A spreading-activation model of the semantic coordination of speech and gesture

Stefan Kopp (skopp@TechFak.Uni-Bielefeld.DE)

Kirsten Bergmann (kirsten.bergmann@Uni-Bielefeld.DE)

Sebastian Kahl (skahl@TechFak.Uni-Bielefeld.DE)

Faculty of Technology, Center of Excellence "Cognitive Interaction Technology" (CITEC),
Collaborative Research Center "Alignment in Communication" (SFB 673)
Bielefeld University, P.O. Box 100 131, D-33501 Bielefeld, Germany

Abstract

In naturally occurring speech and gesture, meaning occurs organized and distributed across the modalities in different ways. The underlying cognitive processes are largely unexplored. We propose a model based on activation spreading within dynamically shaped multimodal memories, in which coordination arises from the interplay of visuo-spatial and linguistically shaped representations under given communicative and cognitive resources. An implementation of this model is presented and first simulation results are reported.

Keywords: Speech, gesture, conceptualization, semantic coordination, activation spreading

Introduction

Gestures are an integral part of human communication and they are inseparably intertwined with speech (McNeill & Duncan, 2000). The detailed nature of this connection, however, is still a matter of considerable debate. The data that underlie this debate have for the most part come from studies on the coordination of overt speech and gestures showing that the two modalities are coordinated in their *temporal* arrangement and in *meaning*, but with considerable variations. When occurring in temporal proximity, the two modalities express the same underlying idea, however, not necessarily identical aspects of it: Iconic gestures can be found to be *redundant* with the information encoded verbally (e.g., 'round cake' + gesture depicting a round shape), to *supplement* it (e.g., 'cake' + gesture depicting a round shape), or even to *complement* it (e.g., 'looks like this' + gesture depicting a round shape). These variations in meaning coordination–in combination with temporal synchrony–led to different hypotheses about how the two modalities encode aspects of meaning and what mutual influences between the two modalities could underlie this. However, a concrete picture of this and in particular of the underlying cognitive processes is still missing.

In previous work (Bergmann & Kopp, 2009) we explored how the surface form of speech and gesture is determined and how this *formulation* process can be simulated in a computational model. In this paper we turn to the preceding stage, namely, *conceptualization* by which meaning is structured, portioned and distributed across the two modalities, yielding different kinds of semantic coordination one can see in real-life natural behavior. We thereby focus on speech along with shape-depicting (iconic) gestures. We start with reviewing the empirical findings on semantic coordination of speech and gesture, and we discuss mechanisms and models that have been put forward to explain it. We argue that building computational models helps to elucidate the mechanisms and to bridge the gap between descriptive models and observable behavior. We propose the first model to present a detailed cognitive account of how meaning can be organized and coordinated in speech and gesture. It is based on tenets of activation spreading in multimodal memory representations and it entails a number of, now explorable, assumptions about conceptualization of speech and gesture. We describe an implementation of this model and present first results on how it can simulate and explain different cases of semantic coordination reported in the literature.

Background

Semantic coordination of speech and gesture

A number of studies have shown that concomittant speech and gesture are coordinated in meaning. One line of evidence coming from cross-linguistic studies suggest that packaging of content for co-speech gestures is influenced by the information packaging for the accompanying speech. For example, Kita and Özyürek (2003) showed that speakers of English who are able to combine manner and path of a movement in a single clause (e.g. 'he rolled down' or 'he swings') accompanied this by a single gesture encoding both semantic features. In contrast, Turkish and Japanese speakers encoded manner and path separately in two clauses (e.g. 'he descended as he rolled') and are more likely to use two separate gestures for these two features. Along the same line, when native speakers of Turkish (L1) speak English as their second language (L2) at different levels of proficiency, their gestures were shown to follow the information packaging strategy they adopt (Özyürek, 2002): Advanced L2 speakers typically encoded manner and path information in one clause and their gestures followed, where as speakers at lower proficiency levels typically used two-clause constructions in speech thus following the structure of Turkish, accompanied by separate gestures for manner and path. A subsequent study (Kita et al., 2007) showed that this effect also occurrs when native speakers of English are forced to produce one- or two-clause descriptions of manner and path.

