# A Structural Model of Semiotic Alignment: The Classification of Multimodal Ensembles as a Novel Machine Learning Task

Alexander Mehler
Department for Computing in the Humanities
Goethe-Universität Frankfurt am Main, Germany
Email: Mehler@em.uni-frankfurt.de

Andy Lücking
Bielefeld University, Germany
CRC 673, B1, “Alignment in Communication”
Email: andy.luecking@uni-bielefeld.de

*Abstract*—**In addition to the well-known linguistic alignment processes in dyadic communication – e.g., phonetic, syntactic, semantic alignment – we provide evidence for a genuine multimodal alignment process, namely *semiotic alignment*. Communicative elements from different modalities “routinize into” cross-modal “super-signs”, which we call *multimodal ensembles*. Computational models of human communication are in need of expressive models of multimodal ensembles. In this paper, we exemplify semiotic alignment by means of empirical examples of the building of multimodal ensembles. We then propose a graph model of multimodal dialogue that is expressive enough to capture multimodal ensembles. In line with this model, we define a novel task in machine learning with the aim of training classifiers that can detect semiotic alignment in dialogue. This model is in support of approaches which need to gain insights into realistic human-machine communication.**

## I. Introduction: Alignment in Communication

Human communication is *coordinated* ([1], [2]) as well as *aligned* ([3], [4]). While the former pertains to “higher-level” coordination, the latter relates to coordination of linguistic representations at various levels, e.g., phonetic or syntactic. Alignment, unlike coordination, is an automatic, resource-free process, making successful dialogue a rather mechanistic process. According to the *Interactive Alignment Model* (IAM, [4]), alignment may “percolate up” from lower to higher levels, since linguistic representations are interconnected. In this way, “global” alignment, i.e, alignment of situation models, can be a result of “local” alignment ([4, p. 173]). The IAM predicts that natural language dialogue is highly repetitive. Since the linguistic representations of one dialogue partner are primed by the utterances of the other dialogue partner and so on, expressions will get “copied” within a dyad, both *intra*- as well as *inter*-personal. Alignment of word use leads to the building of a *dialogue lexicon*, that is, a *routinization* of an unambiguous usage of certain expressions *relative to* the life span of a certain dialogue ([4, p. 175]). However, when Pickering & Garrod speak of an “expression”, they refer to verbal utterances; their model is couched in the psychology of speech production and draws on unimodal verbal evidence. Human dialogue, though, is multimodal: it unfolds in several modalities and employs various communicative means, such as speech, gesture, gaze, proximity, and maybe even odour.

Multimodality opens a new alignment domain, namely *semiotic alignment*, which is the focus of this article. Co-repetition of elements on different modalities – for example, a verbal expression and a gesture – leads to a coupling of the respective elements. We call the resulting cross-modal relationships *Multimodal Ensembles* (MMEs). In the formation of an MME, elements from different semiotic channels become aligned in such a way that they can be seen as a cross-modal “super-sign”. It is reasonable to assume that an MME enters into the dialogue lexicon. In other words, the building of an MME is a *routinization process*. This is a case of *semiotic alignment*, which, as such, goes beyond the current range of the IAM, but nevertheless has to be accounted for in computational models of natural language dialogue.

## II. Semiotic Alignment: The Formation of Multimodal Ensembles

Let us illustrate the formation of MMEs by means of empirically observed data on the interplay of speech and gesture in giving directions.[1]

The direction giver, henceforth called DG, describes the fire escape stairs on both wings of a town hall. DG refers to them bimodally:

(1) hat an den beiden Flügeln blaue [Notfalltreppen]
*has blue [emergency stairs] on both wings*
(A *placing*-gesture is affiliated to bracketed speech)

The noun “Notfalltreppen” (emergency stairs) and the *placing*-gesture relate *in concert* to the discourse referent “fire escape stairs”.

A few seconds later, DG talks about an object that has the same colour as the fire escape stairs:

(2) gleiche Farbe wie [diese Notfalltreppen]
*same colour as [these emergency stairs]*
(A *placing*-gesture is affiliated to bracketed speech)

[1] The data stem from an experiment involving giving directions in a virtual reality environment. For details see [5].

DG re-uses the pair of verbal expression and gesture he previously employed to designate the fire escape stairs. That is, there is not only an uptake of the noun but also of the accompanying *placing*. The discourse referent “fire escape stairs” is quite reliably realized bimodally. Within the given dialogue, an association between the noun “Notfalltreppe” and a gesture is built by repetition. We refer to routines of this kind as *Multimodal Ensembles* (MMEs).

A striking feature of MMEs is that their “ensemblehood” supervenes [**?**] the actual realization of their parts. Consider another example. DG introduces the entrance of a park, which is an arch-shaped gate:

(3) bis irgendwann mal dann der Haupteingang kommt, das ist [ein großer, grauer Bogen]
*until you eventually reach the main entrance, that is [a big, grey arch]*
(An arch-shaped gesture is affiliated to bracketed speech)

In this example, the noun phrase “ein großer, grauer Bogen” (*a big, grey arch*) refers to the gate entrance of the park. The arch-shape of the gate is depicted by a drawing gesture. The noun phrase plus the gesture jointly refer to a discourse referent, namely the arch. Later on in that dialogue, however, DG once again makes reference to the arch entrance, in the following way:

(4) wo [der große Bogeneingang] kommt
*where [the big arch entrance] is*
(An arch-shaped gesture is affiliated to bracketed speech)

Note that the terms “Haupteingang” (*main entrance*) and “Bogen” (*arch*) are fused into the shortened term “Bogeneingang” (*arch entrance*). It is the gesture that remains (more or less) constant and allows for the identification of the second phrase as a repetition of the first. The concept “arch-shaped entrance gate of park” is associated with an arch-shaped gesture. We will take this alignment instance as an empirical example of a multimodal ensemble that is build by routinization during the course of a dialogue.

