Fabien Antoine, Sylvain Galliano, Claude Barras
URL:	http://trans.sourceforge.net/en/presentation.php
File format documentation:	DTD inside the source distribution, demo files inside the binary distribution

Transcriber was originally developed for the (orthographic) transcription of broadcast speech. It uses an XML format which organizes a transcription into one or several sections. Each section consists of one or several speech turns, and each speech turn consists of one or several transcription lines. Background noise conditions can be transcribed independently of the section/turn/line organization of the transcript. All of these units can be timestamped.

```
<Trans version="1" version_date="981211" audio_filename="frint980428" scribe="YM" xml:lang="fr">
 <Topics>
  <Topic id="to1" desc="les titres"/>
 </Topics>
 <Speakers>
  <Speaker id="sp1" name="Simon Tivolle" type="male"/>
  <Speaker id="sp2" name="Patricia Martin" type="female"/>
 </Speakers>
 <Episode program="France Inter" air_date="980428:0700">
  <Section type="filler" startTime="0.000" endTime="4.736">
   <Turn speaker="sp1 sp2" startTime="0.000" endTime="0.387">
    <Sync time="0.000"/>
    <Who nb="1"/> ouais . <Who nb="2"/> sûr ? </Turn>
   <Turn speaker="sp1" startTime="0.387" endTime="4.736">
    <Sync time="0.387"/> ah bon ? <Event desc="rire"/> non . blague , blague de Patricia . <Sync
     time="3.008"/>
    <Event desc="i"/> France-Inter , <Event desc="rire" type="noise" extent="begin"/> il est 7
    heures <Event desc="rire" type="noise" extent="end"/> . </Turn>
  </Section>
  <Section type="nontrans" startTime="4.736" endTime="9.609">
   <Turn startTime="4.736" endTime="9.609">
    <Sync time="4.736"/>
    <Background time="4.736" type="music" level="high"/>
    <Background time="9.609" type="other" level="off"/>
   </Turn>
  </Section>
  <Section type="filler" startTime="9.609" endTime="10.790">
   <Turn speaker="sp2" startTime="9.609" endTime="10.790">
    <Sync time="9.609"/> le journal , Simon Tivolle : </Turn>
  </Section>
  <Section type="report" topic="to1" startTime="10.790" endTime="20.000">
   <Turn speaker="sp1" startTime="10.790" endTime="20.000">
    <Sync time="10.790"/>
    <Event desc="i"/> bonjour ! <Sync time="11.781"/>
    <Background time="11.781" type="music" level="high"/>
    <Sync time="12.237"/> mardi 28 avril . <Sync time="13.344"/> la consultation nationale sur les
    programmes des lycées : <Sync time="16.236"/>
    <Event desc="i"/> grand débat aujourd'hui et demain à Lyon <Sync time="18.521"/> pour tirer les
    enseignements du </Turn>
  </Section>
 </Episode>
</Trans>
```

Figure 7: Transcriber file

3.1.8 Other tools[1]

There are numerous other tools for doing media annotation most of which use their own format. Since we believe the afore-mentioned tools to be the most relevant ones, we restrict ourselves to a short overview of the others here.

