A Formal Scheme for Multimodal Grammars

Philippe Blache & Laurent Prévot
LPL-CNRS, Université de Provence
blache@lpl-aix.fr

Abstract

We present in this paper a formal approach for the representation of multimodal information. This approach, thanks to the to use of *typed feature structures* and *hypergraphs*, generalizes existing ones (typically annotation graphs) in several ways. It first proposes an homogenous representation of different types of information (nodes and relations) coming from different domains (speech, gestures). Second, it makes it possible to specify constraints representing the interaction between the different modalities, in the perspective of developing *multimodal grammars*.

1 Introduction

Multimodality became in the last decade an important challenge for natural language processing. Among the problems we are faced with in this domain, one important is the understanding of how does the different modalities interact in order to produce meaning. Addressing this question requires to collect data (building corpora), to describe them (enriching corpora with annotations) and to organize systematically this information into a homogeneous framework in order to produce, ideally, multimodal grammars.

Many international projects address this question from different perspectives: data representation and coding schemes (cf. ISLE (Dybkjaer, 2001), MUMIN (Allwood, 2005), etc.), corpus annotation (cf. LUNA (Rodriguez, 2007) or DIME (Pineda, 2000), etc.), annotation and editing tools (such as NITE NXT (Carletta, 2003), Anvil (Kipp, 2001), Elan (Wittenburg, 2006), Praat (Boersma, 2009), etc.).

We propose in this paper a generic approach addressing both formal representation and concrete annotation of multimodal data, that relies on *typed-feature structure* (TFS), used as a description language on graphs. This approach is generic in the sense that it answers to different needs: it provides at the same time a formalism directly usable for corpus annotation and a description language making it possible to specify constraints that constitute the core of a *multimodal grammar*.

In the first section, we motivate the use of TFS and present how to concretely implement them for multimodal annotation. We address in the second section one of the most problematic question for multimodal studies: how to represent and implement the relations between the different domains and modalities (a simple answer in terms of time alignment being not powerful enough). In the last section, we describe how to make use of this representation in order to specify multimodal grammars.

2 Typed-feature structures modeling

Information representation is organized in two dimensions: type hierarchies and constituency relations (typically, a prosodic unit is a set of syllables, which in turn are sets of phonemes). The former corresponds to an *is-a* relation, the latter to a *part-of* one. For example *intonational phrase* is a subtype of *prosodic phrase*, and *phonemes* are constituents of *syllables*.

Such an organization is directly represented by means of typed feature structures. They can be considered as a formal annotation schema, used as

Coling 2008: Poster Volume, pages 63–71,
Beijing, August 2010

a preliminary step before the definition of the concrete coding scheme[1]. This step is necessary when bringing together information (and experts) from different fields: it constitutes a common representation framework, homogenizing information representation. Moreover, it allows to clearly distinguish between knowledge representation and annotation. The coding scheme, at the annotation level (labels, features, values), is deduced from this formal level.

The remaining of the section illustrates how to represent objects from different domains by means of TFS. The Figure 1 presents the type hierarchy and the constituency structure of objects taken here as example.

2.1 Phonetics

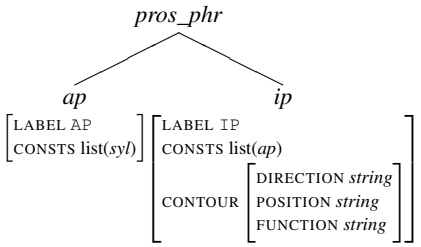
The phoneme is used as primary data: this object is at the lowest level of the constituent hierarchy (most of the objects are set of phonemes). The following feature structure proposes a precise encoding of the main properties describing a phoneme, including articulatory gestures.

$$
\textit{phon}\left[\begin{array}{l}
\text{SAMPA\_LABEL } \textit{sampa\_unit} \\
\text{CAT}\left\{\textit{vowel, consonant}\right\} \\
\text{TYPE}\left\{\textit{occlusive, fricative, nasal, etc.}\right\} \\
\text{ARTICULATION}\left[\begin{array}{l}
\text{LIP}\left[\begin{array}{l}\text{PROTUSION } \textit{string} \\ \text{APERTURE } \textit{aperture}\end{array}\right] \\
\text{TONGUE}\left[\begin{array}{l}\text{TIP}\left[\begin{array}{l}\text{LOCATION } \textit{string} \\ \text{DEGREE } \textit{string}\end{array}\right] \\ \text{BODY}\left[\begin{array}{l}\text{LOCATION } \textit{string} \\ \text{DEGREE } \textit{string}\end{array}\right]\end{array}\right] \\
\text{VELUM } \textit{aperture} \\
\text{GLOTTIS } \textit{aperture}
\end{array}\right] \\
\text{ROLE}\left[\begin{array}{l}\text{EPENTHETIC } \textit{boolean} \\ \text{LIAISON } \textit{boolean}\end{array}\right]
\end{array}\right]
$$

Phonemes being at the lowest level, they do not have any constituents. They are not organized into precise subtypes. The feature structure represent then the total information associated with this type.