Other studies have investigated the cognitive factors that influence frequency and nature of gesturing, including its coordination with speech. Bavelas, Kenwood, Johnson, and Philips (2002) found that speakers are more likely to produce non-redundant gestures when their addressees could see them, as opposed to when their gestures are not visible and hence less essential for their partners. Bergmann and Kopp (2006) report results from an analysis of natural gesturing in direction-giving, indicating that supplementary iconic gestures are more likely in cases of problems of speech production (e.g. disfluencies) or when the information conveyed is introduced into the dialogue (and thus conceptualized for the first time). In line with this, recent work has suggested that speakers indeed produce more gestures at moments of relatively high load on the conceptualization process for speaking (Kita & Davies, 2009), in particular on the linearization and the focusing components of conceptualization (Melinger & Kita, 2007). Hostetter and Alibali (2007) report findings suggesting that speakers who have stronger visual-spatial skills than verbal skills produce higher rates of depictive gestures than other speakers. In a later study, Hostetter and Alibali (2011) found that the speakers with high spatial skills also produced a higher proportion of non-redundant gesture-speech combinations than other speakers, whereas verbal-dominant speakers tended to produce such gestures more in case of speech disfluencies. The authors hypothesize that "*non-redundant gesture-speech combinations occur because mental images are more active in speakers minds at the moment of speaking than are verbal codes*" [p.45]. Taken together, this suggests that non-redundant gesture-speech combinations are the result of speakers having both strong spatial knowledge and weak verbal knowledge simultaneously, and avoiding the effort of transforming the one into the other.

Models of speech and gesture production

Different models of speech and gesture production have been proposed. One distinguishing feature is the point where cross-modal coordination can take place. The Growth Point Theory (McNeill & Duncan, 2000) assumes that gestures arise from idea units combining imagery and categorial content. This combination is unstable and initiates dynamic cognitive events through which speech and gesture unfold. Speech and gesture, in this view, are inseparable and interact throughout the production process.

Assuming that gestures are generated "pre-linguistically", Krauss, Chen, and Gottesman (2000) hold that gesture are generated from a mental representation of a *source concept* comprising a set of semantic features (size, color, shape etc.) that are encoded in propositional and/or spatial format. While there is no influence of language production onto gesture in this model, the readily planned and executed gesture facilitates lexical retrieval through cross-modal priming.

De Ruiter (2000) proposed that speech-gesture coordination arises from a multimodal conceptualization process that selects the information to be expressed in each modality and assigns a perspective for the expression. A propositional representation is transformed into a preverbal message, and an imagistic representation is transformed into a so-called sketch and sent to a gesture planner. Kita and Özyürek (2003) agree that gesture and speech are two separate systems interacting during the conceptualization stage. Based on cross-linguistic evidence, their account holds that language shapes iconic gestures such that the content of a gesture is determined by three factors: (1) a communicative intention, (2) action schemata selected on the basis of features of imagined or real space, (3) bidirectional interactions between speech and gesture production processes at the level of conceptualization, i.e. the organization of meaning. An additional link between the speech formulator and the preverbal message generator allows for feedback from grammatical or phonological encoding to the conceptualizer and thus to gesture.

Hostetter and Alibali (2008) proposed the Gestures as Simulated Action framework that emphasizes how gestures may arise from an interplay of mental imagery, embodied simulations, and language production. According to this view, language production evokes enactive mental representations which give rise to motor activation. Whether a gesture is produced or not depends on the amount of motor activation, the speaker's variable gesture threshold, and the simultaneous engagement of the motor system for speaking.