Simplifying the parts of an MME can also affect both gesture and speech, as the following example shows. Here, DG introduces a sculpture which is located on a round concrete base in the middle of a roundabout:

(5) die Skulptur, die die hat’n [Betonsockel]
*the sculpture, it it has a concrete base*
(A dynamic, two-handed rotund-shaped gesture is affiliated to bracketed speech)

Obviously, the noun “Betonsockel” (*concrete base*) and the rotund-shaped gesture cohere. Later on in the dialogue, the base of the sculpture is referred to again:

(6) rote Röhren auf’m [Sockel]
*red pipes on the [base]*
(A *C*-shaped static left hand gesture is affiliated to bracketed speech)

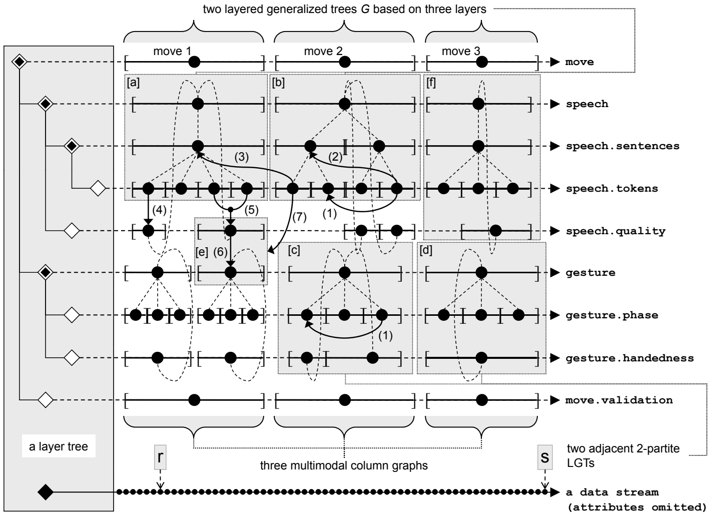

Fig. 1: Neighboring multimodal column graphs spanned by one data stream and several annotation layers.

Not only is the compound noun reduced from “Betonsockel” (*concrete base*) to its head “Sockel” (*base*), but also the gesture is simplified from a bi-manual movement to a left hand static gesture.

According to Pickering & Garrod ([4, p. 181]), “[a] *routine* is an expression, that is ‘fixed’ to a relatively great extent.” As examples (1) to (6) illustrated, an MME is a “relatively fixed” (cross-modal) expression, that is, a routine. What is striking, however, is that dialogue interactants can exploit the mutual informativity of the elements of an MME in order to simplify their form. If this is a rule, we expect that the words in multimodal communication exhibit a different frequency distribution from words in written texts, whose frequency distribution is known to obey Zipf’s law. This hypothesis was tested for the first time in [6], where indeed “frequency fingerprints” of multimodal ensembles were detected. Thus, based on the empirical examples given above and the frequency distribution analysis of [6] we regard semiotic alignment as a genuine and distinctive property of multimodal dialogue, which has to be accounted for in computational models of human communication. Accordingly, in the next section, a graph model is developed which is expressive enough to represent MMEs and to make them usabel as data input to machine learning. Although the empirical references given above are restricted to two modalities, namely speech and gesture, we want to emphasize that the structural model is general enough to cover any formations on the whole multimodal palette.

## III. A Graph-Theoretical Model of Structure Formation in Multimodal Annotations

In this section, we introduce a graph model of MMEs. The aim of this model is to pave the way to their automatic classification. We start from the idea that MMEs can be described as polypartite graphs distributed among multimodal layers such that their intra- or interpersonal occurrences get more and more similar or repetitive in the course of a dialogue. That is, we assume an increasing structural similarity of the manifestations of multimodal ensembles which finally emerge as repetitive occurrences of the same sign. This assumption can be reformulated by the following hypothesis:

> *The longer a dialogue, the more routinized the manifestations of a multimodal ensemble, the more similar its neighboring occurrences (generated by the same or different interlocutors) in that dialogue.*

From this perspective, a machine-learning-model of MME is a classification model of a certain class of graphs which are ordered in time. Thus, in order to classify MME we need to combine three areas: (i) graph modeling of semiotic structures [7], (ii) machine learning of graph structures [8] and (iii) dynamical systems analysis [9]. In this paper, we focus on the first of these areas as a preparatory step towards a classificatory model of the routinization of dialogical communication by means of MMEs. In order to do that we proceed as follows: in Section III-A, we start with a graph model of MMEs in terms of multimodal column graphs. Then, in Section III-B, we describe the learning of MMEs, that is, the classification of instances of this graph model as a novel task in machine learning.