2.2 Prosody

As seen above, prosodic phrases are of two different subtypes: *ap* (accentual phrases) and *ip* (intonational phrases). The prosodic type hierarchy is represented as follows:



Accentual phrases have two appropriate features: the label which is simply the name of the corresponding type, and the list of constituents, in this case a list of syllables. The objects of type *ip* contain the list of its constituents (a set of *ap*s) as well as the description of its contour. A contour is a prosodic event, situated at the end of the *ip* and is usually associated to an *ap*.

The prosodic phrases are defined as set of syllables. They are described by several appropriate features: the syllable structure, its position in the word, its possibility to be accented or prominent:

$$
\textit{syl}\left[\begin{array}{l}
\text{STRUCT } \textit{syl\_struct} \\
\text{POSITION}\left[\begin{array}{l}\text{RANK}\left\{\textit{integer}\right\} \\ \text{SYL\_NUMBER}\left\{\textit{integer}\right\}\end{array}\right] \\
\text{ACCENTUABLE } \textit{boolean} \\
\text{PROMINENCE } \textit{boolean} \\
\text{CONSTITUENTS list}(\textit{const\_syl})
\end{array}\right]
$$

Syllable constituents (objects of type *const_syl*) are described by two different features: the set of phonemes (syllable constituents), and the type of the constituent (onset, nucleus and coda). Note that each syllable constituent can contain a set of phonemes.

$$
\textit{const\_syl}\left[\begin{array}{l}
\text{PHON list}(\textit{phon}) \\
\text{CONST\_TYPE}\left\{\textit{onset, nucleus, coda}\right\}
\end{array}\right]
$$

2.3 Disfluencies

We can distinguish two kinds of disfluencies: *non lexicalized* (without any lexical material, such as lengthening, silent pauses or filled pauses) and *lexicalized* (non-voluntary break in the phrasal flow, generating a word or a phrase fragment). Lexicalized disfluencies have a particular organization with three subparts (or constituents):

- *Reparandum*: the word or phrase fragment, in which the break occurs
- *Break*: a point or an interval that can eventually be filled by a fragment repetition, parenthetical elements, etc.

[1]This approach has been first defined and experimented in the XXXX project, not cited for anonymity reasons.

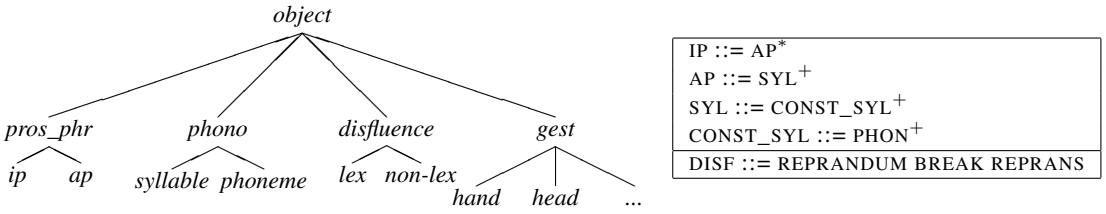
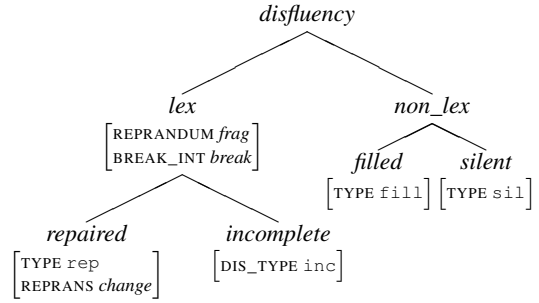
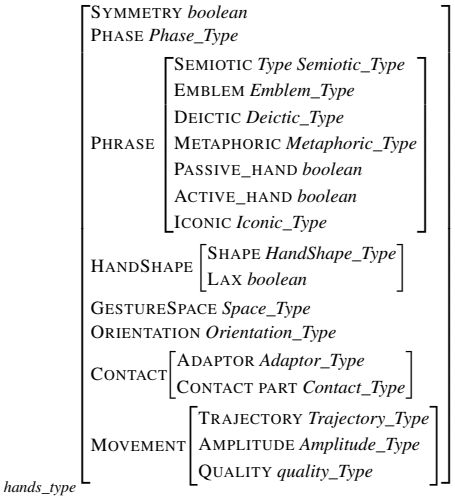


