Towards a Conceptual, Unification-Based Speech-Gesture Interface

Andy Lücking

Goethe-University Frankfurt, D-60325 Frankfurt am Main, Germany
Luecking@em.uni-frankfurt.de

Abstract. A framework for grounding the semantics of co-verbal iconic gestures is presented. A resemblance account to iconicity is discarded in favor of an exemplification approach. It is sketched how exemplification can be captured within a unification-based grammar that provides a conceptual interface. Gestures modeled as vector sequences are the exemplificational base. Some hypotheses that follow from the general account are pointed at and remaining challenges are discussed.

1 Background

People are gesturing in nearly all communication situations, be it in face-to-face dialog [1] or on the telephone [2]. The gestures they produce are informative [3]. The term "gesture" is used here as denoting co-verbal hand or arm movements that relate to the interlocutors' narration [4]. Gestures in this sense have a tripartite structure: they can be distinguished into a preparation, a stroke, and a retraction phase [5]. The stroke phase is the semantically significant phase. Accordingly, "gesture" as used here denotes strokes. Gestures can be distinguished into at least three basic types, viz. deictic gestures (i.e. pointings), iconic gestures (figurative gestures), and beat gestures (rhythmic gestures) [6]. The focus in this article is on iconic gestures.

Gestures relate to one or more words of their accompanying speech – their *affiliate* [7, 8]. The affiliate is singled out on temporal, acoustic, and semantic grounds [9, 10, 11, 8]. However, problems of affiliation are ignored in the following; rather, an account for the meaningfulness of iconic gestures is presented that builds on Goodman's notion of exemplification instead of iconicity as resemblance (Sections 2 and 3). Exemplification is incorporated into a HPSG-style grammar, giving rise to a conceptual, unification-based speech-gesture interface (Section 4). Section 5 covers some extensions and current limitations of the interface. In the concluding Section 6, two hypotheses are posed that follow from the account presented here.

2 Resemblance?

According to the commonplace conception of iconic signs, the meaning of iconic gestures is grounded on resemblance of the form of the gesture and the entity

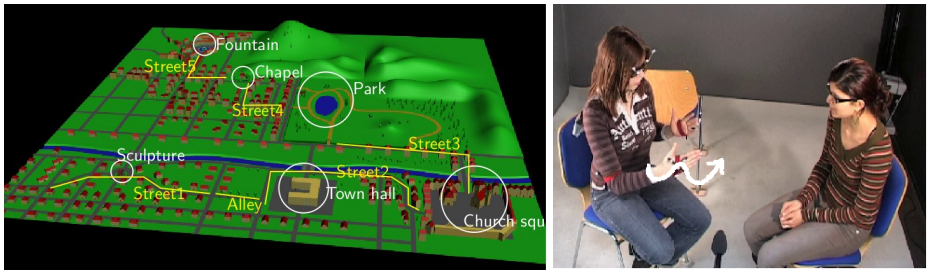


Fig. 1. The route through the SaGA town alongside some landmarks (the figure is taken from [20]) and the gestural depiction of Street 2.

depicted. This view goes back to the semiotic work of C. S. Peirce who hold that icons are representation "whose relation to their objects is a mere community in some quality" [12]. However, there are good reasons that resemblance is not strong enough to give rise to representational signs. Three of them run as follows:

- Resemblance is a relation that is symmetrical, reflexive, and transitive. The sign relation is neither of these [13, 14].
- In some respect or other anything is similar to anything else. Signs are not general to such an extent [13].
- "Some *symbolic means* is required to communicate both the fact that a sign is an icon and the respect in which it is iconic." [15, p. 676; my emphasis] (see also [16]).

Due to these reasons, the semantic account to iconic gestures developed in [8] builds on Goodman's notion of *exemplification* [14], instead of some kind of isomorphy (the usual explication of iconicity, cf. [10, 17]).[1] Exemplification in the semiotic sense used here can be understood quite conventionally as giving examples. For instance, the property denoted by the predicate "green" can be exemplified by a green thing. In the same vein, the predicate "circular" can be exemplified by a gesture that performs a circular trajectory.

Empirical examples suggest that exemplification is indeed operative in multimodal communication. Talking about a route through the virtual SaGA town (see Figure 1; stimulus in the *Speech and Gesture Alignment* corpus (SaGA) [19]), one interlocutor refers to Street 2 with the words "geschlängelte Straße" (*winded street*). As you can see in Figure 1, Street 2 makes a right-left turn. The gesture of the speaker, however, performs an additional turn. Thus, the gesture is not just a "literal" representation of the topic, but it nonetheless exemplifies its affiliated predicate (i.e., *winded*).