- **EMU Speech Database System** [http://emu.sourceforge.net/] – EMU is "a collection of software tools for the creation, manipulation and analysis of speech databases. At the core of EMU is a database search engine which allows the researcher to find various speech segments based on the sequential and hierarchical structure of the utterances in which they occur. EMU includes an interactive labeller which can display spectrograms and other speech waveforms, and which allows the creation of hierarchical, as well as sequential, labels for a speech utterance." (quote from the website) EMU reads and writes ESPS formatted label files as produced by ESPS and Waves+ software from Entropic. The system can also import Praat TextGrids.
- **Wavesurfer** (Developers: Kåre Sjölander and Jonas Beskow) [http://www.speech.kth.se/wavesurfer/] – Wavesurfer is a tool for sound visualization and manipulation, mainly used for the construction of speech corpora. It reads several formats commonly used for such corpora, namely HTK/MLF, TIMIT, ESPS/Waves+, and Phondat. Wavesurfer supports different encodings, Unicode encodings among them.
- **Phon** (Developers: Greg Hedlund and Yvan Rose) [http://childes.psy.cmu.edu/phon/] – Phon is a relatively new software "designed to facilitate phonological and phonetic analysis of data transcribed in CHAT" (quote from the website). The tool uses its own XML-based format, but should be largely compatible with the CHAT format.
- **XTrans** [http://www.ldc.upenn.edu/tools/XTrans/] – XTrans is "a next generation multi-platform, multilingual, multi-channel transcription tool developed by Linguistic Data Consortium (LDC) to support manual transcription and annotation of audio recordings" (quote from the website). It reads and writes a tabular separated text format.
- **TASX Annotator** (Developer: Jan-Torsten Milde, no URL available anymore) – The TASX annotator is a tool similar in design to ANVIL, ELAN and EXMARaLDA. It uses an XML-based data format representing a multi-layered annotation graph. Development of the tool was abandoned some time ago. The tool itself is not offered for download anymore, but some corpora created with it are available.
- **WinPitch** [http://www.winpitch.com/] – WinPitch is a windows based speech analysis tool, comparable in its functionality to (but probably not as widely used as) Praat. It uses an XML-based data format representing a multi-layered annotation graph.
- **Annotation Graph Toolkit** [http://agtk.sourceforge.net/] – The Annotation Graph Toolkit (AGTK) comprises four different tools all of which are based on the same software library and tool architecture. **TableTrans** is for observational coding, using a spreadsheet whose rows are aligned to a signal. **MultiTrans** is for transcribing multi-party communicative interactions recorded using multi-channel signals. **InterTrans** is for creating interlinear text aligned to audio. **TreeTrans** is for creating and manipulating syntactic trees. The tools are intended as a proof-of-concept for the annotation graph framework (see below). Their data format is the XML-based ATLAS interchange format (see below). Development of the AGTK was abandoned at a relatively early stage – the tools have probably not been widely used in practice.
- **Transana** (Developers: Chris Fassnacht and David K. Woods) [http://www.transana.org/] – Transana is a tool for managing, transcribing and analyzing digital audio and video recordings. Transcripts, keywords, video segmentations etc. are stored internally in a MySQL database. An XML export is provided for these data. However, in this export, transcripts are represented as one big stretch of character data, interspersed with RTF formatting instructions. Thus, there is no real content-oriented markup of the transcription. Moreover, when the transcription contains timestamps, the tool

[1] **AnColin** and **iLex** should probably also be mentioned here as tools used for constructing sign language corpora. However, I could not find sufficient information on the tools' data formats – references are Braffort et al. 2004 and Hanke/Storz 2008.

produces a non-well-formed XML export. Transana data are therefore rather problematic in terms of exchangeability and interoperability.

- **F4** [http://www.audiotranskription.de/f4.htm] – F4 may be seen as a typical exponent of a further class of annotation tools, namely simple combinations of a text editor with a media player. Other tools belonging to this class are the **SACODEYL Transcriptor** (http://www.um.es/sacodeyl/), **Casual Transcriber** (https://sites.google.com/site/casualconc/utility-programs/casualtranscriber/), or **VoiceScribe** (https://www.univie.ac.at/voice/page/voicescribe). Such tools are widely used for quickly creating text transcripts linked to media files. However, they have in common that they do not produced structured data which could be systematically exploited for further automatic processing. Rather, they use some free (plain or rich) text format with special constructs for linking into media. For the purposes of CLARIN, such data will not be usable without further curation.

3.2 Generic formats and frameworks

3.2.1 TEI transcriptions of speech

Chapter 8 of the TEI Guidelines is dedicated to the topic of "Transcriptions of Speech". The chapter defines various elements specific to the representation of spoken language in written form, such as

- utterances,
- pauses,
- vocalized but non-lexical phenomena such as coughs,
- kinesic (non-verbal, non-lexical) phenomena such as gestures, etc.

These elements can be used together with elements defined in other chapters to represent multimedia annotations. In particular, elements from chapter 16 ("Linking, Segmentation, and Alignment") can be used to point from the annotation into a recording, to represent simultaneity of events, etc. Furthermore, elements defined in chapter 3 ("Elements Available in All TEI Documents") and chapter 17 ("Simple analytic mechanisms") can be relevant for multimedia annotations.

The entirety of the elements and techniques defined in the TEI guidelines is certainly suited to deal with most issues arising in multimedia annotation. A problem with the TEI approach is rather that (in the general form as formulated the guidelines at least) it contains too many degrees of freedom. For most phenomena and for synchronisation of text and media in particular, there are always several options to express one and the same fact (e.g. the fact that two words are uttered simultaneously). Therefore, if the TEI guidelines are to be used with an annotation tool, more specific rules will have to be applied. However, few tools or corpus projects so far have developed such more specific TEI based formats for multimedia annotation.[2] Among the existing examples are:

- the French CLAPI database [http://clapi.univ-lyon2.fr/analyse_requete.php] which offers a TEI-based export of transcription files,
- the Modyco project [http://www.modyco.fr/corpus/colaje/viclo/] which uses a TEI-based format as an interlingua between ELAN and CHAT annotations (see Parisse/Morgenstern 2010),
- EXMARaLDA which contains options for importing and exporting TEI-conformant data (see Schmidt 2005a),

Besides proving the general applicability of TEI to multimedia annotations, these examples also demonstrate that being TEI-conformant alone does not lead to improved interoperability between annotation formats – none of the TEI examples mentioned is compatible with any of the others in the sense that data could be exchanged between the tools or databases involved.