Inspite of a consistent theoretical picture starting to emerge, many questions about the detailed mechanisms remain open. A promising approach to explicate and test hypotheses are cognitive models that allow for computational simulation. However, such modeling attempts for the production of speech and gestures are almost inexistent. Only Breslow, Harrison, and Trafton (2010) proposed an integrated production model based on the cognitive architecture ACT-R (Anderson, Bothell, Byrne, Lebiere, & Qin, 2004). This account draws on two major assumptions: (1) on Jackendoff's claim that language representations include some irreducibly spatial components; (2) on Goldberg's approach according to which language processing is based on constructions which consist of both semantic and syntactic components. The authors assume such constructions to prescribe spatial representations for what they call *linguistic spatial gestures* and which they assume to provide "*little information not included in the accompanying language*" [p.14]. In this view, constructions are selected first and then words and gestures are determined so as to realize the construction. Accordingly, semantic coordination is predetermined and does not result from a coordination process based on problems with lexicalization or high activation of particular visuo-spatial features. This model hence has difficulties, e.g., to explain gestures that clearly complement or supplement verbally encoded meaning.

A spreading-activation model

We investigate to what extent semantic coordination of speech and gesture can be explained by cognitive principles of activation-based processing on multimodal memory. This account is embedded in a larger production model (Kopp,

Bergmann, & Wachsmuth, 2008) that comprises three stages: conceptualization, where a *message generator* and an *image generator* work together to select and organize information to be encoded in speech and gesture, respectively; formulation, where a *speech formulator* and a *gesture formulator* determine appropriate verbal and gestural forms for this; *motor control* and *articulation* to finally execute the behaviors. Motor control, articulation, and formulation have been subject of earlier work (Bergmann & Kopp, 2009). What is missing is a model for multimodal conceptualization that accounts for the range of semantic coordination we see in real-life speech-gesture combinations.

Basic assumptions

We posit that the semantic coordination of speech and gesture emerges from (1) the communicative goal, (2) the need to activate, retrieve and organize multimodal information to achieve this goal, and (3) the expressive as well as cognitive resources available to the speaker at the moment. To model this process, we make a number of assumptions, partly in line with previous models. First, language production requires a preverbal message to be formulated in a symbolic-propositional representation that is linguistically shaped (Slobin, 1996; Levelt, 1989) (SPR, henceforth). During conceptualization the SPR, e.g. a function-argument structure denoting a spatial property of an object, often needs to be extracted from visuo-spatial representations (VSR), e.g. the mental image of this object. We assume this process to involve the invokation and instantiation of memorized supramodal concepts (SMC, henceforth), e.g. the concept 'round' which links the corresponding visuo-spatial properties to a corresponding propositional denotation. Co-verbal iconic gestures are then shaped by (1) the imagistic content in VSR, (2) the invoked SMCs, and (3) the organization of SPR for linguistic processing. We assume that units or entries of these memory structures can be selectively activated and that activation spreads along links between them. Fig. 1 illustrates the overall relation between the three memory structures.

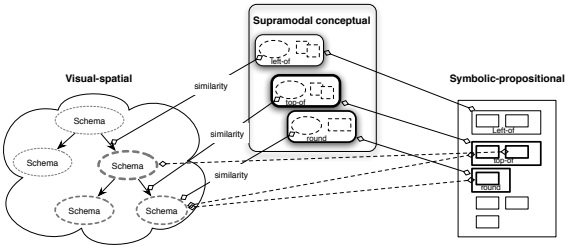

Figure 1: Multimodal memory structures involed in speech-gesture production (activations indicated by bold lines).

Overall production process

Fig. 2 shows an outline of the overall production architecture. Conceptualization consists of cognitive processes that operate upon the abovementioned memory structures to create a, more or less coherent, multimodal message. These processes are constrained by principles of memory retrieval, which we assume can be modeled by principles of activation spreading (Collins & Loftus, 1975). As in cognitive architectures like ACT-R (Anderson et al., 2004), activations float dynamically, spread across linked entities (in particular via SMCs), and decay over time. Activation of more complex SMCs are assumed to decay more slowly than activation in lower VSR or SPR.