A further example is a speaker's depiction of the entry of the SaGA park. The verbal description (which translates to *until you eventually come to the main entrance, which is a large, gray arch*) is accompanied by an arch-shaped

[1] Once the significance of an iconic sign is established, however, it might well be reinterpreted in terms of *a posteriori* recognized resemblance [18].

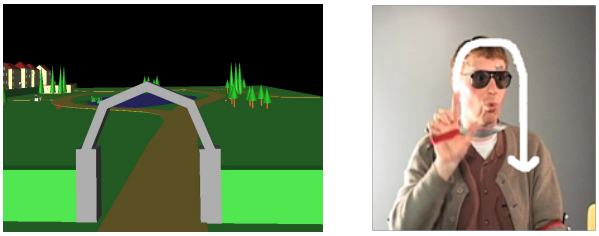


Fig. 2. Entry of the SaGA park and its gestural depiction.

gesture. Again, as you can see in Figure 2, the gesticulated arch and the original arch entry are quite dissimilar (e.g., angular vs. smoothly rounded).

In denotative semantic's terms, these gestures are simply false. However, there is no hint in the dialog context for any pragmatic inferences, self repairs or clarification requests – the number of turns of the winded street, for example, could be asked for by the addressee. The reason for the harmlessness of the denotative imprecision or even falsehood of gestures is due, I suppose, to the principle of multimodal communication that is known as "primacy of speech" [21]. That means that information delivered by speech and gesture are not on a par. Rather, the meaning of a gesture "is seen through" its affiliated speech. At the bottom line, the examples just presented (in addition to the arguments against iconicity as resemblance) illustrate that a general account to iconic gestures should not be given in terms of an (eventually weakened kind of) isomorphism. The meaning of a gesture appears to be dependent on the information delivered by its affiliate. In the following section, a semantic account is presented that tries to capture the co-text dependent role of gesture meaning, namely an account that rests on exemplification.

3 Exemplification

In semiotics, non-denotative signs are no foreign matter. A non-denotative account is exemplification [14]. The key idea is to regard iconic signs as things onto which verbal predicates are applied; the iconic sign is then said to exemplify these predicates. Thus, the denotation relation appears to be reversed.[2] Let us illustrate exemplification by means of a non-linguistic example. The wavelength interval that corresponds to the color 'green' according to the color space spanned by the German language is 497 to 530 nm. The feature structure for the predicate *green* of type *color* in (1) assembles this information. If some entity x exhibits a surface reflection (SR) of light of, say, 511 nm, as expressed in (2),

[2] Hence, "non-denotative" means that the gestures do not denote themselves, it does not mean that denotation plays no role at all.

Fig. 3. Motion carriers (taken from [22]).

(1) $\begin{bmatrix} \textit{color} & \\ \text{RELN} & \text{green} \\ \text{NM} & 497\text{–}530 \end{bmatrix}$ (2) $\begin{bmatrix} \textit{entity} & \\ \text{IND} & x \\ \text{SR} & 511 \end{bmatrix}$ (3) $\begin{bmatrix} \textit{sem-struc} & \\ \text{MODE} & \text{exemplification} \\ \text{ENT} & \begin{bmatrix} \text{IND} & x \\ \text{SR} & \boxed{3}511 \end{bmatrix} \\ \text{PRED} & \begin{bmatrix} \text{RELN} & \text{green} \\ \text{NM} & \boxed{3} \end{bmatrix} \end{bmatrix}$

then the entity and the color predicate can enter into an exemplification relation, as shown in (3).

Note that in order for this example to work in a unification-based grammar framework the nanometer range of the predicate has to be modeled as a complex type that is decomposed into its discrete numbers within the grammar's type hierarchy. Note further that under the proviso that their wavelengths do not overlap x can only exemplify the predicate *green* and no other color predicate.

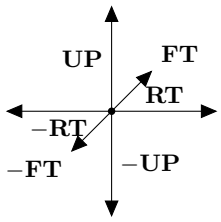
So far, so good. But how to handle iconic gestures? What is their "wavelength"? Which predicates do they exemplify?