### *A. The Graph Model*

When it comes to modeling discourse structures in terms of graph theory, two approaches can be distinguished. Firstly, *resource-oriented models* use graph theory in order to integrate linguistic annotations with their underlying resources [10]. Secondly, *annotation structure-oriented models* focus more narrowly on the structure formation of linguistic annotations [11]. We follow these approaches, but go beyond the latter in one respect: We do not only include multimodal annotation layers and thus go beyond graph models whose vertices solely denote linguistic units, but we also focus on the (possibly discontinuous) structure formation *within as well as between such annotation layers*. This is done in order to come up with a structural model of multimodal discourse as a reference point of its segmentation and classification. The reason is that we aim at investigating subgraphs as models of *recurrent* substructures in multimodal discourse as routinized manifestations of MMEs.

Our graph model of MMEs is mainly based on the notion of a *multimodal column graph*. To present this model stepwise, we use Figure 1 as a golden thread. This is done by means of *generalized trees* (GTs) which have been introduced to map nearly hierarchical structures in semiotic [7] as well as in biological systems [12]. We extend [7] by mapping GTs onto annotation layers (cf. Figure 1). This is done in Definition 1–4:

**Definition 1**. ***Generalized Trees.*** *Let $T = (V, E, x, \mathcal{O})$ be a directed ordered rooted tree with vertex set $V$, edge set $E$, root $x$ and order relation $\mathcal{O} \subset V^2$ which for each vertex $v$ linearly orders the set $\{w \,|\, (v, w) \in E\}$ of vertices to which $v$ is adjacent. Let further $P_{x,v} = (v_{i_0}, e_{j_1}, v_{i_1}, \dots, v_{i_{n-1}}, e_{j_n}, v_{i_n})$, $v_{i_0} = x, v_{i_n} = v$, $e_{j_k} = (v_{i_{k-1}}, v_{i_k}) \in E$, $k = 1, \dots, n$, be the unique path in $T$ from $x$ to $v \in V$. We denote the set of all vertices of $P_{x,v}$ by $V(P_{x,v})$. A* Generalized Tree (GT)

$$GT = (V, E_{[1]}, E_{[2]}, E_{[3]}, E_{[4]}, E_{[5]}, E_{[6]}, E_{[7]}, x)$$

*induced by $T$ is a graph whose partitioned edge set is incrementally defined:*

$$\begin{array}{rrcl}
\textit{kernel edges:} & E_{[1]} & = & E \\
\textit{upwards edges:} & E_{[2]} & \subseteq & E_u = \{(v, w) \,|\, w \in V(P_{x,v}) \setminus \{v\}\} \\
\textit{downwards edges:} & E_{[3]} & \subseteq & E_d = \{(v, w) \,|\, v \in V(P_{x,w}) \setminus \{w\}\} \\
\textit{reflexive edges:} & E_{[4]} & \subseteq & E_r = \{(v, v) \,|\, v \in V\} \\
\textit{lateral edges:} & E_{[5]} & \subseteq & V^2 \setminus (E \cup E_u \cup E_d \cup E_r) \\
\textit{sequential edges:} & E_{[6]} & = & \mathcal{O} \\
\textit{external edges:} & E_{[7]} & = & \emptyset
\end{array}$$

*To simplify our notation, we write $(V, E_{[1..7]}, x)$ when denoting generalized trees. Further, we write $e \in E_{[1..7]}$ if $e \in \cup_{i=1}^{7} E_{[i]}$. Note that we do not claim that all sets $E_{[1..7]}$ are pairwise disjunct. Obviously, each generalized tree $GT = (V, E_{[1..7]}, x)$ induces a rooted directed tree $T(GT) = T$.*

*Example 1. In Figure 1, we observe, amongst others, two GTs denoted by [a] and [b].*

**Definition 2**. ***Extended Graphs and Generalized Trees.*** *Let $\mathbb{G}$ be a set of graphs and $G = (V, E) \in \mathbb{G}$. $\bar{G} = (\bar{V}, \bar{E})$ is called an* extension of $G$ with respect to $\mathbb{G}$ *if there exists a subset $X \subseteq \{(v, w) \,|\, \exists (V', E') \in \mathbb{G} \setminus G \colon v \in V \wedge w \in V'\}$ so that $\bar{E} = E \cup X$ and $\bar{V} = V \cup (\cup_{(v,w) \in X} \{w\})$. Elements of $X$ are called* external edges *of $\bar{G}$. In the case that $G = (V, E_{[1..7]}, x)$ is a GT, we set $E_{[7]} = X$, write $\bar{G} = (\bar{V}, E_{[1..7]}, x)$ and speak of an* Extended Generalized Tree (GT).

*Example 2. In Figure 1, the GT [b] contains several kernel edges (dashed lines), two lateral edges (numbered by (1) and (2)) and two external edges (denoted by (3) and (7)).*

Note that Definition 2 guarantees that the sets $\cup_{i=1}^{6} E_{[i]}$ and $E_{[7]}$ are disjoint. Further, $E_{[7]}$ may be empty so that GTs can be seen to extend themselves.

**Definition 3**. ***Labeled Extended Generalized Trees.*** *A* labeled extended generalized tree *$G = (V, E_{[1..7]}, x, l_V, l_E)$ is an extended GT $G = (V,\ E_{[1..7]}, x)$ such that $l_V : V \rightarrow X$ and $l_E : E \rightarrow Y$ provide vertex and edge labels for two sets of labels $X$ and $Y$.*

In the following definitions, we assume that GTs and their derivations are always labeled although we do not continue to specify the labeling functions $l_V, l_E$ within these definitions.