These difficulties notwithstanding, the TEI guidelines certainly contain important observations which may become useful when defining a comprehensive standard for multimedia annotations. In particular, the fact that they are part of a framework which also (even: chiefly) deals with different types of written language data, is a good (and probably the only) point of contact for bringing together written corpora and multimedia corpora.

[2] Bird/Liberman (2001: 26) say:

> The TEI guidelines for 'Transcriptions of Speech' [...] offer access to a very broad range of representational techniques drawn from other aspects of the TEI specification. The TEI report sketches or alludes to a correspondingly wide range of possible issues in speech annotation. All of these seem to be encompassed within our proposed framework [i.e. Annotation Graphs, T.S.], but it does not seem appropriate to speculate at much greater length about this, given that this portion of the TEI guidelines does not seem to have been used in any published transcriptions to date.

3.2.2 Annotation Graphs / Atlas Interchange Format / Multimodal Exchange Format

Annotation graphs (AGs, Bird/Liberman 2001) are an algebraic formalism intended as "a formal framework [doing] for linguistic annotation". The authors explicitly state that, in a three-layer-architecture as commonly assumed for databases, AGs belong to the logical, not to the physical level. They thus formulate a data model rather than a data format. In principle, one and the same AG can be represented and stored in various physical data structures, e.g. an XML file, a text file or a relational database.

The basic idea of AGs is that "all annotations of recorded linguistic signals require one unavoidable basic action: to associate a label or an ordered set of labels, with a stretch of time in the recording(s)". Bird/Liberman's suggestion is therefore to treat annotations as Directed Acyclic Graphs (DAGs) whose nodes represent (or point to) timepoints in a recording, and whose arcs carry the non-temporal information, i.e. the actual text, of the annotation. They also allude to various ways of adding additional structure to such a DAG (e.g. by typing arcs or partitioning arcs into equivalence classes), but consciously leave the issue open to be resolved by concrete applications. In that sense, many of the data models described in the previous section (e.g. ANVIL, ELAN, EXMARaLDA) can be understood as applications of the AG framework, which specify and restrict a general AG to a subclass which can be efficiently handled by the respective tool.

One way of physically representing an annotation graph is level 0 of the ATLAS Interchange Format (AIF, DTD available from http://xml.coverpages.org/aif-dtd.txt). AIF defines an XML format which represents all the components of an AG as XML elements. The format, or a subset thereof, is used by different tools from the Annotation Graph Toolkit (AGTK, see above).

Taking the idea of AGs as a starting point, the developers of several of the tools described in the previous section devised a multimodal annotation format in which the common denominator information shared by the tools is represented in an AIF file. The format and the underlying analysis of the respective tool formats are described in Schmidt et al. (2008) and Schmidt et al. (2009). The format is supported by various (built-in or stand-alone) import and export filters. Although it is in practice less useful than a direct exchange method between any two tools, it can be seen as a first step towards a standardisation of different AG-based tool formats.

3.2.3 NXT (NITE XML Toolkit)

NXT (Carletta et al. 2003) is a set of libraries and tools designed to support the representation, manipulation, query and analysis of multimedia language data. NXT is different from the afore-mentioned formats and frameworks because its focus is not on transcription or direct annotation of media data, but rather on the subsequent enrichment of data which has already been transcribed with some other tool. Part of NXT is the NITE Object Model, a data model specifying how to represent data sets with multiple, possibly intersecting hierarchies. Serialization of the data model is done through a set of interrelated XML files in which annotations can point to other annotations in different files (i.e. NXT uses standoff annotation).

3.3 Other formats

There are at least two other formats, which are neither associated with a specific tool nor intended as generic formats, but which are widely cited in the literature and have been applied to larger bodies of multimedia data and may therefore be relevant for the scope of this document:

- **BAS Partitur Format** (Schiel et al. 1998) - The BAS (Bavarian Archive of Speech Signals) has created the Partitur format based on their experience with a variety of speech databases. The aim has been to create "an open (that is extensible), robust format to represent results from many different research labs in a common source." The Partitur-Format is probably relevant only for speech corpora, not for the other types of corpora described above.

- **TIMIT** (Fisher et al. 1986) - TIMIT is a corpus of phonemically and lexically transcribed speech of American English speakers of different sexes and dialects, i.e. a phonetic corpus in the sense defined above. Phonetic (phoneme-wise) and orthographic (word- and utterance-wise) transcriptions are stored in individual tabular-separated text files. Timestamps, i.e. start and end offsets for each transcription unit, can be used to refer the different text files to one another.