An answer to the second question can be given from research on the perception of biological motion. The key result of a respective study was that the recognition and classification of biological motion is triggered by abstract motion *carriers* [22]. Figure 3 gives an illustration for walking and running motions. The carriers describe trajectories over the time course of the motion which can be expressed as vector sequences. Being biological motions, gestures can be modeled on the kinetic level in terms of vector representations. The vector representation then is the basis for the exemplification relation.

We already have a clue to an answer to the third question: a gesture exemplifies the predicates it is affiliated with – see the discussion above. How an exemplification account along the lines just outlined can be modeled within a unification-based grammar framework is indicated by means of the *winded* gesture in the following section. In order to process input on different modalities in the first place, a unification-based framework that is built around a multimodal chart parser [23, 24] is assumed.

4 Grammar

Three things are needed in order to pursue an exemplification account as introduced in preceding Section 3:



$mf \in$ **FT**
$mb \in$ **−FT**
$mu \in$ **UP**
$md \in$ **−UP**
$ml \in$ **−RT**
$mr \in$ **RT**

Fig. 4. Mapping of "morphological" predicates onto the gesture space as vector space (taken from [8, p. 171]).

1. A vectorization of gestures;
2. a conceptual, vector-based enrichment of the representation of verbal predicates; and
3. an exemplification interface.

The three ingredients are introduced subsequently.

4.1 Vectorization of gestures

The vectorization of gestures starts from their "morphological" representation as derived, for instance, by annotation. A detailed annotation has been provided by the SaGA corpus, where, amongst others, movements (m) and their directions (*f(orward)*, *b(backward)*, *u(pward)*, *d(ownward)*, *l(eft)*, *r(ight)*) of back of hand, palm, and wrist have been specified. These annotation predicates can be mapped onto the main axes of a vector space, as illustrated in Figure 4.

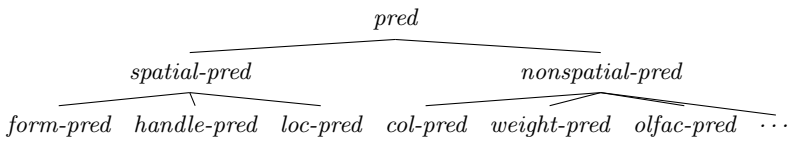
In order to capture the dynamics of motions, vectors are concatenated to *vector sequences* – see [8, p. 171 ff] for details. In effect, the vectorial representation provides a cognitively backed, sub-semantic model of gestures as biological motion. Their semantic significance acts out in the interplay with verbal predicates and the "conceptual vector meaning".

4.2 Conceptual vector meaning

In order to provide a conceptually rich semantic representation of spatial contents, an HPSG framework is enriched with structures in the spirit of the conceptualist approach of [25], where the underlying model is a situation model in the line of [26], which is substantially enriched with a vector space model in the line of [27] (see also [28]). Predicates related to spatio-temporal entities carry an explicit representation of their descriptive content as value of the newly introduced feature CVM (*Conceptual Vector Meaning*). The descriptive CVM content of our example adjective *winded*, for instance, is a sinusoidal vector sequence that has at least two turns (written $\sin_{[2,n]}$).

In order to ensure that only spatial predicates are equipped with the vector-based CVM feature, the type hierarchy bifurcates into spatial and non-spatial predicates, as shown in (4):

(4)



A constraint for the type *spatial-pred* introduces conceptual vector meaning:

(5)
$$\begin{bmatrix} \textit{spatial-pred} & \\ \text{CONT} & \begin{bmatrix} \text{RESTR} & \left\langle \begin{bmatrix} \text{CVM} & \begin{bmatrix} \text{VEC} & \text{vec.type} \\ \text{PATH} & \text{vec} \end{bmatrix} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

The type *vec.type* determines the kind of vector representation in question, that is, for example, whether it is a three-dimensional axis vector or a four-dimensional (spatial plus temporal) motion vector. The value of PATH carries descriptive information of the vector's trajectory. In case of *winded* the vector path is $\sin_{[2,n]}$ and is of type *axis vector*.