In order to utilize EGTs as building blocks of a model of MMEs, we introduce a mapping onto annotation layers (cf. Figure 1). This mapping reflects a hierarchical layer model as inherent to many annotation models of multimodal data [13], [14]. In order to capture this hierarchical ordering, we introduce the notion of a *layered generalized tree* — to keep names simple, we skip the attribute *extended* within the following definitions, but nevertheless start from Definition 2.

**Definition 4**. ***Layered Generalized Trees.*** *A* Layered Generalized Tree (LGT) $G = (V, E_{[1..7]}, x, L)$ *is an extended GT* $(V, E_{[1..7]}, x)$ *together with a mapping* $L\colon V \to \mathbb{L}$ *such that* $(\mathbb{L}, \leq)$ *is a lattice-ordered set and*

$$\begin{array}{rl} \text{layer of the root:} & L(x) = \inf(\mathbb{L}) \\ \text{kernel edges:} & \forall (v,w) \in E_{[1]} \colon L(w) = \inf\{l \in \mathbb{L} \,|\, l > L(v)\} \\ \text{upwards edges:} & \forall (v,w) \in E_{[2]} \colon L(v) > L(w) \\ \text{downwards edges:} & \forall (v,w) \in E_{[3]} \colon L(v) < L(w) \\ \text{sequential edges:} & \forall (v,w) \in E_{[6]} \colon L(v) = L(w) \\ \text{external edges:} & \forall (v,w) \in E_{[7]} \colon L(w) \text{ is undefined} \end{array}$$

*The elements of* $\mathbb{L}$ *are called* layers *of* $G$. $\mathrm{mode}(L) = \inf(\mathbb{L})$ *is called* mode.

*Example 3. In Figure 1, the GT [a] is layered by means of the layers* speech *(i.e., the mode of the LGT [a]),* speech.sentences*, and* speech.tokens.

**Definition 5**. ***Polypartite Layered Generalized Trees.*** *Let* $\mathbb{G} = \{G_1, \ldots, G_n\}$ *be a set of root-identical LGTs* $G_i = (V_i, E_{[1..7]_i}, x, L_i) \in \mathbb{G}$ *such that* $L_1(x) = \ldots = L_n(x)$ *and* $\forall i, j \in \{1, \ldots, n\}, i \neq j \colon \mathbb{L}_i \cap \mathbb{L}_j \setminus \{L_1(x)\} = \emptyset$. *A* Polypartite Layered Generalized Tree (PLGT) $G = (V, E_{[1..7]}, x, L)$ *is a LGT such that* $V = \cup_{i=1}^n V_i$, $\forall k \in \{1,2,3,4\} \colon E_{[k]} = \cup_{i=1}^n E_{[k]_i}$, $L = \cup_{i=1}^n L_i$, $L_i \colon V_i \to \mathbb{L}_i$, $L \colon V \to \mathbb{L} = \cup_{i=1}^n \mathbb{L}_i$, $\inf(\mathbb{L}) = L_1(x)$, *and*

$$\begin{array}{rcl} \textit{lateral edges:}\; E_{[5]} & = & \cup_{i=1}^n (E_{[5]_i} \\ & & \cup \{(v,w) \in E_{[7]_i} \,|\, w \in \cup_{j=1, j\neq i}^n V_j\}) \\ \textit{sequential edges:}\; E_{[6]} & : & \forall v \in V, \forall l \in \mathbb{L} \colon E_{[6]} \textit{ induces a} \\ & & \textit{linear order over} \\ & & \{w \,|\, L(w) = l, (v,w) \in E_{[1]}\} \\ \textit{external edges:}\; E_{[7]} & = & \cup_{i=1}^n \left(E_{[5]_i} \setminus \right. \\ & & \left. \{(v,w) \in E_{[7]_i} \,|\, w \in \cup_{j=1, j\neq i}^n V_j\}\right) \end{array}$$

*We write* $\mathbb{L}(G) = (\mathbb{L}, \leq_{\mathbb{L}})$ *to denote the lattice ordered set* $\mathbb{L}$ *of layers onto which* $G$ *is mapped.*

*Example 4. In Figure 1, [c] and [d] span 2-partite PLGTs mapped onto the partially ordered layers* gesture, gesture.phase *and* gesture.handedness. *Another example is the 2-partite PLGT [f] which is mapped, among other things, onto the layers* speech.sentence *and* speech.quality *which are dominated by the same layer* speech.

Note that in Definition 5, we do not demand that $E_{[6]}$ induces a linear order over the set of all vertices dominated by any focal vertex $v$ (as done in Definition 1). Rather, we demand now that this linear order holds separately for the sets of vertices of the same layer. Obviously, each LGT $G$ also spans a PLGT where $\mathbb{G} = \{G\}$. We retain this manner of speaking to grasp mono-partite graphs under the same notion.

