About Annotation Schemes and Terminology

P. Wittenburg
Max-Planck-Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen
peter.wittenburg@mpi.nl

Introduction

A new EAGLES initiative - International Standards for Language Engineering (ISLE) - was recently funded. It is meant to cover not only Language Engineering but also those linguistic disciplines which exploit Language Resources in general. One of the aims of this initiative is to discuss "annotation schemes specifically for the fields of natural interaction and multi-modal research and development" and - if possible - to develop guidelines for such schemes. While determining linguistic content and the way it ought to be encoded seems to be a very difficult task, choosing adequate structures to represent such encodings seems to be comparatively simple. So the question is, why should this topic be of interest to the language resource community? This paper will try to answer this question, describe the increasing complexity of the subject, and discuss the terminology involved.

We can identify a number of different Language Resources (LR) such as corpora[1], lexicons, ontologies, grammar notes, and notes about the sound systems of a language, amongst others. In this paper about annotation schemes we will focus on corpora, i.e. language resources where the primary content is either a sequence of characters or a sequence of samples taken from human output. Samples can be scalar numbers representing a sound wave or matrices of numbers representing for example a sequence of video frames. Associated with these sequences of characters or samples are descriptions of various types, which we call annotations. We follow the terminology used by Bird & Liberman [1]. It should be noted that lexicons for example exhibit some similarity with respect to the basic structure. Normally, the raw materials are wordforms or fixed expressions extracted from a set of corpora, which are "annotated", of course, with specific content. Nevertheless we restrict ourselves in this paper to corpora, although some of the statements made in this paper also apply to other language resources.

The major force driving the community to accumulate more, and more complex language resources, is the increasing power of information technology. We are able to collect, store, and access huge amounts of textual material and in recent years it has become feasible to digitize audio/video signals with standard equipment, store them, and access the data with computers. Progressively more complex software applications have become available to ease computer access. Researchers and developers are beginning to exploit the new possibilities, such as immediate access to raw data. Computerised access to raw data requires that the information is systematically structured. If we can agree on common structures it will become much easier to re-use these resources outside the areas for which they were collected. Given how much it costs to assemble a language resource, this is an opportunity that we must not miss.

Development Phases

First we would like to briefly describe the major steps in the development of corpora. In a first phase of computer-assisted corpus-development the corpora were text based - mostly orthographic or phonemic transcriptions plus the annotations associated with them or based on written text. From a computational perspective these corpora could be characterised by

1. a proprietary format suitable for the project in mind,
2. an ASCII representation of all phenomena,
3. an identity of file and presentation format, and

[1] A corpus is a collection of structured textual material describing either basic texts or multi-media recordings. The atomic unit of a corpus is one such raw text or one such recording together with its resources describing the raw content.

4. idiosyncratic solutions to encoding certain timing phenomena, such as overlap in speaking and interruptions which were lost due to the nature of the transcription, but should be preserved for analysis.

Typical representatives of this phase are the famous CHILDES corpora including its CHAT format specification [2] and the well-known SHOEBOX tool from SIL [3] mainly constructed for field workers such as anthropologists. Choices that were rational at that time, mainly determined by the computer technology then available, have left these formats with a number of severe restrictions. CHILDES can't be applied to many studies simply because its basic unit is the utterance. In real conversations there is often no such unit as an utterance, but a continuum of related actions. This is just one major deficit. SHOEBOX ended up as a cleverly designed program providing excellent support for many aspects of anthropological field work. Despite this, its basic design is restricted to text representation and it is difficult to include raw audio and video.