4.3 Mode: Exemplification

In addition to the semantic modes *referential*, *predicational*, and *none* as used in the HPSG framework of [29], the mode *exemplification* is introduced. Semantic structures with this mode are appropriate for non-verbal means like gestures. Basically, they connect a non-verbal sign with a predicate, as shown in the constraint in (6):

(6)
$$\begin{bmatrix} \textit{sem-struc} & \\ \text{MODE} & \text{exemplification} \\ \text{EX-PRED} & \textit{pred(icate)} \end{bmatrix}$$

Exemplification becomes operative as a semantic mode if a gesture gets affiliated with verbal material (the affiliation account presented here assumes that affiliates correspond to syntactic constituents, an assumption that probably is too strong, because prosodic constituency, that is one affiliating source, may not coincide with "conventional" constituent structure [30]). The receiving structure for speech and gesture integration is of type *speech-gesture ensemble* (*sg-ensemble*) and is basically defined by the AVM in Figure 5 (slightly simplified from [8, p. 182], where the affiliated expression has to be phonetically marked, following [31].

There is a recursive type hierarchy for sinusoidal vector paths which is constructed out of all possible pairs of movement combinations from Figure 4 as primitive "sin"s. Recursive sinus curves allows to unify the predicate *winded* as speech daughter (S-DTR) with the gesture daughter (G-DTR) with trajectory $mr > mf > ml > mf > mr > mf$ (the morphological description of the dynamics of the example gesture from Figure 1) into a multimodal sign as licensed by (6).

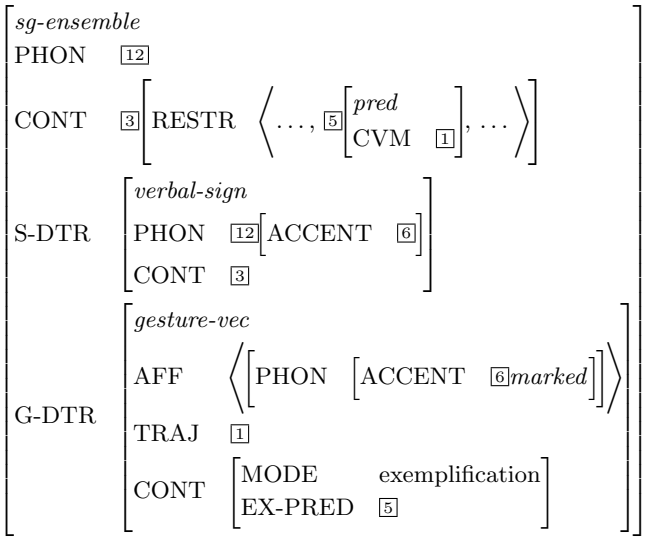


Fig. 5. Speech-gesture interface.

5 Extensions

More complex predicates like the German *zufahren* (*drive towards*), which selects a PP argument headed by the preposition *auf* (*(on)to*), are assumed to be connected with frames in their CONT(ent) feature structure. Thus, in addition to the "core" CVM, there are further predicates available for gestural exemplification, like, for instance, the form of the vehicle used for driving. The structure in Figure 6 is a representation of the *zufahren* construction, where the vectorial *towards* information (a vector sequence $\mathbf{W}$ which over the time course of the driving event approximates the location of goal j – written $\mathbf{W} \mapsto j$) is depicted by a simple forward-movement gesture.

The example from Figure 6 already illustrates three remaining challenges:

- A single gesture might exemplify more than one predicate. Consider, for instance, a gesture that not only performs the path of motion, but simultaneously models the moving vehicle. Currently, there is no means to capture such cases of *multi-exemplification*.
- The treatment of more complex examples like the one from Figure 6 requires unification into lists. Thus, an existentially quantification modification of the unification operation is needed.
- The PATH constraint of the CVM in Figure 6 expresses a rather abstract condition that can be fulfilled by a variety of gesture configurations. The matching conditions of such abstract constraints are called *filters* in [8]; they go beyond mere unification (and thereby the grammar framework) and are stipulated case-wise in terms of additional postulates. A more uniform solution is called for.