**Definition 6**. ***Data Streams, Timestamps and Intervals.*** *A* data stream *is an attributed time-aligned graph* $\mathcal{S} = (S, \preceq, \tau, \alpha)$ *induced by a lattice-ordered set* $(S, \preceq)$ *where* $\preceq$ *is a linear order relation over* $S$. Time-aligned *means that* $\tau \colon S \to \mathbb{Q}_0^+$ *is a total injective mapping of vertices* $s \in S$ *onto timestamps* $\tau(s)$ *modeled as rational numbers such that*

$$\begin{array}{c} \forall r, s, t \in S \colon \; ((r \prec s \prec t \wedge \nexists x \in S, x \neq s \colon (r \prec x \prec t)) \\ \to \tau(t) - \tau(s) = \tau(s) - \tau(r)) \end{array}$$

$\mathcal{S}$ *is* attributed *in the sense that* $\alpha \colon S \to \mathrm{Pot}([Q_0^+]^\infty)$ *is a mapping of vertices* $s \in S$ *onto sets of elements of* $[Q_0^+]^\infty = \cup_{i=1}^\infty (Q_0^+)^i$ *which models feature value structures in terms of scalars, vectors or matrices. Elements* $r, s \in S$ *are called* stamps *which span so-called* intervals $[r, s] = \{x \,|\, \tau(r) \leq \tau(x) \leq \tau(s)\}$. *Analogously, elements* $p, q \in \mathbb{Q}_0^+$ *are called* timestamps *which span so-called* time intervals $[p, q]$.

*Example 5. Figure 1 contains exactly one data stream. This may be a video stream or a stream of tracking data where each stamp is attributed by a data matrix of tracking data.*

Obviously, the function $\tau$ of a data stream defines a clock speed. Based on this preparatory definition we can now define so-called stream-aligned PLGTs:

**Definition 7**. ***Stream-Aligned PLGT.*** *Let* $\mathcal{S} = (S, \preceq, \tau, \alpha)$ *be a data stream. A* stream-aligned PLGT $G(\mathcal{S}) = (V, E_{[1..7]}, x, L, \theta_{\mathcal{S}})$ *induced by* $\mathcal{S}$ *is a PLGT* $G = (V, E_{[1..7]}, x, L)$ *together with a total function* $\theta_{\mathcal{S}} \colon V \to S^2$ *such that*

$$\begin{array}{rcl} \textit{kernel edges:}\; E_{[1]} & : & \forall (v,w) \in E_{[1]}, \theta_{\mathcal{S}}(v) = (r,s), \\ & & \theta_{\mathcal{S}}(w) = (t,u) \colon \\ & & \tau(r) \leq \tau(t) \leq \tau(u) \leq \tau(s) \\ \textit{sequential edges:}\; E_{[6]} & : & \forall (v,w) \in E_{[6]}, \theta_{\mathcal{S}}(v) = (r,s), \\ & & \theta_{\mathcal{S}}(w) = (t,u) \colon \\ & & \tau(r) < \tau(s) \leq \tau(t) < \tau(u) \;\vee \\ & & \tau(r) \leq \tau(s) < \tau(t) \leq \tau(u) \end{array}$$

**Definition 8**. ***Adjacency and Parallelism.*** *Let* $\mathcal{S} = (S, \preceq, \tau, \alpha)$ *be a data stream and* $[r, s]$ *an interval in* $\mathcal{S}$. *Two disconnected stream-aligned PLGTs* $G_1(\mathcal{S}) = (V_1, E_{[1..7]_1}, x, L_1, \theta'_{\mathcal{S}})$, $G_2(\mathcal{S}) = (V_2, E_{[1..7]_2}, y, L_2, \theta''_{\mathcal{S}})$ *are called* adjacent *in* $[r, s]$ *iff*

$$\begin{array}{lcl} L_1 \colon V_1 \to \mathbb{L} & \wedge & L_2 \colon V_2 \to \mathbb{L} \\ & \wedge & \tau(r) \leq \min(\inf\{\tau(t) \,|\, v \in V_1 \wedge \theta'_{\mathcal{S}}(v) = (t,u)\}, \\ & & \inf\{\tau(t) \,|\, v \in V_2 \wedge \theta''_{\mathcal{S}}(v) = (t,u)\}) \\ & \wedge & \tau(s) \geq \max(\sup\{\tau(u) \,|\, v \in V_1 \wedge \theta'_{\mathcal{S}}(v) = (t,u)\}, \\ & & \sup\{\tau(u) \,|\, v \in V_2 \wedge \theta''_{\mathcal{S}}(v) = (t,u)\}) \end{array}$$

*In the case that* $L_1 \colon V_1 \to \mathbb{L}_1$, $L_2 \colon V_2 \to \mathbb{L}_2$, *where* $\mathbb{L}_1 \cap \mathbb{L}_2 = \emptyset$, *we call* $G_1(\mathcal{S})$ *and* $G_2(\mathcal{S})$ parallel *in* $[r, s]$.

*Example 6. In Figure 1, the PLGTs [c] and [d] are adjacent in the interval* $[r, s]$. *Further, they are parallel to the PLGT [b] in* $[r, s]$, *but not to the PLGT [a] in* $[r, s]$.