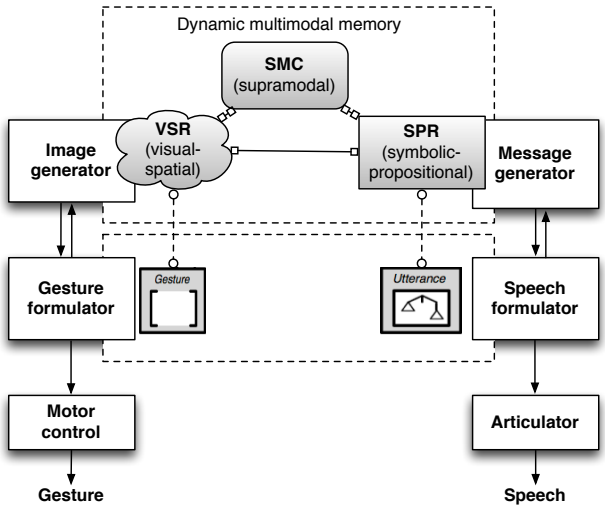

Figure 2: Overall production architecture.

Production starts with the *message generator* and *image generator* inducing local activations of modal entries, evoked by a communicative goal. VSRs that are sufficiently activated invoke matching SMCs, leading to an instantiation of SPRs representing the corresponding visuo-spatial knowledge in linguistically shaped ways. The *generators* independently select modal entries and pass them on to the *formulators*. As in ACT-R, highly activated features or concepts are more likely to be retrieved and thus to be encoded. Note that, as activation is dynamic, feature selection depends on the time of retrieval and thus available resources. The *message generator* has to map activated concepts in SPR onto grammatically determined categorical structures, anticipating what the *speech formulator* is able to process (cf. (Levelt, 1989)). Importantly, interaction between *generators* and *formulators* in **each** modality can run top-down **and** bottom-up. For example, a proposition being encoded by the *speech formulator* results in reinforced activation of the concept in SPR, and thus increased activation of associated concepts in VSR.

In result, semantic coordination emerges from the local choices generators and formulators take, based on the activation dynamics in multimodally linked memory representations. Redundant speech and gesture result from focused activation of supramodally linked mental representations, whereas non-redundant speech and gesture arise when activations scatter over entries not connected via SMCs.

Computational simulation

We have implemented the activation-based model of semantic coordination within our larger speech and gesture production architecture (Bergmann & Kopp, 2009). Newly implemented parts are the VSR, SPR and SMC memory structures, the activation dynamics upon these structures, and the generator modules operating on them.

Representations

To realize the VSR and part of the SMC, we employ a model of visuo-spatial imagery called *Imagistic Description Trees* (IDT) (Sowa & Kopp, 2003). The IDT model was designed, based on empirical data, to cover the meaningful visuo-spatial features in shape-depicting iconic gestures. Important aspects include (1) a tree structure for shape decomposition with abstract object schemas as nodes, (2) extents in different dimensions as an approximation of shape, and (3) the possibility of dimensional information to be underspecified. The latter occurs, e.g., when the axes of an object schema cover less than the three dimensions of space or when an exact dimensional extent is left open but only a coarse relation between axes like "dominates" is given. This allows to represent the visuo-spatial properties of SMCs such as 'round', 'left-of' or 'longish'. Applying SMC to VSR is realized through graph unification and similarity matching between object schemas, yielding similarity values that assess how well a certain SMC applies to a particular visuo-spatially represented entity (cf. Fig. 1). SPR are implemented straight forward as predicate-argument sentences.

Activation dynamics

Each memory entry in VSR, SPR and SMC has a time-dependent activation value a_t. Activation dynamics results from simple update and spreading rules applied to these values in each iteration of a stepwise cognitive simulation process. At each step all of the following updates are performed:

- Activation update for memory entries: $a_{t+1} = a_t - d + r$, with decay d, random noise r (order of magnitude 10^{-1})
- Activation spreading within VSR: $a_{t+1} = \frac{a_t}{c \cdot l}$, where c is the number of outgoing links (fan-out effect) and l is the depth in the hierarchical IDT structure (fade-out effect)
- Activation spreading from SPR towards VSR via SMC: $a^{vsr}_{t+1} = \frac{a^{vsr}_t + a^{spr}_t}{2} + \alpha \cdot (a^{smc}_t - a^{vsr}_t) + r - d$, where α controls the rate of convergence towards the SMC activation.
- Activation spreading from VSR towards SPR via SMC: $a^{spr}_{t+1} = \frac{a^{vsr}_t + a^{spr}_t}{2} + \alpha \cdot (a^{smc}_t - a^{spr}_t) + r - d$