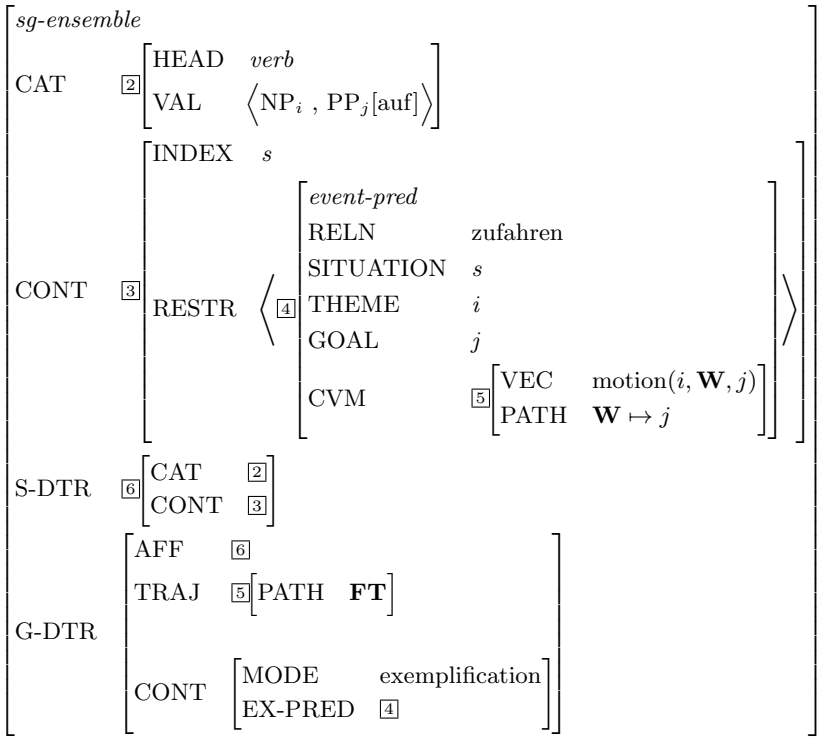


Fig. 6. *Drive towards* construction, affiliated with a gesture.

6 Conclusion

The grammar interface and its interdisciplinary background sketched in this article is intended as a foundational framework for the study of the interplay of verbal and non-verbal signs in multimodal communication, with an encore of speech-gesture integration. Being still a starting point, it provides a platform for theoretically-guided, systematic research on multimodality.

The difference between exemplification of CVMs of core predicates vs. the exemplification of CVMs from frame predicates provides a systematic, new perspective on semantic redundancy vs. complementary [32].

From the co-text sensitivity of gesture interpretation, which tries to capture the primacy of speech principle (see Section 2 above), follows the diachronic hypothesis, that the interpretation of co-verbal gesture co-changes with the meaning change of its affiliates. Seen from an intercultural perspective, this also relates to conceptual differences between apparently corresponding words from different languages [33] Thus, the account presented here can be made subject to empirical falsification.

Acknowledgement

Support by the German Research Foundation and the research project *Speech-gesture alignment* of the CRC *Alignment in Communication*, Bielefeld University, is gratefully acknowledged.

Bibliography

[1] Clark, H.H.: Using Language. Cambridge University Press, Cambridge (1996)

[2] Bavelas, J., Gerwing, J., Sutton, C., Prevost, D.: Gesturing on the telephone: Independent effects of dialogue and visibility. Journal of Memory and Language **58**(2) (2008) 495–520

[3] Iverson, J., Goldin-Meadow, S.: Why people gesture when they speak. Nature **396** (1998) 228

[4] Kendon, A.: How gestures can become like words. In Poyatos, F., ed.: Cross-cultural Perspectives in Non-Verbal Communication. Hogrefe, Toronto (1988) 131–141

[5] Kendon, A.: Gesticulation and speech: Two aspects of the process of utterance. In Key, M.R., ed.: The Relationship of Verbal and Nonverbal Communication. Volume 25 of Contributions to the Sociology of Language. Mouton Publishers, The Hague (1980) 207–227

[6] McNeill, D.: Hand and Mind—What Gestures Reveal about Thought. Chicago University Press, Chicago (1992)

[7] Kranstedt, A., Kopp, S., Wachsmuth, I.: MURML: A multimodal utterance representation markup language for conversational agents. In: Proceedings of the AAMAS02 Workshop on Embodied Conversational Agents – let's specify and evaluate them, Bologna, Italy (2002)

[8] Lücking, A.: Prolegoma zu einer Theorie ikonischer Gesten. PhD thesis, Universität Bielefeld (2011)

[9] Loehr, D.: Apects of rhythm in gesture in speech. Gesture **7**(2) (2007) 179–214

[10] Rieser, H.: Aligned iconic gesture in different strata of mm route-description. In: LonDial 2008: The 12th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL), King's College London (2008) 167–174