3.4 Interoperability of tools and formats

Interoperability between tools and formats, at this point in time, usually means that a converter exists to directly transform one tool's format into that of another (i.e. interoperability is usually *not* achieved via a pivot format or a tool-external standard). In most cases, such converters are built into the tools in the form of an import or an export filter. Filters may be implemented as XSLT stylesheets or as pieces of code in some other programming language. The following table provides an overview of import and export filters integrated in the most widely used tools:

Tool	Imports	Exports
ANVIL	ELAN, Praat	---
CLAN	ELAN, Praat	ELAN, EXMARaLDA, Praat
ELAN	CHAT, Praat, Transcriber, (ShoeBox, Toolbox, FLeX)	CHAT, Praat, (ShoeBox, Toolbox, TIGER)
EXMARaLDA	ANVIL, CHAT, ELAN, FOLKER, Praat, Transcriber, (TASX, WinPitch, HIAT-DOS, syncWriter, TEI, AIF)	CHAT, ELAN, FOLKER, Praat, Transcriber, (TASX, TEI, AIF)
FOLKER	EXMARaLDA	ELAN, EXMARaLDA (TEI)
Praat	---	---
Transcriber	CHAT (ESPS/Waves, Timit, several NIST formats)	Several formats, but none of the ones discussed here

The following matrix pairs off the different tools, showing where a direct interoperability in the form of an import or export filter exists.

	Imports							Exports						
	ANVIL	CHAT	ELAN	EXMARaLDA	FOLKER	Praat	Transcriber	ANVIL	CHAT	ELAN	EXMARaLDA	FOLKER	Praat	Transcriber
ANVIL		-	+	-	-	+	-		-	-	-	-	-	-
CLAN	-		+	-	-	+	-	-		+	+	-	+	-
ELAN	-	+		-	-	+	-	-	+		-	-	+	-
EXMARaLDA	+	+	+		+	+	+	-	+	+		+	+	-
FOLKER	-	-	-	+		-	-	-	-	+	+		-	-
Praat	-	-	-	-	-		-	-	-	-	-	-		-
Transcriber	-	+	-	-	-	-		-	-	-	-	-	-	

Taking transitivity into account (if tools A and B are interoperable, and tools B and C are interoperable, then A and C are interoperable via B), there seems to be, in principle, a way of exchanging data between any two of these seven tools. Furthermore, for some pairs of tools, there is more than one way of exchanging data (e.g. ELAN imports CHAT, and CLAN also exports ELAN). In practice, however, interoperability in its present form has to be handled with great care for the following reasons:

- Information may be lost in the conversion process because the target format has no place for storing specific pieces of information contained in the source format (e.g. when exporting Praat from ELAN, information about speaker assignment will be lost).
- For similar reasons, information may be reduced or structurally simplified in the conversion process (e.g. EXMARaLDA transforms structured annotations into simple annotations when importing certain tier types from ELAN).
- Some converters rely on simplified assumptions about the source or target format and may fail when faced with their full complexity (e.g. CLAN's EXMARaLDA export will fail when the source transcriptions are not fully aligned).

- Since there is no common point of reference for all formats and data models, different conversion routes between two formats will usually lead to different results (e.g. importing Praat directly in ANVIL will not result in the same ANVIL file as first importing Praat in ELAN and then importing the result in ANVIL).

Lossless roundtripping between tools is therefore often not possible, and any researcher working with more than one tool or format must handle interoperability issues with great care. Thus, although the existing interoperability of the tools is useful in practice, a real "standardisation" would still be an important improvement.

3.5 Transcription conventions / Transcription systems

3.5.1 Systems for phonetic transcription

- **IPA** (International Phonetic Alphabet) – The International Phonetic Alphabet can be regarded as one of the longest-standing standards in linguistics. Its development is controlled by the International Phonetic Association. IPA defines characters for representing distinctive qualities of speech, i.e. phonemes, intonation, and the separation of words and syllables. IPA extensions also cater for additional speech phenomena like lisping etc. According to Wikipedia, there are 107 distinct letters, 52 diacritics, and four prosody marks in the IPA proper. Unicode has a code page (0250-02AF) for IPA symbols (IPA symbols that are identical with letters of the Latin alphabet, are part of the respective Latin-x codepages).

- **SAMPA, X-SAMPA** ((Extended) Speech Assessment Methods Phonetic Alphabet, Gibbon et al. 1997, Gibbon et al. 2000) – SAMPA and X-SAMPA are mappings of the IPA into a set of symbol included in the 7-bit-ASCII set. The mapping is isomoprhic so that a one-to-one transformation in both directions can be carried out (see, for instance, http://www.theiling.de/ipa/). Many speech corpora and pronunciation lexicons have been transcribed using SAMPA. As Unicode support in operating systems and applications gains ground, SAMPA and X-SAMPA will probably become obsolete over time.

