The Bielefeld Jigsaw Map Game (JMG) Corpus

Andy Lücking
Goethe-Univ. Frankfurt a. M.

Peter Menke
Univ. Bielefeld

Olga Abramov
Univ. Bielefeld

Alexander Mehler
Goethe-Univ. Frankfurt a. M.

October 9, 2011

Spoken language still poses a challenge to mechanisms developed for information processing and retrieval. Applications in this area often require a large amount of annotated data, which is hardly obtainable for spoken language. We present a corpus of 78 semi-natural dialogues (length: ≈20 h) completely transcribed and annotated on various linguistic levels. The dialogues stem from a psycholinguistic, task-oriented coordination game, the Jigsaw Map Game (JMG) (Weiß, Pfeiffer, Schaffranietz, and Rickheit 2008). In order to solve the task of the JMG, the speakers produced spontaneous utterances about objects from a predefined object set, that have to be located according to a map. What makes this data special is the combination of fixed utterance topics and nonetheless unconstrained language use. Primarily developed and annotated to study *alignment in communication* (Pickering and Garrod 2004), the corpus represents a useful resource for natural language processing and studies on spoken language. We describe 1. the theoretical motivation of the JMG, 2. the experimental setting used for data gathering, 3. the levels of annotation and its correction, as well as 4. the applications of the JMG corpus in the area of lexical alignment.

1 Introduction

Garrod and Pickering (2004) emphasise the role of dialogue for understanding the "easiness" of human communication compared to the complexity of information transferred. What helps to communicate according to (Clark 1996) is the information shared by the speakers, that is, their *common ground*. Pickering and Garrod (2004) distinguish between *explicit* common ground (achieved through negotiation), and *implicit* common ground.

They argue that implicit common ground between dialogue partners is established almost *mechanistically* by means of *alignment*. That is, people, when communicating, automatically build up an implicit common ground *via* mutually *aligning* their mental representations on all linguistic levels. The basic mechanism that implements alignment is priming, which predicts that the probability to produce a linguistic element (where the elements may range from phonetic, articulatory features to whole utterances) increases if the speaker herself (intra-personal alignment *via* monitoring) or the addressee (inter-personal alignment) produced this element before. The alignment theory, though debated (cf. for instance (Schiller and de Ruiter 2004)), allows an understanding of coordinated language use beyond negotiation.

The Bielefeld CRC 673 "Alignment in Communication" addresses this topic from an interdisciplinary perspective. Linguists, psycholinguists and computer scientists complement each other in studying alignment within the collaborative research project. Psycholinguistic experiments designed to study alignment are recorded, transcribed and annotated with relevant linguistic information in order to enable the computational machinery to analyse the data quantitatively. In the following, we describe the Jigsaw Map Game corpus of dialogues that was systematically annotated on various linguistic levels. In Section 2 we describe the experimental design which has been used to elicit the dialogues. In Section 3 we describe the levels of annotation and their evaluation. Section 4 describes the quantitative characteristics of the JMG corpus. In Section 5 we present some successful applications of the corpus in the area of alignment research. The conclusions are drawn in Section 6.

2 Experimental Design

One of the empirical challenges when studying alignment in natural dialogues is the openness of dialogical communication with respect to topic selections and language use, *inter alia*. A way to overcome this problem is to introduce semi-controlled experimental dialogues that restrict the topic space of the dialogue to some extent (Garrod 1999; Schober and Brennan 2003). In the JMG setting this has to be accomplished in particular in order to fix a reference point for measuring alignment on the lexical level. The three main experimental approaches proposed in the literature are: the *referential communication task* (Krauss and Weinheimer 1966), the *maze game* (Garrod and Anderson 1987) and the *map task* (Anderson et al. 1991). Each of those experimental paradigms has a certain principal limitation compared to "real life" face-to-face communication. The referential communication task allows to examine referential processes, however the roles of the communication partners remain fixed. Maze games allow to study the impact of agreement on the success of the game in relatively free verbal interactions. But the setting of the maze game with its spatial separation of interlocutors who sit alone in a room just facing monitors is far from natural. When aiming to examine the basic processes in natural face-to-face conversations none of these paradigms is sufficient on its own. Accordingly, Weiß, Pfeiffer, Schaffranietz, and Rickheit (2008) and Weiß, Pustylnikov, Mehler, and Hellmann (2009) developed an experimental design, namely the