**Definition 9**. ***Multimodal Column Graphs.*** *Let* $\mathcal{S} = (S, \preceq, \tau, \alpha)$ *be a data stream and* $[r, s]$ *an interval in* $\mathcal{S}$. *Let further* $\mathbb{G}(\mathcal{S}) = \{G_1(\mathcal{S}), \ldots, G_n(\mathcal{S})\}$ *be a set of parallel or* *adjacent stream-aligned PLGTs* $G_i(\mathcal{S}) = (V_i, E_i, x_i, L_i, \theta_{i_\mathcal{S}})$, $i \in \{1, \ldots, n\}$, *in* $[r, s]$. *We use* $\mathbb{G}(\mathcal{S})$ *to build a graph*

$$G_{\mathbb{G}(\mathcal{S})} = (\cup_{i=1}^n V_i, \cup_{i=1}^n E_i, \cup_{i=1}^n L_i, \cup_{i=1}^n \theta_{i_\mathcal{S}}, [r, s])$$

*define*

$$\mathcal{L}(G_{\mathbb{G}(\mathcal{S})}) = \{(\mathbb{L}_i, \leq_{\mathbb{L}_i}) \,|\, G_i(\mathcal{S}) \in \mathbb{G}(\mathcal{S})\}$$

*and call*

$$\text{modes}(G_{\mathbb{G}(\mathcal{S})}) = \{\inf(\mathbb{L}_i) | (\mathbb{L}_i, \leq_{\mathbb{L}_i}) \in \mathcal{L}(G_{\mathbb{G}(\mathcal{S})})\} \ ,$$

*the* set of modes *of* $G_{\mathbb{G}(\mathcal{S})}$. *The cardinality* $|\text{modes}(G_{\mathbb{G}(\mathcal{S})})|$ *is called the* degree of modality *of* $G_{\mathbb{G}(\mathcal{S})}$. *If* $|\text{modes}(G_{\mathbb{G}(\mathcal{S})})| > 1$, $G_{\mathbb{G}(\mathcal{S})}$ *is called* Multimodal Column Graph (MCG) induced by $\mathbb{G}(\mathcal{S})$ in $[r, s]$.

MCGs are our generic graph-theoretical model of *Multimodal Ensembles* (MMEs). That is, we propose to represent MMEs as families of graphs with a kernel hierarchical structure such that these graphs are, firstly, vertically ordered by a mapping onto layers (of some modes) and, secondly, horizontally ordered by a mapping onto some data stream. In simple cases, a MME is manifested by a single lexical item parallel to an elementary gesture. In more complex cases, a MME consists of composite verbal and gestural units. In any event, the notion of adjacency and parallelism allows us to relax the condition that the constituents of a MME of different mode occur in exactly the same interval. Rather, these occurrences may overlap to some degree so that adjacency can be seen as a fuzzy relation.

**Definition 10**. ***Reference Graph and Reference Layer.*** *Let* $\mathcal{S} = (S, \preceq, \tau, \alpha)$ *be a data stream and* $RG = (\{m_1, \ldots, m_n\}, \emptyset, L_{RG}, \theta_\mathcal{S})$ *be a graph such that* $L_{RG}(m_1) = \ldots = L_{RG}(m_n) = \text{rel}$ *and* $\forall i \in \{1, \ldots, n\}$, $\theta_\mathcal{S}(m_i) = (r_i, s_i)$, $\theta_\mathcal{S}(m_{i+1}) = (r_{i+1}, s_{i+1}) : \tau(r_i) \leq \tau(s_i) < \tau(r_{i+1}) \leq \tau(s_{i+1})$. *We call* $RG$ *a* Reference Graph *and* rel *a* Reference Layer.

**Definition 11**. ***Multimodal Sequence Graphs.*** *Let* $\mathcal{S} = (S, \preceq, \tau, \alpha)$ *be a data stream,* $RG = (\{m_1, \ldots, m_n\}, \emptyset, L_R, \theta_\mathcal{S})$ *a reference graph with the reference layer* rel. *Further, let*

$$\mathbb{M}(\mathcal{S}, T(\text{rel})) = \{G_{\mathbb{G}_1(\mathcal{S})}, \ldots, G_{\mathbb{G}_n(\mathcal{S})}\}$$

*be a set of multimodal column graphs (MCGs)*

$$G_{\mathbb{G}_i(\mathcal{S})} = (V_i, E_i, L_i, \theta_{i_\mathcal{S}}, [r_i, s_i])$$

*induced by* $\mathbb{G}_i(\mathcal{S})$ *in the interval* $[r_i, s_i]$ *(cf. Def. 10) where* $\theta_{i_\mathcal{S}}(m_i) = (r_i, s_i)$ *and* $\forall i, j \in \{1, \ldots, n\} : \mathcal{L}(G_{\mathbb{G}_i(\mathcal{S})}) = \mathcal{L}(G_{\mathbb{G}_j(\mathcal{S})})$. *Further, we set*

$$\begin{aligned} V_T &= \cup_{(\mathbb{L}_i, \leq_{\mathbb{L}_i}) \in \mathcal{L}(G_{\mathbb{G}_i(\mathcal{S})}), G_{\mathbb{G}_i(\mathcal{S})} \in \mathbb{M}(\mathcal{S}, T(\text{rel}))} \mathbb{L}_i \\ E_T &= \{(\text{rel}, l) \,|\, l \in \text{modes}(G_{\mathbb{G}_i(\mathcal{S})})\} \cup \\ &\quad \left( \cup_{(\mathbb{L}_i, \leq_{\mathbb{L}_i}) \in \mathcal{L}(G_{\mathbb{G}_i(\mathcal{S})}), G_{\mathbb{G}_i(\mathcal{S})} \in \mathbb{M}(\mathcal{S}, T(\text{rel}))} \leq_{\mathbb{L}_i} \right) \end{aligned}$$