Figure 1: *Type and constituent hierarchies*

- *Reparans*: all that follow the break and recovers the reparandum (in modifying or completing it) or simply left it uncompleted.

The general disfluency type hierarchy, with the appropriate features at each level is given in the following figure:





2.4 Gestures

Besides verbal communication, gestures constitute the main aspect of multimodality. In multimodal annotation, this is probably the most difficult and time-consuming task. Moreover, only few works really focus on a precise description of all the different domains of verbal and non verbal modalities. The TFS-based approach proposed here answers to the first need in such a perspective: a common representation framework.

We give in this section a brief illustration of the representation of one gesture (hands). It relies on adaptation of different proposals, especially (Kipp03) or MUMIN (Allwood, 2005), both integrating McNeill's gesture description (McNeill05).

The following structure encodes the description of gesture phases, phrases (representing different semiotic types), the hand shape as well as its orientation, the gesture space, and the possible contact with bodies or objects. A last feature also describes the movement itself: trajectory, quality (fast, normal or slow) and amplitude (small, medium and large).

2.5 Application

We have experimented this modeling in the complete annotation of a multimodal corpus (see (Blache, 2010)). In this project, a complete TFS model has been first designed, covering all the different domains (prosody, syntax, gestures, discourse, etc.). From this model, the annotations have been created, leading to a 3-hours corpus of narrative dialogs, fully transcribed. The corpus is fully annotated for some domains (phonetics, prosody and syntax) and partly for others (gestures, discourse, disfluencies, specific phenomena). The result is one of the first large annotated multimodal corpus.

3 Graphs for Multimodal Annotation

Graphs are frequently used in the representation of complex information, which is the case with multimodality. As for linguistic annotation, one of the most popular representations is *Annotation Graphs* (Bird, 2001). They have been proposed in particular in the perspective of anchoring different kinds of information in the same reference,

making it possible to align them[2]. In AGs, nodes represent positions in the signal while edges bear linguistic information. Two edges connecting the same nodes are aligned: they specify different information on the same part of the input. Implicitly, this means that these edges bear different features of the same object.

Such a representation constitutes the basis of different approaches aiming at elaborating generic annotation formats, for example LAF (and its extension GrAF (Ide, 2007)). In this proposal, edge labels can be considered as nodes in order to build higher level information. One can consider the result as an *hypergraph*, in which nodes can be subgraphs.

We propose in this section a more generalized representation in which nodes are not positions in the signal, but represent directly objects (or set of objects). All nodes have here the same structure, being them nodes or hypernodes. The main interest of this proposal, on top of having an homogeneous representation, is the possibility to anchor information in different references (temporal, spatial or semantic).

3.1 Nodes

As seen above, multimodal annotation requires the representation of different kinds of information (speech signal, video input, word strings, images, etc.). The *objects*[3] that will be used in the description (or the annotation) of the input are of different nature: temporal or spatial, concrete or abstract, visual or acoustic, etc. A generic description requires first a unique way of locating (or indexing) all objects, whatever their domain. In this perspective, an index (in the HPSG sense) can be specified, relying on different information:

- LOCATION: objects can in most of the cases be localized in reference to a temporal or a spatial situation. For example, phonemes have a temporal reference into the speech signal, physical objects have spatial localization that can be absolute (spatial coordinates), or relative (with respect to other objects).
- REALIZATION: data can either refer to *concrete* or physical objects (phonemes, gestures, referential elements, etc.) as well as *abstract* ones (concepts, emotions, etc.).
- MEDIUM: specification of the different modalities: *acoustic, tactile* and *visual*.[4]
- ACCESSIBILITY: some data are directly accessible from the signal or the discourse, they have a physical existence or have already been mentioned. In this case, they are said to be "*given*" (e.g. gestures, sounds, physical objects). Some other kinds of data are deduced from the context, typically the abstract ones. They are considered as "*accessible*".