The first formula models the decay and random noise of each entry's activation, the second realizes local spreading of activation within VSR, the latter two at a global level between VSR and SPR. Especially the global multimodal activation spreading is important as it ensures that linked visuo-spatial and propositional codes align and mutually stabilize. Fig. 3 (left) shows the activations of two linked entries. At point $t = 200$ one entry gets temporarily activated and the activation of the linked entry follows. The second important property of this rule is that activation of the more global SMC a^{smc}_t spreads to both linked entries, such that both are "pulled" towards this value. This can be seen in Fig. 3 (right) where the SMC's activation is increased by 2.0 at point $t = 100$. Note that activation of SMCs decays **more slowly** than activation of VSR and SPR entries. Activations of linked entries thus stabilizes at a higher level, such that stable multimodal information packages emerge for a limited period of time. The duration of this time period depends on the decay rate of SMC activations. Finally, memory retrieval depends on the activation of the entries being retrieved. We adopt the ACT-R approach to map activation onto retrieval probability: $p = 1/(1 + e^{\frac{-(a_t - s)}{r}})$, where s is a threshold and r the noise in the activation levels.

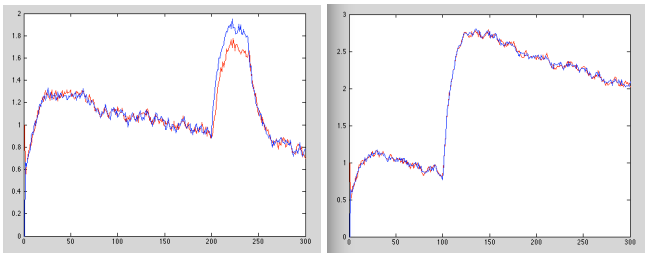

Figure 3: Activations of two memory entry linked via an SMC: temporary activation of one entry (left); activation of the linking SMC (right).

Generators

The *message generator* has to package activated SPR information in a way that the *speech formulator* can produce an appropriate construction. We employ an LTAG-based (Lexicalized Tree Adjoining Grammar) sentence planner for speech formulation (cf. (Bergmann & Kopp, 2009)). To make sure that all facts necessary to generate a specific construction are available, the *message generator* applies networks that reflect the encoding options provided by the *speech formulator*'s LTAG grammar (this conforms the view that the conceptualizer learns to anticipate the formulator's abilities (Levelt, 1989)). These message networks consist of *type nodes* for entities, *properties* of entities and *relations* between them. These are connected via weighted links reflecting the combination of particular linguistic types in a language. For instance, *relation nodes* are strongly linked to two (or more) *entity nodes*, while links between entity and property nodes are weaker.The message generator matches the activated propositions in SPR against nodes of possible message networks and determines their initial activations. Activation, again, spreads via the weighted links and finally results in an overall activation pattern of a pre-verbal message. This has been implemented for a limited part of our domain of investigation (corresponding to NPs about buildings and their properties).

The *image generator* retrieves visuo-spatial information from activated VSR and SMC entries in memory. It is in charge of unifying this information into an imagistic representation, from which the *gesture formulator* can derive a gesture form specification (based on Bayesian decision networks learned from empirical data (Bergmann & Kopp, 2009)). For instance, information about shape is combined with information about the object's size or position. Depending on the knowledge encoded here, the *gesture formulator* is able to plan a shape-depicting gesture or rather a localizing deictic or placing gesture.

Simulation results

The implemented model offers–and simulates–detailed explanations of how semantic coordination between speech and gesture arises (see next Section). In particular, it allows us to manipulate the interaction between modality-specific production processes. As a first exploration, we report results on how processing time as a cognitive resource affects the observable meaning coordination.

