A general scheme for broad-coverage multimodal annotation

Philippe Blache
Laboratoire Parole et Langage
CNRS & Aix-Marseille Universités
blache@lpl-aix.fr

Abstract

We present in this paper a formal and computational scheme in the perspective of broad-coverage multimodal annotation. We propose in particular to introduce the notion of *annotation hypergraphs* in which primary and secondary data are represented by means of the same structure.

This paper addresses the question of resources and corpora for natural human-human interaction, in other words broad-coverage annotation of natural data. In this kind of study, most of domains have do be taken into consideration: prosody, pragmatics, syntax, gestures, etc. All these different domains interact in order to build an understandable message. We need then large multimodal annotated corpora of real data, precisely annotated for all domains. Building this kind of resource is a relatively new, but very active research domain, illustrated by the number of workshops (cf. (Martin, 2008)), international initiatives, such as MUMIN (Allwood, 2005), annotation tools such as NITE NXT (Carletta, 2003), Anvil (Kipp, 2001), etc.

1 A characterization of primary data

Different types of primary data constitute the basis of an annotation: speech signal, video input, word strings, images, etc. But other kinds of primary data also can be used, for example in the perspective of semantic annotations such as concepts, references, types, etc. Such data are considered to be atomic in the sense that they are not built on top of lower level data. When looking more closely at these kinds of data, several characteristics can be identified:

- **Location**: primary data is usually localized with respect to a timeline or a position: gestures can be localized into the video signal, phonemes into the speech one, words into the string or objects into a scene or a context. Two different kinds of localisation are used: *temporal* and *spatial*. In the first case, a data is situated by means of a time interval whereas spatial data are localised in terms of relative or absolute positions.
- **Realization**: primary data usually refer to *concrete* (or physical) objects: phonemes, gestures, referential elements into a scene, etc. However, other kinds of primary data can be *abstract* such as concepts, ideas, emotions, etc.
- **Medium**: The W3C recommendation EMMA (*Extensible Multi-Modal Annotations*) proposes to distinguish different medium: *acoustic, tactile* and *visual*. This classification is only relevant for data corresponding to concrete objects.
- **Production**: the study of information structure shows the necessity to take into account accessibility of the objects: some data are directly accessible from the signal or the discourse, they have an existence or have already been mentioned. In this case, they are said to be "*produced*". For example, gestures, sounds, physical objects fall in this category. On the other hand, other kinds of data are deduced from the context, typically the abstract ones. They are considered as "*accessible*".

In the remaining of the paper, we propose the following definition:

Primary data*: atomic objects that cannot be decomposed. They represent possible constituent on top of which higher level objects can be built. Primary data does not require any interpretation to be identified, they are of direct access.*

This primary data typology is given in figure (1). It shows a repartition between *concrete* vs. *abstract* objects. Concrete objects are usually those taken into account in corpus annotation. As a consequence, annotation usually focuses on speech and gestures, which narrows down the set of data to those with a temporal localization. However, other kinds of data cannot be situated in the

Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009, pages 174–177,
Suntec, Singapore, 6-7 August 2009.

	Phonemes	Words	Gestures	Discourse referents	Synsets	Physical objects
Produced	+	+	+	+/-	-	+
Accessible	-	-	-	+/-	+	-
Concrete	+	+	+	+/-	-	+
Abstract	-	-	-	+/-	-	+
Temporal	+	+	+	+/-	-	-
Spatial	-	-	+/-	+/-	-	+
Acoustic	+	+/-	-	-	-	-
Visual	-	-	+	+/-	-	+
Tactile	-	-	+/-	+/-	-	+

Figure 1: *Primary data description*

timeline (e.g. objects in the environment of the scene) nor spatially (e.g. abstract data).

We need to propose a more general approach of data indexing that has to distinguish on the one hand between temporal and spatial localization and on the other hand between data that can be located and data that cannot.

2 Graph representation: nodes and edges semantics

One of the most popular linguistic annotation representation is *annotation graphs* (Bird, 2001) in which nodes are positions whereas edges bear linguistic information. This representation is elaborated on the basis of a temporal anchoring, even though it is also possible to represent other kinds of anchoring. Several generic annotation format has been proposed on top of this representation, such as LAF and its extension GrAF (cf. (Ide, 2007)). In these approaches, edges to their turn can be interpreted as nodes in order to build higher level information. One can consider the result as an hypergraph, in which nodes can be subgraphs.