A generic structure node can be given, gathering the index and the some other object properties.

$$
\begin{bmatrix}
\text{ID} \\
\text{DOMAIN} \left\{ \textit{prosody, syntax, pragmatics, ...} \right\} \\
\text{INDEX} \begin{bmatrix}
\text{LOCATION} \left\{ \begin{array}{l} \text{TEMPORAL} \begin{bmatrix} \text{START } \textit{value} \\ \text{END } \textit{value} \end{bmatrix} \\ \text{SPATIAL } \textit{coord} \end{array} \right\} \\
\text{REALIZATION} \left\{ \textit{concrete, abstract} \right\} \\
\text{MEDIUM} \left\{ \textit{acoustic, tactile, visual} \right\} \\
\text{ACCESSIBILITY} \left\{ \textit{given, accessible} \right\}
\end{bmatrix} \\
\text{FEATURES } \textit{object\_type}
\end{bmatrix}_{node}
$$

This structure relies on the different information. Besides INDEX, some other features complete the description:

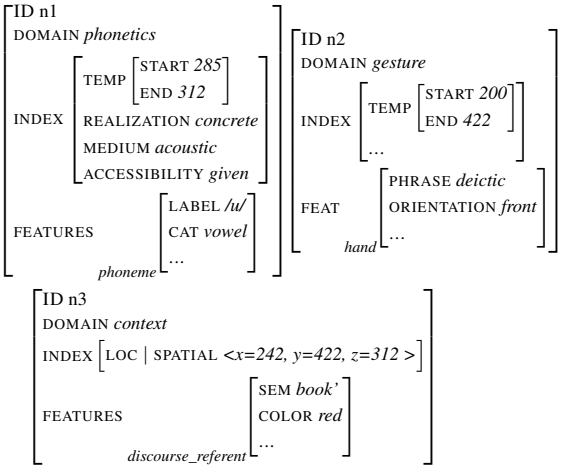
- ID: using an absolute ID is useful in the perspective of graph representation, in which nodes can encode any kind of information (atomic or complex, including subgraphs).
- DOMAIN: specification of the domain to which the information belongs. This feature is useful in the specification of generic interaction constraints between domains.
- FEATURES: nodes have to bear specific linguistic indications, describing object properties. This field encodes the type of information presented in the first section.

[2] Another important interest of AGs is that they can constitute the basis for an exchange format, when thinking on annotation tools interoperability (a proposal is currently elaborated under auspices of the MITRE program, see `http://www.mitre.org/`).

[3] We call *object* any annotation that participates to the description: phoneme, words, gestures, but also phrases, emotions, etc.

[4] See the W3C EMMA recommendation (*Extensible Multi-Modal Annotations*, `http://www.w3.org/2002/mmi/`.

The following examples illustrate the representation of atomic nodes from different domains: a phoneme (node *n1*) and a gesture (node *n2*), that are temporally anchored, and a physical object (node *n3*) which is spatially situated. This last object can be used as a referent, for example by a deictic gesture.

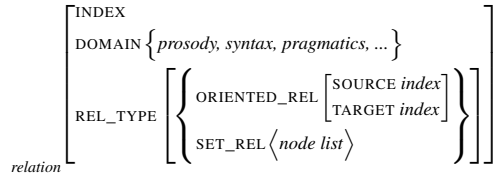


3.2 Relations

Linguistic information is usually defined in terms of relations between (sets of) objects, which can be atomic or complex. For example, a phrase is defined by syntactic relations (government, agreement, linearity, etc.) between its constituents. In some cases, these relations can concern objects from the same domain (e.g. syntax in the previous example). In other cases, different domains can be involved. For example, a long break (greater than 200ms) usually precedes a left corner of a new phrase.

The nature of the relation can also be different according to the kind of information to be encoded. Many relations are binary and oriented (precedence, dependency, etc.). Some others only consists in gathering different objects. A construction (in the sense of *Construction Grammars*, see (Fillmore96)) is precisely that: a set of object or properties that, put together, form a specific phenomenon. It is then useful in our representation to distinguish between *oriented relations* and *set relations*. Oriented relations (for example precedence) connect a source and a target, that can be eventually formed with set of objects. Set relations are used to gather a set of objects, without orientation or order (e.g. the constituency relation).