The production process is initiated by setting the communicative intention "introduce churchwindow-1". Upon receiving this goal, the *image generator* activates visuo-spatial imagery of the church window in VSR, and the *message generator* activates symbolic representations of non-spatial semantic concepts in SPR. These activations spread through memory and lead to invokation of SMCs for, e.g., 'round' (bound to churchwindow-1) and 'at-top-of' (the church-tower), as well as instantiation of the corresponding SPR entries. SMCs along with their linked entries in VSR and SPR attain highest and most slowly decaying activation values.

After a preset number of processing cycles, both generators retrieve modality-specific information from memory with a probability depending on current activation values, leading to 'round' and 'at-top-of' concepts being encoded in speech and gesture in a less coordinated way: the *message generator* may retrieve only information about the salient shape of the window, but not about its position relative to other entities. Accordingly, a sentence like "The church has a round window" gets formulated. The *image generator*, on the other hand, may receive information about the entity's position as well. This can result in shape depicting gestures, like drawing the shape of the window in the air, or a static posturing gesture where the hands becoming a model of the circular shape. As the position of the entity is also available, the gesture would be performed in that part of gesture space. So, the gesture would be non-redundant to speech, supplementing it with the position of the entity.

If more time is available, however, the contents expressed either verbally or gesturally tend to converge. The *message generator* will start to (re-)activate those entries being retrieved and selected by the *speech formulator*. This results in multimodal representations being better coordinated when the modality-specific formulators start with their generation work, as it is more likely that both generators receive the same information about shape and position of the entity. Accordingly, the *speech formulator* is now enabled to plan a sentence like "The church has a round window at the top" which–like the gesture(s) described previously–encodes both, shape information and the entity's relative position.

To quantify these observations, we ran a simulation experiment in which we manipulated the available time (in terms of memory update cycles) before the model had to come up with a sentence and a gesture. We analyzed the resulting sentences and gestures for semantic redundancy/non-redundancy. We defined two conditions: A time-constrained condition with a certain number *N* of cycles and a condition with twice as many cycles. We ran the model 100 times in each condition. Fig. 4 shows that non-redundant (supplementary) gestures dominate in those runs with stricter temporal limitations, while redundant ones become more likely when time available is increased. Notably, the information conveyed by gesture was similar in both conditions. So, the higher redundancy in the less time-constrained condition is mostly due to the fact that the verbal utterances were richer in content.

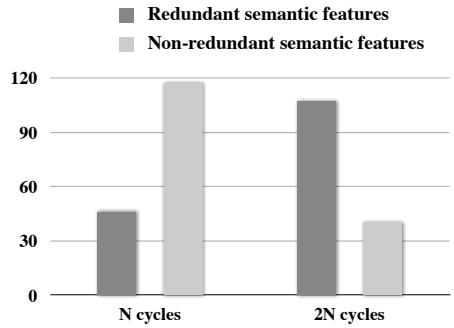

Figure 4: Number of semantic gesture features encoded redundantly vs. non-redundantly with speech in 100 simulation runs in more (left) or less (right) time-constrained conditions (note that a gesture may carry more than one feature).

Discussion and conclusions

We have presented the first model to explain semantic coordination between speech and gesture in terms of (1) how visuo-spatial and symbolic-propositional memory entries are dynamically linked, (2) how activation spreads in these concept structures, and (3) how this interacts with modality-specific processes of conceptualization and formulation. We believe that this model offer mechanisms and thus possible explanations for many empirical findings and hypotheses put forth in literature: The hypothesis that gestures are more likely if activation in visuo-spatial memory is higher, is directly explained by the activation-based retrieval probabilities when the image generator accesses memory; the hypothesis that non-redundant gestures are more likely when spatial codes are not transformed into verbal codes is accounted for by entries in VSR and SPR not being linked via SMC, leading to less coordinated conceptual structures and activations. Fi-

nally, the shaping of gesture by speech is accounted for, first, through SPR and SMC schematizing VSR in linguistically shaped ways and, second, through choices in linguistic formulation reinforcing activations in SPR and thus VSR.