- **ToBi** (Tones and Break Indices) – „ToBI is a framework for developing community-wide conventions for transcribing the intonation and prosodic structure of spoken utterances in a language variety. A ToBI framework system for a language variety is grounded in research on the intonation system and the relationship between intonation and the prosodic structures of the language (e.g., tonally marked phrases and any smaller prosodic constituents that are distinctively marked by other phonological means)." (quote from http://www.ling.ohio-state.edu/~tobi/). ToBi systems are available or under development for different varieties of English, German, Japanese, Korean, Greek, different varieties of Catalan, Portuguese, Serbian and different varieties of Spanish.

3.5.2 Systems for orthographic transcription

There are numerous systems for doing orthographic transcription. Many of them are language specific or have at least been used with one language only. Also, many of them are documented only sparsely or not at all. The following list therefore includes only such systems for which an accessible documentation exists and which are known to be used by a larger community and/or for larger bodies of data.

- **CA** (Conversation Analysis, Sacks et al. 1978) – In an appendix of Sacks et al. 1978, the authors sketch a set of conventions for notating transcripts to be used in conversation analysis. The conventions consist of a set of rules about how to format and what symbols to use in a type-written transcript. They have been transferred later to be used with text-processors on computers, but there is no official documentation of a computerized CA, let alone a document specifying the symbols to be used as Unicode characters. Although never formulated in a more comprehensive manner, the CA conventions have been widely used and have inspired or influenced some of the systems described below.

- **CHAT** (Codes for the Human Analysis of Transcripts, MacWhinney 2000) – Besides being a text-based data format (see above), CHAT is also a transcription and coding convention. Analogous to the CLAN tool, it was originally developed for the transcription and coding of child language data, but now also contains a CA variant for use in conversation analysis (in a way, this *could* be seen as the (or one) computerized variant of CA - see above). Since it is so closely tied to the CHAT format and the CLAN tool, many aspects relevant for computer encoding (e.g. Unicode compliancy) have been treated in

sufficient detail in the conventions. A special system for the transcription of bilingual data, LIDES (Barnett et al. 2000), was developed on the basis of CHAT.

- **DT/DT2** (Discourse Transcription, DuBois et al. 1993) – DT is the convention used for transcription of the Santa Barbara Corpus of Spoken American English. It formulates rules about how to format and what symbols to use in a plain text transcription, including timestamps for relating individual lines to the underlying recording. DT2 is an extension of DT. It contains a table which specifies Unicode characters for all transcription symbols.

- **GAT/GAT2/cGAT** (Gesprächsanalytisches Transkriptionssystem, Selting et al. 2009) – GAT is a convention widely used in German conversation analysis and related fields. It uses many elements from CA transcription, but puts a special emphasis on the detailed notation of prosodic phenomena. The original GAT conventions explicitly set aside all aspects of computer encoding of transcriptions. To a certain degree, this has been made up for in the recently revised version, GAT 2. cGAT is based on a subset of the GAT 2 conventions and formulates explicit rules, including Unicode specifications of all transcription symbols, for computer-assisted transcription in the FOLKER editor (see above).

- **GTS/MSO6** (Göteborg Transcription Standard, Modified Standard Orthography, Nivre 1999 and Nivre et al. ????) – According to its authors, GTS is a "standard for machine-readable transcriptions of spoken language first used within the research program Semantics and Spoken Language at the Department of Linguistics, Göteborg University." It consists of two parts, one language independent part called GTSG (GTS General), and one language dependent part. The MSO. GTS, however does not necessarily require MSO. GTS in combination with MSO is the basis for the Göteborg Spoken Language Corpus.

- **HIAT** (Halbinterpretative Arbeitstranskriptionen, Ehlich/Rehbein 1976, Ehlich 1993, Rehbein et al. 2004) – HIAT is a transcription convention originally developed in the 1970s for the transcription of classroom interaction. The first versions of the system (Ehlich/Rehbein 1976) were designed for transcription with pencil or typewriter and paper. HIAT's main characteristic is the use of so-called Partitur (musical score) notation, i.e. a two-dimensional transcript layout in which speaker overlap and other simultaneous actions can be represented in a natural and intuitive manner. HIAT was computerized relatively early in the 1990s in the form of two computer programs – HIAT-DOS for DOS (and later Windows) computers, and syncWriter for Macintoshes. However, standardization and data exchange being a minor concern at the time, these data turned out to be less sustainable than their non-digital predecessors. The realisation in the HIAT community that data produced by two functionally almost identical tools on two different operating systems could not be exchanged and, moreover, the prospect that large existing bodies of such data might become completely unusable on future technology was one of the major motivations for initiating the development of EXMARaLDA. The most recent version of the conventions (Rehbein et al. 2004) therefore contains explicit instructions for carrying out HIAT transcription inside EXMARaLDA (or Praat).