On top of this distinction, it is also necessary to give an index to the relations, in order to make their reference possible by other objects. As for nodes, an index is used, even though its form is simple and does not need a complex anchor. Finally, for the same reasons as for nodes, the specification of the domain is necessary. The following feature structure gives a first view of this organization:



Besides these information, a relation description has to be completed with other information:

- TYPE: different types of relations can be implemented in such representation, such as dependency, precedence, constituency, anaphore, etc.
- SCOPE: a relation can be specific to a construction or at the opposite valid whatever the context. For example, the precedence relation $[V \prec Clit_{[nom]}]$ is only valid in the context of interrogative constructions whereas the relation exluding the realization of a backchannel[5] after a connective is valid whatever the context. We distinguish then between *local* and *global* scopes.
- POLARITY: a relation can be negated, implementing the impossibility of a relation in a given context.
- CONSTRUCTION: in the case of a local relation, it is necessary to specify the construction to which it belongs.
- STRENGTH: some relation are mandatory, some other optional. As for constraints, we distinguish then between *hard* and *soft* relations, depending on their status.

Finally, a last property has to be precisely defined: the synchronization between two objects

[5] A backchannel is a reaction, verbal or gestual, of the adressee during a conversation.

coming from different domains (for example gestures and words). In some cases, both objects have to be strictly aligned, with same boundaries. For example, a syllable has to be strictly aligned with its set of phonemes: the left syllable boundary (resp. the right) has to be the same as that of the first syllable phoneme (resp. the last). In other cases, the synchronization must not be strict. For example, a deictic gesture is not necessarily strictly aligned with a referential pronoun. In this case, boundaries of both objects only have to be roughly in the same part of the signal.

We propose the definition of alignment operators adapted from (Allen, 1985) as follows:

$=$	*same*	boundaries have to be equal
$<_\Delta$	*before*	$b_1 <_\Delta b_2$ means b_1 value is lower than b_2, with $b_2 - b_1 < \Delta$
$>_\Delta$	*after*	$b_1 >_\Delta b_2$ means that the boundary b_1 follows b_2, with $b_1 - b_2 < \Delta$
$\approx_\Delta$	*almost*	boundaries are neighbors, without order relation, with $\mid b_1 - b_2 \mid \leq \Delta$

This set of operators allow to specify *alignment equations* between different objects. The advantage of this mechanism is that an equation system can describe complex cases of synchronization. For example, a construction can involve several objects from different domains. Some of these objects can be strictly aligned, some others not.

The final TFS representation is as follows:

$$\text{relation}\begin{bmatrix}\text{INDEX}\\ \text{DOMAIN}\{\textit{prosody, syntax, pragmatics, ...}\}\\ \text{REL\_TYPE}\begin{bmatrix}\{\text{ORIENTED\_REL}\begin{bmatrix}\text{SOURCE } \textit{index}\\ \text{TARGET } \textit{index}\end{bmatrix}\\ \{\text{SET\_REL}\langle \textit{node list}\rangle\}\end{bmatrix}\\ \text{TYPE}\{\textit{dependency, precedence, etc.}\}\\ \text{SCOPE}\{\textit{global, local}\}\\ \text{POLARITY}\{\textit{plus, minus}\}\\ \text{CONSTRUCTION } \textit{contruction\_type}\\ \text{STRENGTH}\{\textit{hard, soft}\}\\ \text{ALIGNMENT}\langle \textit{alignment\_equations}\rangle\end{bmatrix}$$

The following feature structure shows an example of a global relation indicating that a verbal nucleus usually comes with a minor raising of the intonation (only main features are indicated here). This information is represented by an implication relation, which is oriented from the syntactic category to the prosodic phenomenon. Alignment equations stipulate a strict synchronization between object.

$$\text{relation}\begin{bmatrix}\text{INDEX}\\ \text{REL\_TYPE} \mid \text{ORIENTED\_REL}\begin{bmatrix}\text{SOURCE } VN_1\\ \text{TARGET } mr_2\end{bmatrix}\\ \text{TYPE}\{\textit{implication}\}\\ \text{STRENGTH}\{\textit{soft}\}\\ \text{ALIGNMENT}\langle lb_1{=}lb_2;\ rb_1{=}rb_2\rangle\end{bmatrix}$$