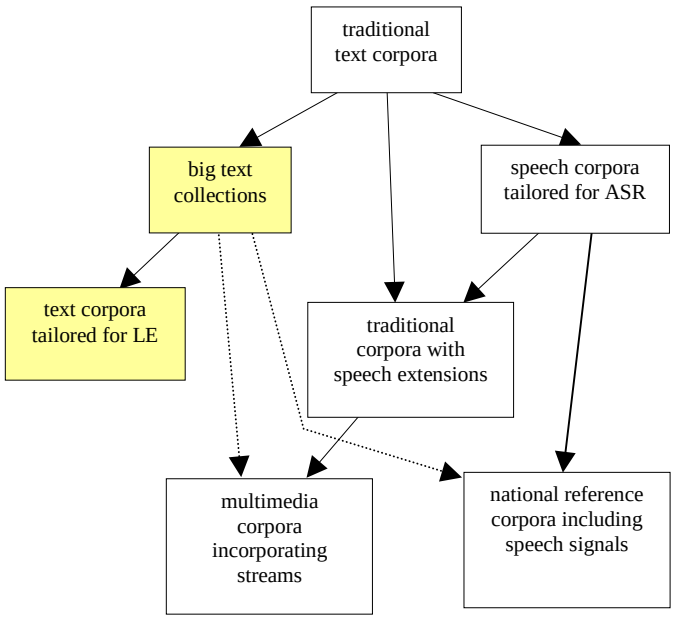


Figure 1 gives an overview of the major phases in corpus development. Traditional textual corpora such as CHILDES were in general small, hand-encoded representations of the results of scientific investigations. When publishers recognized the potential of computers they started to assemble big collections of highly structured texts. Some of these were elaborated with extensive, (semi) automatically added annotations. At the same time, the community of speech engineers needed specially tailored corpora covering audio recordings and closely linked transcriptions, i.e. simple annotations. Language engineers automatically creating annotations also needed specially tailored formats. After a phase of extending traditional formats with speech, it became obvious that there was a requirement for a true multi-media format which could combine several independent data streams. National reference corpora now under development are largely comparable with the big text collections except that the link with the speech signal has to be added.

A second phase of corpora development can be identified with the need to handle large amounts of data and support more complex structuring of the data. Language engineers, publishers and others started to assemble large text corpora as complex hierarchical structures. The interested community understood that to facilitate a variety of activities, such as efficient access to structure elements, long-term documentation, and remote access, it was necessary to adopt appropriate formalisms. SGML was chosen to serve these functions. The Web community has evolved SGML into the slightly simpler XML standard, which is now preferred. Part of the Language Engineering community was looking for a more efficient storage format, which their linguistic programs - such as parsers and POS-taggers - could access for reading and writing. They created the Tipster format [4] which was implemented using relational database technology. Both, SGML/XML and Tipster were and are heavily used by science and industry.

Another branch in this second phase is represented by speech engineers, who had to train their statistical speech recognisers. In general they needed fairly simple data structures where orthographic and/or phonemic transcriptions were tightly linked to the speech signal. Parts of the corpus structure were put into the file hierarchy. TIMIT [5] is a classical example.

In linguistic research, individual researchers started looking for immediate access to the original speech signal when working with a traditional corpus which might be annotated in - say - the CHAT format. Carnegie-Mellon-University (CHILDES) and the Max-Planck-Institute (ESF Corpus [6]) devised

suitable tools, but the solutions based on an old and restricted format were too restricted to be seen as anything but a stop-gap.

The third and current phase of corpora development involves the design of two new types of corpora and tools.

1. One type is based on multi-media information where various independent streams such as speech, gesture, facial expression, and eye movements amongst others can be annotated and linked. The gesture corpus tool developed at the Max-Planck-Institute can be seen as one example [7].

2. Another type is focussed on building large corpora including speech signals which are intended to serve as national reference corpora for speech engineers as well as for linguists and language engineers. The well-known British National Corpus can be taken as the first and architypical example [8].

Both types of corpora are complex. In multi-media based corpora, complicated time relationships between the annotations in up to 40 or 50 manually created annotation layers demand a specialised new structure. For the national reference projects, the sheer volume of the semi-automatically generated structured data presents the crucial problems. SGML/XML remains the preferred mark-up format for these reference corpora. Compared to the early text-based text collections, the reference projects only the inclusion of the speech signal has to be solved. Multi-media corpora, however, required new types of annotation schemes had to be invented.

Also the media industry and sciences involved in related research projects are actively discussing their needs for annotating broadcast material. The evolving MPEG7 standard [9] is intended to describe the hierarchical structure of such movies down to automatically extracted features describing low-level aspects of single frames. Since MPEG7 will serve different needs than those of the community dealing with language resources, we will not discuss it in detail.