In order to explore farther this direction, we propose a more general interpretation for nodes that are not only positions in the input: nodes are complex objects that can be referred at different levels of the representation, they encode all annotations. In order to obtain an homogeneous representations, the two node types used in hypergraphs (*nodes* and *hypernodes*) share the same information structure which relies on the following points:
- **Index**: using an index renders possible to represent any kind of graphs, not only trees. They give to nodes the possibility of encoding any kind of information.
- **Domain**: prosody, semantics, syntax, gesture, pragmatics, etc. It is important to indicate as precisely as possible this information, eventually by means of sub-domains
- **Location**: annotations generally have a spatial or a temporal situation. This information is optional.
- **Features**: nodes have to bear specific linguistic indications, describing its properties.

Hypernodes bear, on top of this information, the specification of the subgraph represented by its constituents and their relations. We propose to add another kind of information in the hypernode structure:

- **Relations**: secondary data are built on top of primary one. They can be represented by means of a set of properties (constituency, linearity, coreference, etc.) implemented as edges plus the basic characteristics of a node. A secondary data is then graph with a label, these two elements composing an hypernode.

The distinction between node and hypernodes makes it possible to give a homogeneous representation of primary and secondary data.

3 An XML representation of annotation hypergraphs

We propose in this section an XML encoding of the scheme presented above.

3.1 Atomic nodes

The first example of the figure (2) illustrates the representation of a *phoneme*. The node is indexed, making its reference possible in higher level structures. Its label corresponds to the tag that would be indicated in the annotation. Other elements complete the description: the linguistic domain (specified by the attributes type and sub-type), the specification of the medium, the object localization (by means of anchors). In this example, a phoneme being part of the acoustic signal, the anchor is temporal and use an explicit timeline reference.

The same kind of representation can be given for transcription tokens (see node `n21` in figure (2)). The value of the node is the orthographic form. It is potentially aligned on the signal, and then represented with a temporal anchoring. Such

```
<node ID="n1" label="u">
   <domain type="phonetics" subtype="phoneme"
         medium="acoustic"/>
   <anchor type="temporal" start="285" end="312"/>
</node>

<node ID="n21" label="book">
   <domain type="transcription" subtype="token"/>
   <anchor type="temporal" start="242" end="422"/>
</node>

<node ID="n24" label="N">
   <domain type=" morphosyntax" subtype="word"/>
   <anchor type="temporal" start="242" end="422"/>
   <features ms="ncms---"/>
</node>

<node ID="n3" label="deictic">
   <domain type="gestures" subtype="hand"/>
   <anchor type="temporal" start="200" end="422"/>
   <features hand="right" deictic_type="space"
         object="ref_object"/>
</node>

<node ID="n4" label="discourse-referent">
   <domain type="semantics" subtype="discourse_universe"
         medium="visual"/>
   <anchoring type="spatial" x="242" y="422" z="312"/>
   <features isa="book" color="red" />
</node>
```

Figure 2: *XML encoding of atomic nodes*

anchoring makes it possible to align the orthographic transcription with the phonetic one. In the case of written texts, temporal bounds would be replaced by the positions in the texts, which could be interpreted as an implicit temporal anchoring.

The next example presented in node `n24` illustrates the representation of part-of-speech nodes. The domain in this case is *morphosyntax*, its subtype is "`word`". In this case too, the anchoring is temporal, with same bounds as the corresponding token. In this node, a feature element is added, bearing the morpho-syntactic description.

The atomic node described in node `n3` represents another physical object: a *deictic gesture*. Its domain is gesture and its subtype, as proposed for example in the MUMIN scheme (see (Allwood, 2005)) is the part of the body. The anchoring is also temporal and we can observe in this example a synchronization of the gesture with the token "`book`".

The last example (node `n4`) presents an atomic node describing a physical object present in the scene (a book on a shelf of a library). It belongs to the semantics domain as a discourse referent and is anchored spatially by its spatial coordinates. One can note that anchoring can be *absolute* (as in the examples presented here) or *relative* (situating the object with respect to other ones).

3.2 Relations

Relations are represented in the same way as nodes. They are of different types, such as constituency, linearity, syntactic dependency, semantic specification, etc. and correspond to a certain domain. The example `r1` in figure (3) illustrates a `specification` relation between a noun (node `n21`, described above) and its determiner (node `n20`). Non-oriented binary relations also occur, for example cooccurrency. Relations can be expressed in order to represent a set of objects. The next example (relation `r2`) presents the case of three constituents of an higher-level object (the complete description of which being given in the next section).

Finally, the alignment between objects is specified by two different values: *strict* when they have exactly the same temporal or spatial marks; *fuzzy* otherwize.

3.3 Hypernodes

Hypernodes encode subgraphs with the possibility of being themselves considered as nodes. Their structure completes the atomic node with a set of relations. Hypernodes encode different kinds of objects such as phrases, constructions, referential expressions, etc. The first example represents a *NP*. The node is indexed, bears a tag, a domain, an anchoring and features. The set of relations specifies two types of information. First, the *NP* node has three constituents: `n20` (for example a determiner), `n22` (for example an adjective) and `n24` (the noun described in the previous section). The alignment is said to be *strict* which means that the right border of the first element and the left border of the last one have to be the same. The resulting structure is an hypernode describing the different characteristics of the *NP* by means of features and relations.