4 Representation with Hypergraphs

Nodes and relations can be combined and form higher level nodes, representing constructions which are a set of objects (the constituents) plus a set of relations between them. Such nodes are in fact *hypernodes* and bear two kinds of information: the properties characterizing the object plus a set of relations between the constituents (representing a subgraph). In the syntactic domain, for example, they represent phrases, as follows:

$$\begin{bmatrix}\text{DOMAIN } \textit{syntax}\\ \text{INDEX} \mid \text{LOCATION} \mid \text{TEMPORAL}\begin{bmatrix}\text{START } 122\\ \text{END } 584\end{bmatrix}\\ \text{FEATURES}\begin{bmatrix}\text{CAT } VP\end{bmatrix}\\ \text{RELATIONS}\left\{\begin{bmatrix}\text{INDEX } r_1\\ \text{REL\_TYPE} \mid \text{SET\_REL}\langle V, NP, Adv\rangle\\ \text{TYPE } \textit{constituency}\\ \text{STRENGTH } \textit{hard}\end{bmatrix};\right.\\ \left.\begin{bmatrix}\text{INDEX } r_2\\ \text{REL\_TYPE} \mid \text{ORIENTED\_REL}\begin{bmatrix}\text{SOURCE } NP\\ \text{TARGET } V\end{bmatrix}\\ \text{TYPE } \textit{dependency}\\ \text{STRENGTH } \textit{hard}\end{bmatrix}\right\}\end{bmatrix}$$

In the same way, the interaction between different objects from different domains can involve several relations. For example, a deictic construction can be made of the conjunction of an anaphoric pronoun, a deictic gesture and a physical object (for example a book on a shelf). Such a construction can be described by the following structure:

$$\begin{bmatrix}\text{INDEX} \mid \text{LOCATION} \mid \text{TEMPORAL}\begin{bmatrix}\text{START } 841\\ \text{END } 1520\end{bmatrix}\\ \text{FEATURES}\begin{bmatrix}\text{SEM } book'\end{bmatrix}\\ \text{RELATIONS}\left\{\begin{bmatrix}\text{INDEX } r_3\\ \text{SET\_REL}\langle Pro_1, Dx\_gest_2, Ph\_object_3\rangle\\ \text{TYPE } \textit{constituency}\\ \text{ALIGNMENT}\langle lb_1 \approx_\Delta lb_2;\ rb_1 \approx_\Delta rb_2\rangle\end{bmatrix};\right.\\ \left.\begin{bmatrix}\text{INDEX } r_4\\ \text{ORIENTED\_REL}\begin{bmatrix}\text{SOURCE } Pro_1\\ \text{TARGET } Ph\_object_3\end{bmatrix}\\ \text{TYPE } \textit{reference}\end{bmatrix}\right\}\end{bmatrix}$$

This construction indicates some properties (limited here to the semantic value) and two re-

lations between the different objects: one constituency, indicating the different objects involved in the construction and their (fuzzy) alignment and a reference relation between the pronoun and a physical object (here, a book).

This structure represents an hypergraph: it is a graph connecting different nodes, each of them being to its turn described by another graph, as shown above. The main interest of such a representation is its flexibility: all kinds of information can be described, at any level. Graphs being less constrained than trees, and edges (or relations) being typed, we can gather different levels, different domains and different granularities. For example, an agreement relation can be specified thanks to the deictic construction, besides the constituency one, making it possible to instanciate the agreement value of the pronoun.

Note that hypergraphs are also investigated in other knowledge representation, their properties are well known (Hayes, 2004) and the implementation of specific hypergraphs as the one presented here could be done in RDF graphs for example as suggested in (Cassidy, 2010).

5 Constraints for Multimodal Grammars

In the same way as typed feature structures can implement constraints and constitute a description language on linguistic structures (cf. HPSG,), the same approach can be generalized to multimodal information. SOme recent works have been done in this direction (see (Alahverdzhieva, 2010; **?**)). The representation we propose can implement generic information about multimodal constructions. We illustrate in the following this aspect with two phenomena: *backchannels* and *dislocation*.

Several studies on conversational data (see for example (Bertrand09)) have described backchannels (that can be vocal or gestual) and their context. They have in particular underline some regularities on the left context:

- backchannels usually follow: major intonative phrases (IP), flat contours, end of conversational turn (i.e. saturated from a semantic, syntactic and pragmatic point of view)
- backchannels never appear after connectives

These constraints can be implemented by means of a feature structure (representing an hypernode) with a set of precedence relations. The different objects involved in the description of the phenomenon (IP, flat contour, conversational turn, connective) are indicated with an indexed ID, referring to their complete feature structure, not presented here.