In this paper we will not elaborate on the very important work which has been done by a variety of different groups such as TEI, CES, and EAGLES amongst others, to devise standards for tag sets (labels) for structural elements and for encoding schemes to annotate linguistic content such as morphology. It should be noted that all these standards will be in use for the foreseeable future for both reference corpora and multi-media corpora. Where researchers are aware of these standards they should be applied - granting that they are applicable. The Max-Planck-Institute will be using these standards in as many of its of its multi-media projects as it can, implementing them as constraints for the data entry module. This approach is essential if we are to build up a body of generally accessible corpora. For some projects - as in, for example, setting up a corpus to describe an endangered language - the standards available are inadequate.

Multi-Media Corpora

In the following we focus on multi-media corpora, since they incorporate the greatest amount of inherent complexity, since the structure of textual collections has been studied for a long time now, and since the joint EC/NSF project is about corpora in natural interaction/multi-modal environments. What we intend is to briefly indicate some dimensions of complexity which we are confronted with in such multi-media corpora.

In multi-media corpora we use time series such as audio or video recordings as the primary data. These time series form a stable - unalterable - basis for all the annotations. All annotations including the transcription may be subject to change, because the scientific question could change, requiring additional coding. Media time is assumed to go 0 to some minutes or hours and represents a continuum. Here, we assume that we have solved the problem of having created only one common time axis and can aligne all media streams with this axis. To achieve this is not a trivial task, but will not be discussed here in detail. In general the annotations then are stretched from the beginning to the end of the recording. Any definition of a linguistic unit is temporary, i.e. specific to one specific project. We must therefore assume that users will want to be flexible when linking their annotations with the original media.

In order to continue we first have to introduce some basic terminology. The terms chosen may be subject of discussion. Annotation is a process which will happen on various layers or tiers. The

orthographic transcription of speech may be one tier as could the morphologic or syntactic description of what has been said. There could be other tiers to code gestures or facial expressions and so on. A single chunk of information annotated in some tier and linked with media time is called a tag. Events are seen as tags with a short time interval.

The user annotates independent streams of information which can be extracted from the primary data. This independence of streams has all sorts of consequences, including that all sorts of timing relations between tags are possible. This point was made by Brugman [10] when he was designing his first multi-media tool and was recently exhaustively described by Bird&Liberman [1]. Both groups have pointed out that tags can be parts of hierarchies. While Brugman mentions types of dependencies, Bird&Liberman speak about three types of hierarchies: "token-based", "type-based" and "graph-based".

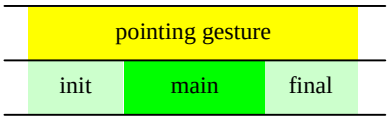


Figure 2 shows an example of two tiers which exhibit a type-based hierarchy.

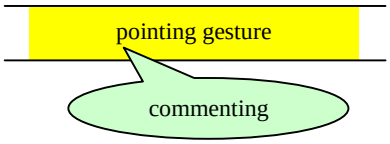


Figure 3 shows an example of a tag to which a comment is associated.

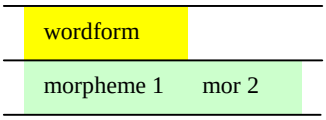


Figure 4 shows an example of a SHOEBOX like hierarchy

When for example gestures are coded, a single "gesture" would be split into "gesture phases", i.e. by definition the left-most and the right-most time are as indicated in figure 2. Additionally, it makes sense to require that tags on one tier are not overlapping. B&L cite the TIMIT example where /she/ spans the interval from t1 to t3, /sh/ from t1 and t2, and /iy/ from t2 to t3. Another type of hierarchy can be identified where, for example, coders want to associate comments to a certain tag. Typically one will need reference structures to implement such hierarchies. In SHOEBOX corpora token-based hierarchies are represented by using the column structure (s. figure 4).

A completely different type of hierarchy could be part of the encoding such as the annotation of syntactical relationships with the help of bracket structures. Here, it is the task of a specialised tool to carry out interpretations of such encodings and for example to visualize the implicit hierarchies. But the bracket structures are not part of the annotation structure. However, for example in morphological annotations it might be necessary to express a relation between two wordforms with help of general mechanisms. It is not yet clear where exactly the boundary between general mechanisms anchored in the annotation scheme and special tools are. To be able to represent linguistic dependencies or relations the annotation scheme therefore has to provide the possibility to set references between tags within one tier. As B&L also have pointed out there have to be relations between structure elements on different tiers. Of course, all these references will in general be associated with labels.