[11] Alahverdzhieva, K., Lascarides, A.: Analysing language and co-verbal gesture in constraint-based grammars. In Müller, S., ed.: Proceedings of the 17th International Conference on Head-Driven Phase Structure Grammar (HPSG), Paris (2010) 5–25

[12] Peirce, C.S.: On a new list of categories. In: Proceedings of the American Academy of Arts and Sciences Series. Volume 7. (1867) 287–298

[13] Bierman, A.K.: That there are no iconic signs. Philosophy and Phenomenological Research **23**(2) (1962) 243–249

[14] Goodman, N.: Languages of Art. An Approach to a Theory of Symbols. 2 edn. Hackett Publishing Company, Inc., Idianapolis (1976)

[15] Burks, A.W.: Icon, index, and symbol. Philosophy and Phenomenological Research **9**(4) (1949) 673–689

[16] Eco, U.: A Theory of Semiotics. Indiana University Press, Bloomington (1976)

[17] Giorgolo, G.: A formal semantics for iconic spatial gestures. In Aloni, M., Bastiaanse, H., de Jager, T., Schulz, K., eds.: Logic, Language and Meaning. Volume 6042 of Lecture Notes in Computer Science. Springer, Berlin and Heidelberg (2010) 305–314

[18] Sonesson, G.: The ecological foundations of iconicity. In Rauch, I., Carr, G.F., eds.: Semiotics Around the World: Synthesis in Diversity. Proceedings of the Fifth International Congress of the IASS, Berlin and New York, Berkeley, 1994, de Gruyter (1997) 739–742

[19] Lücking, A., Bergmann, K., Hahn, F., Kopp, S., Rieser, H.: The Bielefeld speech and gesture alignment corpus (SaGA). In: Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, Malta, 7th International Conference for Language Resources and Evaluation (LREC 2010) (2010) 92–98
[20] Lücking, A., Bergman, K., Hahn, F., Kopp, S., Rieser, H.: Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its applications. Journal of Multimodal User Interfaces (2012) accepted.
[21] de Ruiter, J.P.: On the primacy of language in multimodal communication. In: Proceedings of workshop on multimodal corpora, Lisbon (2004)
[22] Johansson, G.: Visual perception of biological motion and a model for its analysis. Perception Psychophys. **14** (1973) 201–211
[23] Johnston, M.: Unification-based multimodal parsing. In: Proceedings of the 36th Annual Meeting on Association for Computational Linguistics – Volume I, Montreal, Quebec, Canada, Annual Meeting of the ACL, Association for Computational Linguistics (1998) 624–630
[24] Johnston, M., Cohen, P.R., McGee, D., Oviatt, S.L., Pittman, J.A., Smith, I.: Unification-based multimodal integration. In: Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics, Madrid, Spain, European Chapter Meeting of the ACL, Association for Computational Linguistics (1997) 281–288
[25] Jackendoff, R.: Semantic Structures. Current studies in linguistics ; 18. MIT Press, Cambridge, MA (1991)
[26] Kratzer, A.: An investigation of the lumps of thought. Linguistics and Philosophy **12**(5) (1989) 607–653
[27] Zwarts, J.: Vectors as relative positions: A compositional semantics of modified pps. Journal of Semantics **14**(1) (1997) 57–86
[28] Zwarts, J., Verkuyl, H.: An algebra of conceptual structure; an investigation into jackendoff's conceptual semantics. Linguistics and Philosophy **17** (1994) 1–28
[29] Sag, I.A., Wasow, T., Bender, E.M.: Syntactic Theory: A Formal Introduction. 2 edn. CSLI Publications, Stanford (2003)
[30] Klein, E.: Prosodic constituency in HPSG. In Cann, R., Grover, C., Miller, P., eds.: Grammatical Interfaces in HPSG. Stanford University Press, Stanford (2000) 169–200
[31] Engdahl, E., Vallduví, E.: Information packaging in HPSG. In Engdahl, E., Vallduví, E., eds.: Edinburgh Working Papers in Cognitive Science. Volume 12 of Studies in HPSG. University of Edinburgh, Edinburgh (1996) 1–31
[32] Bergmann, K., Kopp, S.: Verbal or visual? how information is distributed across speech and gesture in spatial dialog. In Schlangen, D., Fernández, R., eds.: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue. Brandial'06, Potsdam, Universitätsverlag Potsdam (2006) 90–97
[33] Kita, S., Özyürek, A.: What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. Journal of Memory and Language **48**(1) (2003) 16–32