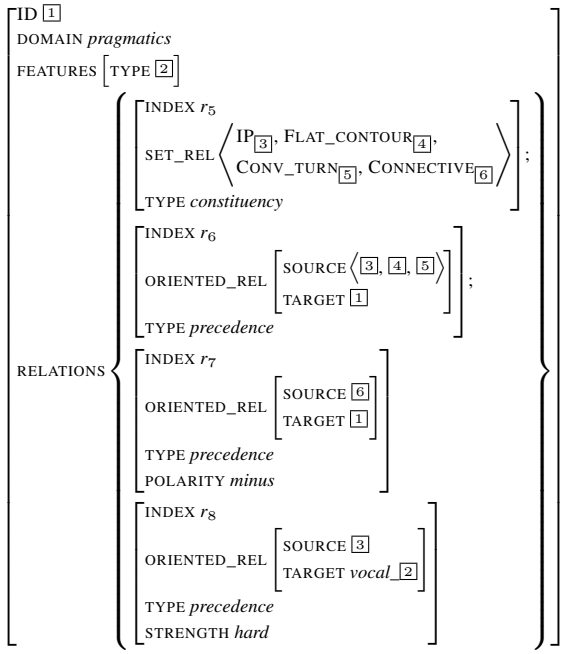


Figure 2: *Backchannel Constraint*

This structure (cf. Figure 2) represents a constraint that backchannels have to satisfy. The first relation specifies the constituents and their indexes, with which the different precedence constraints are represented. The relation *r6* indicates all kinds of object that should precede a backchannel. This constraint subsumes the most specific relation *r8* stipulating that a vocal backchannel is always preceded with an *IP* (this is a *hard* constraint). The relation *r7* excludes the possibility for a backchannel to be preceded with a connective.

The second example (cf. Figure 3) proposes a constraint system describing dislocated structures. We propose in this description to distinguish two syntactic constituents that form the two parts of the dislocation: the dislocated phrase (called *S1*) and the sentence from which the phrase has been

extracted (called *S2*). Usually (even if not always), *S2* contains a clitic referring to *S1*. We note in the following this clitic with the notation *S2//Clit*. For readability reasons, we only present in this structure the relations.

This structure describes the case of a left dislocation (with *S1* preceding *S2*, the constraint being hard). In such cases, *S1* is usually realized with a minor raising contour. The constraint *r13* implements the anaphoric relation between the clitic and the dislocated element. Finally, the relation *r14* indicates an agreement relation between the clitic and *S1* and in particular the fact that the case has to be the same for both objects.

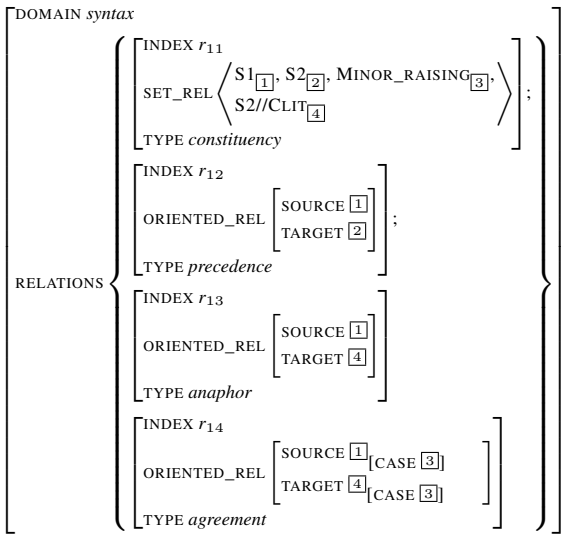


Figure 3: *Dislocation Constraint*

6 Conclusion

Linguistic annotation in general, and multimodality in particular, requires high level annotation schemes making it possible to represent in an homogeneous way information coming from the different domains and modalities involved in human communication.

The approach presented in this paper generalizes previous methods (in particular annotation graphs) thanks to two proposals: first in providing a way to index objects without strict order relation between nodes and second in specifying a precise and homogeneous representation of the objects and their relations. This approach has been developed into a formal scheme, *typed feature structures*, in which all the different domains can be represented, and making it possible to implement directly hypergraphs. TFS and hypergraphs are particularly well adapted for the specification of interaction constraints, describing interaction relations between modalities. Such constraints constitute the core of the definition of future multimodal grammars.