- **ICOR** (Interaction Corpus, http://icar.univ-lyon2.fr/documents/ICAR_Conventions_ICOR_2007.doc) – ICOR is the transcription convention used for transcriptions in the French CLAPI database. As formulated in the cited document, it is a convention for producing plain text files. However, the fact that CLAPI offers TEI versions of all ICOR transcripts shows that there is a conversion mechanism for turning ICOR text transcriptions into XML documents.

3.5.3 Systems for sign language transcription

- **HamNoSys** (Hamburger Notationssystem für Gebärdensprachen, Prillwitz et al. 1989)
- Stokoe notation

3.5.4 Commonly used combinations of formats and conventions

Although, in principle, most of the tools described in section 3.1 could be used with most transcription systems described in this section, most communities in the humanities have an outspoken preference

towards a specific combination of tool and transcription system. The following lists some widely used such combinations:

- CLAN + CHAT is widely used for doing transcription and coding of child language,
- CLAN + CA is widely used for doing transcription in conversation analysis,
- EXMARaLDA + HIAT is widely used for doing transcription in functional-pragmatic discourse analysis,
- FOLKER + GAT is widely used for doing transcription in interactional linguistics, German conversation analysis and related fields,
- ICOR + TEI (though strictly speaking not a combination of a convention and a tool) is the basis of the French CLAPI database.

4 Summary / Recommendations

4.1 Media Encoding

4.2 Media Annotation

There is, to date, no widely dominant method, let alone an official standard, for doing media annotation. Considering the heterogeneity of the corpus types and research communities involved, this is probably not too surprising. However, the number of tools and formats actually used has a manageable proportion. Moreover, there are obvious conceptual commonalities between them, and they interoperate reasonably in practice. Using these (i.e. the first seven discussed in section 3.1.) tools as a basis, and combining the approaches of AG and TEI, a real standard, suitable to be used in a digital infrastructure, does not seem to be an unrealistic goal.
Until such a standard is defined, however, researchers doing media annotation need recommendations about which tools and formats to use in order to maximize their chances of being standard compliant eventually. The following criteria may serve as guidelines for such a recommendation:

- The format should be XML based, since XML has become the virtually unrivalled solution for representing structured documents and further processing will be easier if files are in XML.
- The tool and format should support Unicode, since only this – equally unrivalled standard – ensures that data from different languages can be reliably exchanged.
- The format should be based on a format-independent data model (or it should be possible to relate the format to such a data model) since this greatly facilitates the integration of data into an infrastructure which may require diverse physical representations of one and the same piece of data.
- The tool should be stable, its development should be active, since the definition of a standard or the integration of a tool (format) into an infrastructure may require adaptations of the tool or at least information from its developer(s).
- The tool should run on different platforms, since the targeted user community of the infrastructure will also have diverging platform preferences.
- It should be possible to import the tool format into or export it to other tools, since this demonstrates a basic ability to interoperate.

According to these criteria, there are at least four tools in the above list which can be **recommended without restrictions**, i.e. they meet all of the criteria:

- **ANVIL**
- **ELAN**
- **EXMARaLDA**
- **FOLKER**

From the more widely used tools, three remain which do not meet all of the criteria, but which can still be **recommended with some restrictions**:

- **Praat** – the only objection to Praat is that its data format is not XML based. This makes an initial transformation of Praat data somewhat more difficult. However, since the format itself is rather simple and the underlying data model well understood, and since, furthermore several tools provide converters for transforming Praat data into XML, this objection is a minor one.
- **CHAT** – CHAT's main drawback is the lack of an explicit data model, making CHAT data somewhat more closely tied to, and more dependent on, the application they were created with. However, considering CHAT's wide user base and the huge amounts of CHAT data available in CHILDES and Talkbank, it would probably not be a good idea to exclude CHAT from a standardisation effort. Some of the problems arising from the lack of a data model can be alleviated by making full use of the CLAN tool's internal checking mechanisms (see below) and maybe also by transforming CHAT data into the Talkbank XML format.

- **Transcriber** – Transcriber's main problem is that the development is not active anymore. A new version (based on annotation graphs) has been announced for quite some time now, but the envisaged release data (2nd quarter of 2009) has long passed without a new version having been released. However, the tool has been sufficiently used in practice and its format seems to be sufficiently well understood also by other applications to make it worthwhile considering Transcriber in a standardisation effort.