*and demand that* rel $\notin V_T$ *in order to build a directed rooted tree*

$$T(\text{rel}) = (V_T \cup \{\text{rel}\}, E_T, \text{rel})$$

*which is called* Layer Tree. *Now, let* $\sqsubseteq \subset \mathbb{M}(\mathcal{S}, T(\text{rel}))^2$ *be an order relation over* $\mathbb{M}(\mathcal{S}, T(\text{rel}))$ *such that*

$$G_{\mathbb{G}_i((\mathcal{S})} \sqsubseteq G_{\mathbb{G}_j(\mathcal{S})} \text{ iff } \tau(r_i) \leq \tau(s_i) < \tau(r_j) \leq \tau(s_j)$$

*If* $\sqsubseteq$ *is a linear order relation over* $\mathbb{M}(\mathcal{S}, T(\text{rel}))$, *we call* $(\mathbb{M}(\mathcal{S}, T(\text{rel})), \sqsubseteq)$ *a* Multimodal Sequence Graph (MSG) *with vertex set* $\mathbb{M}(\mathcal{S}, T(\text{rel}))$ *and edge set* $\sqsubseteq$.

*Example 7. In Figure 1, the layer* `move` *spans a reference graph based on its vertices named* `move 1, 2,` *and* `3`, *respectively. The stream-alignment of LGTs mapped onto this layer, that is, their mapping onto some data stream induces, in turn, intervals which allow us to delineate three multimodal column graphs (cf. Figure 1). Obviously, these three MCGs are linearly ordered so that Figure 1 demonstrates a multimodal sequence graph whose layer tree is visualized on the left of Figure 1 with* rel =`move`. *We also observe (but do not prove) that* $(\mathbb{M}(\mathcal{S}, T(\text{rel})), \sqsubseteq)$ *is a lattice ordered set.*

So far we have basically considered digraphs and, thus, disregarded hypergraphs whose edges cover more than one start and end vertex. Our aim was to model *constituency relations* (by means of kernel and sequential edges), *dependency relations* (by means of upwards, downwards, lateral and external edges), *cross-modal relations* (by means of external edges), stratification (by means of layered GTs and column graphs) and time-related structures (by means of data streams and sequence graphs). These notions can be exemplified as follows:

- Figure 1 contains two edges labeled by 1 which demonstrate discontinuous dependencies: On the level of lexical items, e.g., within the layer `speech.tokens`, this is exemplified by lexical cohesion, e.g., the reiteration of semantically related units [15]. Think, for example, of the DG (see Section II) uttering "`The long bolt. No, the screw.`" In this case, `screw` is sense-related to `bolt` indicating an alternative description of the same object. As cohesion relations naturally go beyond sentence boundaries, reference relations as well as substitutions or ellipses are, likewise, examples of discontinuous dependencies. On the level of gestures (e.g. within the layer named `gesture.phase`), such dependencies are exemplified by a pointing gesture which is interrupted by an iconic one. In order to annotate the cohesion of the discontinuous parts of the pointing gesture, we can use an external edge which manifests their continuation.
- The lateral Edges 2 and 3 in Figure 1 are exemplified by nomination/denomination (or expansion/condensation) relations. In the case of Edge 3, this relates to adjacent LGTs, while Edge 2 interrelates elements of the same LGT. Think, for example, of the DG performing a definition act (e.g. uttering "`I will call the long green panel a plate`") in order to introduce a dialogue-specific label (i.e. `plate`) which later on is reiterated. Another example is given by rhetorical relations of text spans of different levels of linguistic resolution.

- From the point of view of LGT [a], Edge 4 manifests an *external* edge while it is a *lateral* edge in terms of the encompassing PLGT in which it connects different LGTs. Edge 4 is exemplified by a verbal manifestation of a reference-semantic object quality (e.g. height, shape or color) by means of a single token (e.g. the long bolt). Analogously, Edge 5 connects the same LGTs, but by a hyperedge with two start vertices. Such edges are exemplified by phrases like "The large, not the small bolt" where both attributes have the same reference point. A more elaborate example is a multimodal multinuclear rhetorical relation which links a gesture (in the role of a nonverbal nucleus) with at least two related text spans.
- Finally, Edge 7 of Figure 1 connects a vertex with a graph based on Edge 6 which connects two vertices of different modes. Within the present approach, such a link is exemplified by a multimodal nomination/denomination in which the DG introduces a name for a *multimodal* sign which is reiterated later on.

Apart from Edge 5 and 7, the present graph model allows mapping all these relations. In order to model heterogeneous, but nevertheless directed relations of the sort of Edge 5, we need to redefine our graph-theoretical apparatus in terms of hypergraphs. Edge 7 demands a more elaborate graph model in which graphs are nested and, thus, can be referred to as vertices on their own. In any case, we have to state that none of the existing multimodal annotation tools allows visualizing such relations. This visualization gap also relates to *discontinuous dependency relations* and, thus, to the Edges 1-3. In this sense, multimodal annotation tools which cover the full range of multimodal structure formation are still an open field of research. In any event, this small set of examples shows some aspects of the complexity of structure formation one has to face when dealing with multimodal communication.