From a practical point of view, the proposal described in this paper is currently under experimentation within the OTIM project (see (Blache, 2010)). An XML scheme has been automatically generated starting from TFS formal scheme. The existing multimodal annotations, created with ad hoc annotation schemes, are to their turn automatically translated following this format. We obtain then, for the first time, a large annotated multimodal corpus, using an XML schema based on a formal specification.

References

Alahverdzhieva, K. and A. Lascarides (2010) "Analysing Language and Co-verbal Gesture and Constraint-based Grammars", in Proceedings of the *17th International Conference on Head-Driven Phase Structure Grammar*.

Allen F. and P. J. Hayes (1985) "A common-sense theory of time", in *9th International Joint Conference on Artificial Intelligence*.

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

Bertrand R., M. Ader, P. Blache, G. Ferré, R. Espesser, S. Rauzy (2009) "Représentation, édition et exploitation de données multimodales : le cas des backchannels du corpus CID", in *Cahiers de linguistique française*, 33:2.

Blache P., R. Bertrand, and G. Ferré (2009) "Creating and Exploiting Multimodal Annotated Corpora: The ToMA Project". in Kipp, Martin, Paggio and Heylen (eds.) *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*, LNAI 5509, Springer.

Blache P. et al. (2010) "Multimodal Annotation of Conversational Data", in proceedings of *LAW-IV - The Linguistic Annotation Workshop*

Bird S., Day D., Garofolo J., Henderson J., Laprun C. & Liberman M. (2000) "ATLAS : A Flexible and Extensible Architecture for Linguistic Annotation", in procs of *LREC00*

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

Boersma P. & D. Weenink (2009) *Praat*: doing phonetics by computer, http://www.praat.org/

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*

Carpenter B. (1992) *The Logic of Typed Feature Structures*. Cambridge University Press.

Cassidy S. (2010) *An RDF Realisation of LAF in the DADA Annotation Server*. Proceedings of ISA-5, Hong Kong, January 2010.

Dipper S., M. Goetze and S. Skopeteas (eds.) (2007) *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics and Information Structure*, Working Papers of the SFB 632, 7:07

Dybkjaer L., S. Berman, M. Kipp, M. Wegener Olsen, V. Pirrelli, N .Reithinger, C. Soria (2001) "Survey of Existing Tools, Standards and User Needs for Annotation of Natural Interaction and Multimodal Data", *ISLE Natural Interactivity and Multimodality Working Group Deliverable D11.1*

Fillmore C. & P. Kay (1996) *Construction Grammar*, Manuscript, University of California at Berkeley Department of linguistics.

Gruenstein A., J. Niekrasz, and M. Purver. (2008) "Meeting structure annotation: Annotations collected with a general purpose toolkit". In L. Dybkjaer and W. Minker, editors, *Recent Trends in Discourse and Dialogue*, Springer-Verlag.

Hayes J. and Gutierrez C. (2004) Bipartite graphs as intermediate model for RDF. Proceedings of ISWC 2004, 3rd International Semantic Web Conference (ISWC2004), Japan.

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

Ide N. and Suderman K. (2009) Bridging the Gaps: Interoperability for GrAF, GATE, and UIMA. Proceedings of the Third Linguistic Annotation Workshop, held in conjunction with ACL 2009, Singapore.

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

Kipp, M. (2003) *Gesture Generation by Immitation: From Human Behavior to Computer Character Animation*, PhD Thesis, Saarland University.

Lascarides, A. and M. Stone (2009) "A Formal Semantic Analysis of Gesture", in *Journal of Semantics*, 26(4).

McNeill, D. (2005) *Gesture and Thought*, The University of Chicago Press.

Pineda, L., and G. Garza (2000) "A Model for Multimodal Reference Resolution", in *Computational Linguistics*, Vol. 26 no. 2

Rodriguez K., Stefan, K. J., Dipper, S., Goetze, M., Poesio, M., Riccardi, G., Raymond, C., Wisniewska, J. (2007) "Standoff Coordination for Multi-Tool Annotation in a Dialogue Corpus", in procs of the *Linguistic Annotation Workshop at the ACL'07 (LAW-07)*

Wegener Knudsen M.and al. (2002) *Survey of Multimodal Coding Schemes and Best Practice*, ISLE

Wittenburg, P.; Brugman, H.; Russel, A.; Klassmann, A. and Sloetjes, H. (2006) "ELAN: a Professional Framework for Multimodality Research". In proceedings of LREC 2006