For similar reasons (i.e. development not active), some of the tools can only be **recommended with severe restrictions**:

- **TASX**'s development seems to have been abandoned, the tool itself not officially available anymore
- **AG toolkit**'s development was abandoned at a relatively early stage
- **WinPitch**'s development status is unclear

Otherwise, however, these three tools meet most of the criteria specified above.

A few of the tools mentioned in 3.1.8 **disqualify** for a standardisation effort, because their file formats lack a sound structural basis. This is the case for **Transana** as well as for **F4** and similar tools.[3]

Two further recommendations can be made to data creators or curators:
First, they should be encouraged to use the full palette of tool-internal mechanisms for ensuring consistency, validating data structures, etc., since this is likely to increase the reliable processability and exchangeability of data. More specifically, this means:

- In **ANVIL**, specification files should be carefully designed and used consistently.
- In **CLAN**, the check command should be used on completed data, and a Talkbank XML version should be generated for completed data.
- In **ELAN**, linguistic types for tiers should be carefully designed and used consistently. If possible, linguistic types should be associated with categories of the ISOCat registry.
- In **EXMARaLDA**, basic transcriptions should be checked for structure errors and segmentation errors. A segmented transcription should be generated for completed basic transcriptions.
- In **FOLKER**, the mechanism for checking the syntax and temporal structure of transcriptions should be enabled.

Second, they should be encouraged to use an established (i.e. tested and document) combination of tools and conventions as described in section 3.5.4. Where none of the established combinations meets their demands, new conventions should be developed as extensions or modifications of existing ones.

[3] The remaining tools described in 3.1.8, but not yet touched on in this section (Wavesurfer, EMU, Phon and XTrans), are probably all more on the "recommendable" side of the scale, but I (TS) do not feel competent to judge this. It should maybe also be mentioned that there are some legacy tools around whose development was abandoned 10 years or more ago and which can, for all practical purposes, be regarded as obsolete. However, larger bodies of data still exist which were created with these tools and which may have to be curated for integration into a digital infrastructure. **syncWriter**, **HIAT-DOS** and **MediaTagger** are three such tools (see Schmidt/Bennöhr 2008 for a description of the first two).

5 References

[Barnett et al. 2000]
Barnett, R.; Codo, E.; Eppler, E.; Forcadell, M.; Gardner-Chloros, P.; van Hout, R.; Moyer, M.; Torras, M.; Turell, M.; Sebba, M.; Starren, M.; Wensing, S. (2000): The LIDES Coding Manual: A Document for Preparing and Analyzing Language Interaction Data Version 1.1--July 1999. Special issue of International Journal of Bilingualism (4,2), 131-271.

[Bird/Liberman 2001]
Bird, S. & Liberman, M. (2001): A formal framework for linguistic annotation. In: Speech Communication 33, 23-60.

[Braffort et al. 2004]
Braffort A., Choisier A., Collet C., Dalle P., Gianni F., Lenseigne B., Segouat J. (2004) : toward an annotation software for video of sign language, including image processing tools and signing space modelling. In Proc. of 4th International Conference on Language Resources and Evaluation - LREC 2004, volume 1, p. 201–203, Lisbon, Portugal.

[Carletta et al. 2003]
Carletta, J.; Evert, S.; Heid, U.; Kilgour, J.; Robertson, J.; VoormannH. (2003): The NITE XML Toolkit: flexible annotation for multi-modal language data. Behavior Research Methods, Instruments, and Computers, special issue on Measuring Behavior, 35 (3), 353-363.

[DuBois et al. 1993]
DuBois, J.; Schuetze-Coburn, S.; Cumming, S. & Paolino, D. (1993): Outline of Discourse Transcription. In: Edwards, J. A. & Lampert, M. D. (ed.): Talking Data: Transcription and Coding in Discourse Research, 45-89. Hillsdale, NJ: Erlbaum.

[Ehlich/Rehbein 1976]
Ehlich, K. & Rehbein, J. (1976): Halbinterpretative Arbeitstranskriptionen (HIAT). In: Linguistische Berichte 45, 21-41.

[Ehlich 1993]
Ehlich, K. (1993) HIAT: A Transcription System for Discourse Data. In: Edwards, J. A. & Lampert, M. D. (ed.): Talking Data: Transcription and Coding in Discourse Research, 123-148. Hillsdale, NJ: Erlbaum.

[Fisher et al. 1986]
Fisher, W.; Doddington, G.; Goudie-Marshall, K. (1986): The DARPA Speech Recognition Research Database: Specifications and Status. Proceedings of DARPA Workshop on Speech Recognition. pp. 93–99.

[Gibbon et al. 1997]
Gibbon, D.; Moore, R.; Winski, R. (eds.) (1997): Handbook of Standards and Resources for Spoken Language Systems. Berlin: Mouton de Gruyter.