### B. Alignment as Classification

Based on our graph model we now define a classification task in terms of alignment in multimodal communication, whether intra- or inter-personal. The idea is to think of multimodal column graphs as formal models of the multimodal manifestation units of such alignment processes, that is, of MMEs. In order to maintain a general terminology, we define this task as follows:

**Task: MME Classification and Mining:** *Let* $(\mathbb{M}(\mathcal{S}, \mathrm{rel}), \sqsubseteq)$ *be an MSG and* $\delta \colon \mathbb{M}(\mathcal{S}, \mathrm{rel})^2 \rightarrow [0, 1]$ *a metric. Further, let* $\mathbb{C} = \{C_i \,|\, i \in I\}$ *be a reference partitioning of* $\mathbb{M}(\mathcal{S}, \mathrm{rel}) = \{G_{\mathbb{G}_1(\mathcal{S})}, \ldots, G_{\mathbb{G}_n(\mathcal{S})}\}$ *such that* $\mathbb{M}(\mathcal{S}, \mathrm{rel}) = \cup_{i \in I} C_i$ *and* $\forall C_i \in \mathbb{C} \; \forall G_{\mathbb{G}_i(\mathcal{S})}, G_{\mathbb{G}_j(\mathcal{S})} \in C_i : \delta(G_{\mathbb{G}_i(\mathcal{S})}, G_{\mathbb{G}_j(\mathcal{S})}) \ll 1$ *— that is, the classes of* $\mathbb{C}$ *contain MCGs which are "similar" in the sense of the metric* $\delta$*. Based on these preliminaries, we define two tasks*

- MME Classification *is the task of learning* $\mathbb{C}$.
- MME-Alignment Mining *is the task of detecting multimodal column graphs* $G_{\mathbb{G}_i(\mathcal{S})}, G_{\mathbb{G}_j(\mathcal{S})}$ *whose similarity (in the sense of* $\delta$*) is both higher than expected by chance and the higher the later they occur (i.e, the higher the indices* $i, j$*).*

The basic idea of this definition is that dialogical alignment is a process of structural stabilization which results in a decreasing entropy of the spectrum of MMEs as the dialogue unfolds. That is, we expect that alignment occurs not only in terms of more and more shared and stable choices of elementary units, but also in terms of aligned MMEs whose recurrent usage increases the similarity of adjacent column graphs used to represent them. Exploring such units equals the segmentation of recurrent fragments of multimodal discourse. Obviously, to perform Task 1 we need a powerful graph similarity measure (or metric $\delta$) *beyond the apparatus of similarity or distance measures operating on trees [16] or feature vectors [17].* In this sense, Task 1 represents a novel task in machine learning, that is, the detection and classification of MMEs.

## References

[1] D. Lewis, *Conventions. A Philosophical Study*. Cambridge, Massachusetts: Harvard U.P., 1969.

[2] H. H. Clark, *Using Language*. Cambridge: Cambridge U. P., 2000.

[3] S. Garrod and A. Anderson, "Saying what you mean in dialogue: a study in conceptual and semantic co-ordination," *Cognition*, vol. 27, no. 2, pp. 181–218, 1987.

[4] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences*, vol. 27, pp. 169–226, 2004.

[5] S. Kopp, H. Rieser, I. Wachsmuth, K. Bergmann, and A. Lücking, "Speech-gesture alignment," Project Panel at the 3rd Int. Conf. of the Int. Society for Gesture Studies, 2007.

[6] A. Lücking, A. Mehler, and P. Menke, "Taking fingerprints of speech-and-gesture ensembles: Approaching empirical evidence of intrapersonal alignment in multimodal communication," in *Proc. of LONDIAL 2008*, King's College London, June 2–4 2008, pp. 157–164.

[7] A. Mehler, "Generalised shortest paths trees: A novel graph class applied to semiotic networks," in *Analysis of Complex Networks: From Biology to Linguistics*, M. Dehmer and F.-E. Streib, Eds. Weinheim: Wiley-VCH, 2009.

[8] H. Bunke, S. Günter, and X. Jiang, "Towards bridging the gap between statistical and structural pattern recognition," in *Proc. of the 2nd Int. Conf. on Advances in Pattern Recognition*. Berlin: Springer, 2001.

[9] D. Hinrichsen and A. J. Pritchard, *Mathematical Systems Theory I — Modelling, State Space Analysis, Stability and Robustness*. Berlin/New York: Springer, 2005.

[10] S. Evert, J. Carletta, T. J. O'Donnell, J. Kilgour, A. Vögele, and H. Voormann, "The nite object model," IMS, University of Stuttgart, Version 2.1, 2003.

[11] T. Schmidt, *Computergestützte Transkription — Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln*. Frankfurt am Main: Peter Lang, 2005.

[12] M. Dehmer, A. Mehler, and F. Emmert-Streib, "Graph-theoretical characterizations of generalized trees," in *Proc. of the 2007 International Conference on Machine Learning: Models, Technologies & Applications (MLMTA'07), June 25-28, 2007, Las Vegas*, 2007.

[13] M. Kipp, *Gesture Generation by Imitation — From Human Behavior to Computer Character Animation*. Boca Raton: Dissertation.com, 2004.

[14] H. Brugman, P. Wittenburg, S. C. Levinson, and S. Kita, "Multimodal annotations in gesture and sign language studies," in *Proc. of LREC 2002*, 2002, pp. 176–182.

[15] A. Budanitsky and G. Hirst, "Evaluating WordNet-based measures of lexical semantic relatedness," *Computational Linguistics*, vol. 32, no. 1, pp. 13–47, 2006.

[16] P. Bille, "A survey on tree edit distance and related problems," *Journal of Theoretical Computer Science*, vol. 337, no. 1-3, pp. 217–239, 2005.

[17] R. Serafin and B. D. Eugenio, "FLSA: Extending latent semantic analysis with features for dialogue act classification," in *Proc. of ACL '04*, 2004, pp. 692–699.