Our simulation study showed that the model also offers a natural account for the finding that non-redundant gesture are more likely when conceptualization load is high, based on the assumption that memory-based cross-modal coordination consumes resources (memory, time) and is reduced or compromised when, e.g., time is limited. This examplifies how a model like ours can help to make hypothesis testable by giving rise to predictions that can be explored in computational simulations as well as in appropriately set up empirical experiments. While the model presented here mainly accounts for information distribution, work is underway to extend the model to account also for different ways to package information over multiple clauses, e.g., depending on available linguistic or gestural resources. This will enable to simulate cross-linguistic differences in co-speech gesturing. Another issue for future work will be to go beyond object-related gestures accompanying NP constructions, and to address descriptions of action events with a more complex internal structure and thus a more demanding semantic coordination to be achieved by the cognitive processes involved in multimodal conceptualization.

Acknowledgements

This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 "Alignment in Communication" and the Center of Excellence 277 "Cognitive Interaction Technology" (CITEC).

References

Anderson, J., Bothell, D., Byrne, M., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060.

Bavelas, J., Kenwood, C., Johnson, T., & Philips, B. (2002). An experimental study of when and how speakers use gestures to communicate. *Gesture*, *2*(1), 1–17.

Bergmann, K., & Kopp, S. (2006). Verbal or visual: How information is distributed across speech and gesture in spatial dialog. In D. Schlangen & R. Fernandez (Eds.), *Proceedings of the 10th workshop on the semantics and pragmatics of dialogue* (pp. 90–97).

Bergmann, K., & Kopp, S. (2009). Increasing expressiveness for virtual agents - autonomous generation of speech and gesture for spatial description tasks. In *Proc. of AAMAS 2009* (p. 361-368).

Breslow, L., Harrison, A., & Trafton, J. (2010). Linguistic spatial gestures. In D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the 10th international conference on cognitive modeling* (pp. 13–18).

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407-428.

Hostetter, A., & Alibali, M. (2007). Raise your hand if you're spatial—relations between verbal and spatial skills and gesture production. *Gesture*, *7*, 73–95.

Hostetter, A., & Alibali, M. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin and Review*, *15/3*, 495–514.

Hostetter, A., & Alibali, M. (2011). Cognitive skills and gesture-speech redundancy. *Gesture*, *11*(1), 40–60.

Kita, S., & Davies, T. S. (2009). Competing conceptual representations trigger co-speech representational gestures. *Language and Cognitive Processes*, *24*(5), 761-775.

Kita, S., & Özyürek, A. (2003). 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*, 16–32.

Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. *Language and Cognitive Processes*, *22*, 1212–1236.

Kopp, S., Bergmann, K., & Wachsmuth, I. (2008). Multimodal communication from multimodal thinking - towards an integrated model of speech and gesture production. *Semantic Computing*, *2*(1), 115-136.

Krauss, R., Chen, Y., & Gottesman, R. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), *Language and gesture* (pp. 261–283). Cambridge, UK: Cambridge University Press.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.

McNeill, D., & Duncan, S. (2000). Growth points in thinking-for-speaking. In D. McNeill (Ed.), *Language and gesture* (pp. 141–161). Cambridge, UK: Cambridge University Press.

Melinger, A., & Kita, S. (2007). Conceptualisation load triggers gesture production. *Language and Cognitive Processes*, *22*(4), 473-500.

Özyürek, A. (2002). Speech-gesture relationship across languages and in second language learners: Implications for spatial thinking and speaking. In *Proceedings of the 26th annual boston university conference on language development* (pp. 500–509).

Ruiter, J. de. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture* (pp. 284–311). Cambridge, UK: Cambridge University Press.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking". In J. J. Gumperz & S. C. Levison (Eds.), *Rethinking linguistic relativity* (p. 70-96). Cambridge Univ. Press.

Sowa, T., & Kopp, S. (2003). A cognitive model for the representation and processing of shape-related gestures. In *Proc. European Cognitive Science Conference.*