[Gibbon et al. 2000]
Gibbon, D.; Mertins, I.; Moore, R.; (eds.) (2000): Handbook of Multimodal and Spoken Language Systems: Resources, Terminology and Product Evaluation. Boston: Kluwer Academic Publishers.

[Hanke/Storz 2008]
Hanke T., Storz J. (2008): ILEX - a database tool for integrating sign language corpus linguistics and sign language lexicography. In Proc. of 6th International Conference on Language Resources and Evaluation, LREC 2008, p. W25–64–W25–67, Marrakesh.

[MacWhinney 2000]
MacWhinney, B. (2000): The CHILDES project: tools for analyzing talk. Mahwah, NJ: Lawrence Erlbaum.

[Nivre 1999]
Nivre, J. (1999): Modified Standard Orthography (MSO6). Department of Linguistics, Göteborg University.

[Nivre et al. ????]
Nivre, J.; Allwood, J.; Grönqvist, L.; Ahlsén, E.; Gunnarsson, M.; Hagman, J.; Larsson, S.; Sofkova, S. (????): Göteborg Transcription Standard. Version 6.3. Department of Linguistics, Göteborg University.
[http://www.ling.gu.se/projekt/tal/doc/transcription_standard.html]

[Parisse/Morgenstern 2010]
Parisse, C.; Morgenstern, A. (2010): A multi-software integration platform and support for multimedia transcripts of language. In: Proceedings of the LREC 2010 workshop 'Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality', 106 – 110;

[Prillwitz et al. 1989]
Prillwitz, S. et al (1989): HamNoSys. Version 2.0; Hamburger Notationssystem für Gebärdensprache. Eine Einführung. (Internationale Arbeiten zur Gebärdensprache und Kommunikation Gehörloser; 6) Hamburg : Signum 1989 - 46 S.

[Rehbein et al. 2004]
Rehbein, J.; Schmidt, T.; Meyer, B.; Watzke, F. & Herkenrath, A. (2004) Handbuch für das computergestützte Transkribieren nach HIAT. In: Arbeiten zur Mehrsprachigkeit, Folge B (56). [http://www.exmaralda.org/files/azm_56.pdf]

[Sacks et al. 1978]
Sacks, H.; Schegloff, E. & Jefferson, G. (1978) A simplest systematics for the organization of turn taking for conversation. In: Schenkein, J. (ed.): Studies in the Organization of Conversational Interaction, 7-56. New York: Academic Press.

[Schiel et al. 1998]
Schiel, F.; Burger, S.; Geumann, A. & Weilhammer, K. (1998) The Partitur Format at BAS. In: Proceedings of the First International Conference on Language Resources and Evaluation, 1295-1301. Paris: ELRA.

[Selting et al. 2009]
Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., Hartung, M., Kern, F., Mertzlufft, C., Meyer, C., Morek, M., Oberzaucher, F., Peters, J., Quasthoff, U., Schütte, W., Stukenbrock, A., Uhmann, S. (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: Gesprächsforschung (10), pp. 353-402,

[Schmidt 2005a]
Schmidt, T. (2005a): Computergestützte Transkription - Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln. Frankfurt a. M.: Peter Lang.

[Schmidt 2005b]
Schmidt, T. (2005b): Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech. Arbeiten zur Mehrsprachigkeit, Folge B (62). [http://www.exmaralda.org/files/SFB_AzM62.pdf]

[Schmidt/Bennöhr 2008]
Schmidt, T. & Bennöhr, J. (2008): Rescuing Legacy Data. In: Language Documentation and Conservation (2), 109-129.

[Schmidt et al. 2008]
Schmidt, T.; Duncan, S.; Ehmer, O.; Hoyt, J.; Kipp, M.; Magnusson, M.; Rose, T. & Sloetjes, H. (2008): An exchange format for multimodal annotations. In: Proceedings of the Language and Evalutation Conference 2008

[Schmidt et al. 2009]
Schmidt, T.; Duncan, S.; Ehmer, O.; Hoyt, J.; Kipp, M.; Magnusson, M.; Rose, T. & Sloetjes, H. (2009) An Exchange Format for Multimodal Annotations. In: Michael Kipp, Jean-Claude Martin, P. P. & Heylen, D. (ed.): Multimodal Corpora, Lecture Notes in Computer Science 207-221. Springer.

[Schmidt/Schütte 2010]
Schmidt, T. & Schütte, W. (2010) FOLKER: An Annotation Tool for Efficient Transcription of Natural, Multi-party Interaction. In: Nicoletta Calzolari (Conference Chair), Khalid Choukri, B. M. J. M. J. O. S. P. M. R. D. T. (ed.): Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta: European Language Resources Association (ELRA).