New projects to develop multi-media annotation and exploitation tools such as EUDICO [11] which started in 1997 and Talkbank [12] which started in 1999 speak about direct user collaboration during the process of annotation. The following scenario is envisaged: user X is sitting at one location and is working at a certain tier and in parallel user Y is coding another tier at another location. They both want to see each others annotation since it might affect their own way of coding. They also want to write and store comments which probably refer to each other and to some tag or set of tags in the annotation structure. This increases complexity in two ways: (1) We possibly might end up with multiple tiers of the same type (such as two transcriptions) differing, however, with respect to their codes and tier structure. The number of tiers will increase and one can expect that cross-references will be used. (2) The number of comments will increase and references have to support this as indicated in figure 3. Comments themselves can be subject of being refered to.

During implementation another problem arises: users are creating temporary annotations which have to be separated from those which are checked and treated as being accepted. The underlying annotation mechanisms have to anticipate this type of usage.

In fact most of these phenomena which can occur have been thoroughly investigated by and described in B&L [1].

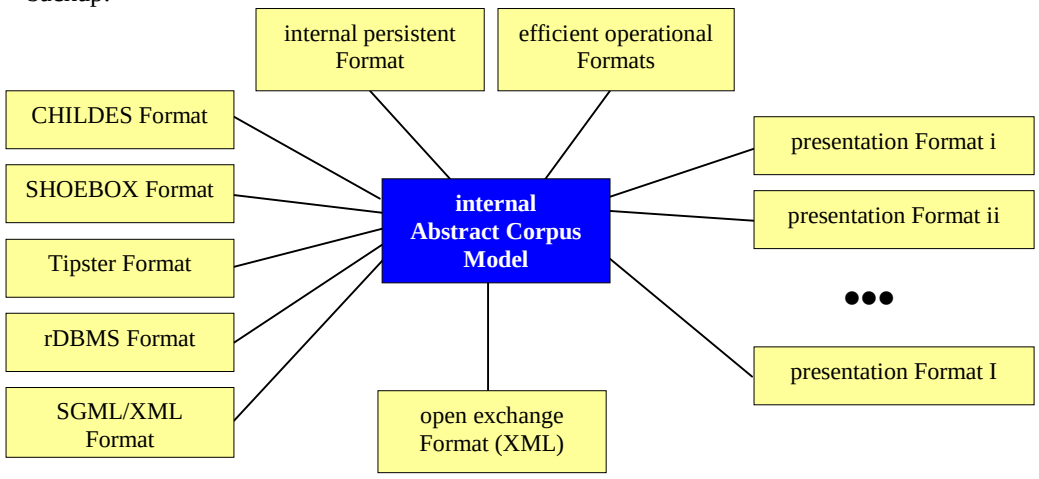
4. Layers of Representation

As mentioned above the first computer based corpora formats such as CHILDES and SHOEBOX, collapsed the file format and the presentation format. This is not a viable concept anymore for a variety of reasons:

(1) Data complexity has increased to such an extent that it is difficult to restrict visualization to precisely those phenomena the user is interested at a given moment.

(2) File structures have to represent the whole complexity of the annotation in such a way that consistent editing is supported and that annotation can easily be imported, which is not always compatible with immediate display.

(3) For long-term documentation purposes there is a strong preference for exporting data as XML-based file . But while XML-structured files are meant to be human-readable, this is just a last-ditch backup.



In Figure 5 the rich topology of representation layers in EUDICO is shown as an example. Native import formats are supported to provide connections to existing and widely used corpora, while operational and internal formats are used to improve efficiency, with an open exchange format for documentation and exchange purposes, and various visualization formats to support the user. All these formats inter-convert via a common object model which is claimed to be powerful enough to model all relevant details.

(4) For various processing purposes we have to rewrite the data into specialised optimal structures. For example, some programs generate full-text indexes to speed searches of large files.