The second example illustrates the case of a referential expression. Let's imagine the situation where a person points out at a book on a shelf, saying "*The book will fall down*". In terms of information structure, the use of a definite *NP* is possible because the referent is accessible from the physical context: the alignment of the *NP* (`n50`) and the deictic gesture (`n3`, see previous section) makes the coreference possible. This construction results in a discourse referent bringing together all the properties of the physical object (`n3`) and that of the object described in the discourse

```
<relation id="r1" label="specification">
   <domain type="syntax" subtype="oriented_rel"/>
   <edge from="n20" to="n24">
</relation>
```

```
<relation id="r2" label="constituency">
   <domain type="syntax" subtype="set_rel"/>
   <node_list>
      <node id="n20"/> <node id="n22"/> <node id="n24"/>
   </node_list>
   <alignment type="strict"/>
</relation>
```

Figure 3: *XML encoding of relations*

```
<node ID="n50" label="NP">
   <domain type="syntax" subtype="phrase"/>
   <anchor type="temporal" start="200" end="422"/>
   <features cat="NP" agr="ms" sem_type="ref"/>
   <relations>
      <relation id="r1" type="constituency">
         <domain type="syntax" subtype="set_rel"/>
         <node_list>
            <node id="n20"/> <node id="n22"/> <node id="n24"/>
         </node_list>
         <alignment type="strict"/>
      </relation>
      <relation id="r2" type="specification">
         <domain type="syntax" subtype="oriented_rel"/>
         <edge from="n20" to="n24">
      </relation>
   </relations>
</node>
```

```
<node ID="n51" label="ref_expression">
   <domain type="semantics" subtype="discourse_referent"/>
   <features referent="book'" color="red" />
   <relations>
      <relation id="r3" type="constituency">
         <domain type="semantics" type="set_rel"/>
         <node_list>
            <node id="n50"/> <node id="n3"/> <node id="n4"/>
         </node_list>
         <alignment type="fuzzy"/>
      </relation>
      <relation id="r4" type="pointing">
         <domain type="gesture" type="oriented_rel"/>
         <edge from="n3" to="n4">
         <alignment type="strict"/>
      </relation>
   </relations>
</node>
```

Figure 4: *XML encoding of hypernodes*

(n50). In this expression, the alignment between the objects is fuzzy, which is the normal situation when different modalities interact. The second relation describes the pointing action, implementing the coreference between the noun phrase and the physical object. This representation indicates the three nodes as constituents.

4 Conclusion

Understanding the mechanisms of natural interaction requires to explain how the different modalities interact. We need for this to acquire multimodal data and to annotate them as precisely as possible for all modalities. Such resources have to be large enough both for theoretical and computational reasons: we need to cover as broadly as possible the different phenomena and give the possibility to use machine learning techniques in order to produce a new generation of multimodal annotation tools. However, neither such resource, and a fortiori such tools, already exist. One reason, besides the cost of the annotation task itself which is still mainly manual for multimodal information, is the lack of a general and homogeneous annotation scheme capable of representing all kinds of information, whatever its origin.

We have presented in this paper the basis of such a scheme, proposing the notion of *annotation hypergraphs* in which primary as well as secondary data are represented by means of the same node structure. This homogeneous representation is made possible thanks to a generic description of primary data, identifying four types of basic information (index, domain, location, features). We have shown that this scheme can be directly represented in XML, resulting in a generic multimodal coding scheme.

References

Allwood J., L. Cerrato, L. Dybkjaer, & al. (2005) "The MUMIN Multimodal Coding Scheme", *NorFA yearbook*

Bird S., M. Liberman (2001) "A formal framework for linguistic annotation" *Speech Communication*, Elsevier

Carletta, J., J. Kilgour, and T. O'Donnell (2003) "The NITE Object Model Library for Handling Structured Linguistic Annotation on Multimodal Data Sets" in procs of the EACL Workshop on Language Technology and the Semantic Web

Ide N. & K. Suderman (2007) "GrAF: A Graph-based Format for Linguistic Annotations", in proceedings of the *Linguistic Annotation Workshop at the ACL'07 (LAW-07)*

Kipp M. (2001) "Anvil-a generic annotation tool for multimodal dialogue" in procs of 7th European Conference on Speech Communication and Technology

Martin, J.-C., Paggio, P., Kipp, M., Heylen, D. (2008) Proceedings *of the Workshop on Multimodal Corpora : From Models of Natural Interaction to Systems and Applications* (LREC'2008)