So - farewell to the good old days, where finding suitable corpus representations was comparatively simple. All the major new projects are developing different type of representations. MATE [13] initially imports data into an XML-based representation. For processing this data is reorganised into a relational database, to facilitate searches. EUDICO [11] (see also figure 5) uses a general internal object oriented representation (ACM = Abstract Corpus Model), an internal persistent format to store the state of the software as necessary, an API to be able to import and operate on major existing file formats, and the project discusses the necessity to support an open exchange format, and various presentation/visualization formats dependent on the users' needs. TalkBank's documentation is not yet very explicit, but they do plan to support a variety of native corpus formats, all recoded into the CODON layer, which is to act as a general format to represent all kinds of information, and they talk about various visualization formats required to support the user. Bird & Liberman are involved in TalkBank and CODON is based on their annotation graphs as presented in [1].

For large reference corpora where SGML/XML is to be used to structure the data and ensure well-formedness and consistency, other formats will have to be generated to support efficient searching or to give the user different views of the data. The Dutch-Flemish National Corpus Project intends to use EUDICO representation layers for exploitation.

This chapter shows how the original all-purpose single corpus format has now been replaced by a complex landscape of related formats design to meet a variety of needs, to omit irrelevant elements or to organize the data appropriately for various specific applications.

5. About Annotation Terminology

Now that we have described the more complicated structures of current multi-media based corpora and answered our original question about why we bother about annotation structure, we now want to discuss some of the terminology used in the various contributions.

Fundamental to our discussion is what Bird and Liberman called the logical structure of linguistic annotations [1]. The term is comparable with the term "corpus structure", if more precise. As they explain, this logical structure has to describe all relevant structural phenomena which can occur in such annotations. It acts as a kind of reference basis for all further steps. The term "corpus model" comes close, in that it has to be based on the logical structure. But it focuses on operational aspects such as class hierarchies (in the case of object-oriented modeling) which transform the structural phenomena into an operational model and thus make the assumptions testable. "Common or abstract models" add the claim of being universal in that they model all structures and operations which can occur. A common model therefore is the operational nucleus of any system/tool which claims to be format-independent, i.e. a tool which claims to be able to map all relevant corpus formats into its internal representations.

The term "annotation architecture" seems to be used in two ways: (1) Some apply this term to the high-level structure of annotations. (2) Others use this term in relation to a tool which offers the possibility of defining the structure of the annotations. In both cases, the term is related to the overall structure of annotations or an annotation system. The term "annotation scheme" is seen as being almost identical with the term annotation architecture. It is seen as a blue-print of the basic structure of the annotations. It might be that annotation schemes also include high-level notes about the encoding principles. Sometimes the term "data architecture" is used instead of "annotation architecture". It refers to the same subject, but is less biased towards a linguistic context. The term "framework" appears in two contexts: (1) B&L [1] use the term "formal framework" to describe an exhaustive logical structure for annotations, to give a useful picture of all the phenomena which can and will occur. (2) The term "annotation framework" is used in an operational sense, i.e. as a tool - an annotation template - which allows one to enter concrete annotations. Such tools may not allow the user to specify the annotation structure and the encoding constraints and may not include data entry checks to ensure that the data entered is compliant with these specifications. Given that engineers generally understand a "framework" to mean something into which you can plug compatible modules, the term "annotation framework" is irritating.

We have to differentiate a group of terms which describe the way certain linguistic phenomena (such as morphology or syntax) are annotated. The term "annotation scheme" has already been mentioned. It covers statements about the structure of annotations as well as high-level notes about the encoding principles. The term "annotation schema" is seen as covering all important details about the encoding. The term is taken from database literature where it describes not only the structure of a database, but also the labels and types of the attributes and their range of value. "Encoding schemes" or "encoding schemas" are those elements which describe the content of linguistic annotations: the labels to be used to describe linguistic phenomena such as part-of-speech, and the codes to be used to describe the part-of-speech category to which the object referred to belongs. While the term "annotation" in this context includes structural descriptions, the term "encoding" here only refers to the coding of linguistic content.

The term "linguistic object" is used in the context of linguistic annotations in corpora to denote all the items which appear as referable units. These can be elements of the structure such as tiers but also elements of the encoding such as a wordform in an orthographic tier. A distinction should be made between objects which are identified by structural definitions, and those which become referable units as a consequence of the way lower-level encoding is interpreted by a higher level tool. In CHILDES a wordform is not a referable unit on structural level, since it is part of a string which stands for an utterance. Only when we interpret the string and carry out segmentations by using a set of delimiters (tokenization) can we identify a wordform as a unit. In this case the encoding process is required to provide the mechanism that identifies such a unit, because the delimiters are constructs defined at the annotation structure level.

The traditional term "format" with its different flavors such as "annotation format", "file format", "presentation format", "open exchange format" etc all are associated with some way of representing annotation structures on external media such as disks, displays, or even paper. Of course, the complexity of an annotation architecture has to be represented somehow on media to make it accessible

to others or to a user working on a computer. But there are many ways to present or store a given complex annotation. A special term is the “open exchange format” since it refers to a representation which is largely/fully specified by international standards, so designed that everyone can interpret the representation based on the specification notes of those standards and that this interpretation may still be practical many years after the data has been encoded. In general, a snapshot exported to an open exchange format is seen as a static corpus, in contrast to corpora which are continually being modified and extended.

A third group of terms has already been discussed in chapter 3. The terms “layers of representations”, “layered data architectures”, “layered representation system”, “multi-level architectures” etc all refer to the fact that current systems mostly work with different types of representations of a given annotation. The reasons for this split lies in the complexity of the annotation schemes. For visualization we need types of representation that differ from those suitable for fast searches or for documentation purposes. Only some of these representations have as yet been specified by standards.

6. Summary

We have described a few major phases in the development of corpora types and argued that current discussion about annotation structures is dominated by two groups of specialists, those who have to build "national reference corpora" which include speech signals, and those who are creating multi-media corpora incorporating the annotation of independent data streams. The multi-media corpora have more complex structures and call for new approaches to corpus design and handling. The complexity has been thoroughly investigated by a few groups and suggestions have been made for annotation schemes and models to handle and represent this complexity.

The complexity of the annotations in multi-media corpora also has consequences in so far as the old identity between file format and visualization format is being broken up into multi-layered representations. Each type of representation is optimized for its particular application, i.e. for visualization the annotations will be structured in a different way than for running statistical operations on them or for documentation purposes. We have also tried to clarify many of the terms dealing with annotation structures which can be found in literature.

For the joint EC/NSF project we see the need for more thorough discussion of all the aspects mentioned in this paper, and the need to establish guidelines which can be used by the people who have to build multi-media corpora. The consequences for tool architecture also have to be discussed to allow us to establish a list of requirements. Although the paper doesn't focus on encoding schemes we also need to discuss whether we need an EAGLES-like efforts to standardize encoding schemes for multi-modal annotations encoding behaviors like gesture and facial expression.

7. References

[1] S. Bird, M. Liberman 1999: *A Formal Framework for Linguistic Annotation*, Technical Report MS-CIS-99-01, University of Pennsylvania

[2] B. MacWhinney 1991: *The CHILDES Project: Tools for Analyzing Talk*, Hillsdale NJ: Lawrence Erlbaum Associates.

[3] Shoebox: http://www.sil.org/computing/catalog/shoebox.html

[4] Tipster: http://www.itl.nist.gov/iaui/849.02/related_projects/tipster

[5] TIMIT: http://morph.ldc.upenn.edu/catalog/LDC93S1.html

[6] MED: http://www.mpi.nl/world/tg/spoken-childes/spoken-childes.html

[7] Gesture Corpus: http://www.mpi.nl/world/research/research.html

[8] BNC: http://info.ox.ac.uk/bnc

[9] MPEG7: http://drogo.cselt.stet.it/mpeg/standards/mpeg-7/mpeg-7.htm

[10] H. Brugman, S. Kita 1995: *Impact of digital video technology on transcription: a case of spontaneous gesture transcription*, Ars Semeiotica Volume 18, Gunter Narr Verlag Tübingen

[11] EUDICO: http://www.mpi.nl/world/tg/lapp/eudico/eudico.html

[12] Talkbank: http://www.talkbank.org

[13] MATE: http://mate.mip.ou.dk