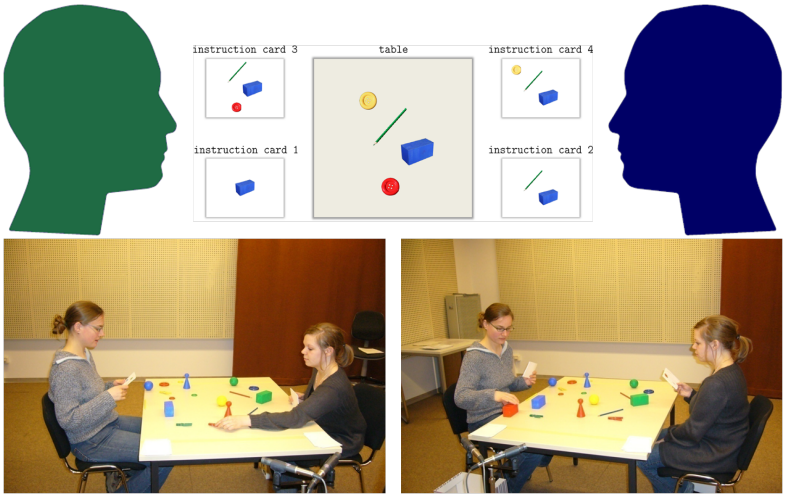


Figure 1: Schematic overview of the JMG (taken from Mehler, Lücking, and Weiß (2010)): agent A (left) instructs agent B (right) according to the card 1. Then, agent B takes the card 2 and instructs agent A, and so on until there are no more cards, and the map is complete. The lower pictures show in imprint from a JMG trial: A instructs B to place a red *clothespin* (left), B instructs A to place a red *cuboid*.

Jigsaw Map Game (JMG) that enabled to naturalise experimental dialogues by encouraging face-to-face interaction with naïve participants mutually perceiving their behaviour while communicating in a multimodal way. At the same time, important dialogue parameters, such as communicative goals, task-relevant knowledge, dialogue organisation (e.g. turn-taking) as well as group-related parameters of communicative situations are controlled, since the JMG basically defines a task-oriented setting. In short, the JMG setting is designed in order to account for the trade-off between natural communication and experimental control (Schober and Brennan 2003). The main features of the JMG paradigm can be summarised as follows:

- semi-spontaneous dialogues
- iterative role switching (role rotative changes between instructor and follower)
- face-to-face interactions
- object localisation (similar to map task)
- referential communication (object identification and placement)

The participants of a JMG dialogue are sitting facing each other at a table that serves as the interaction space for the construction of the target map. The participants' task is to cooperatively position objects like cuboids, cones or buttons on the table according to a predefined object arrangement, the target map (see Figure 1). The

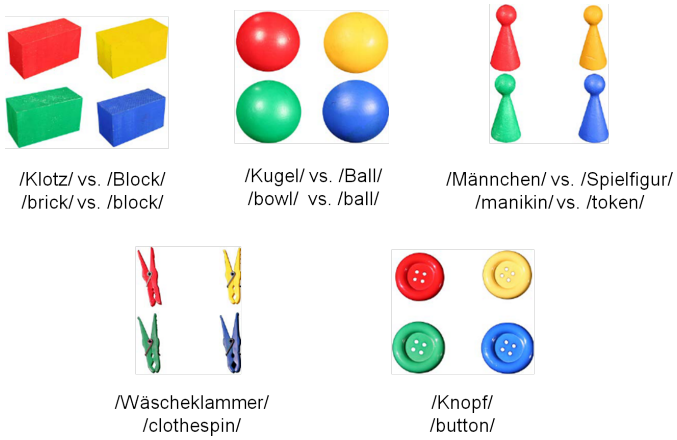


Figure 2: The upper row lists the critical objects with two alternative names. The lower row shows two examples of uncritical objects with a single denotation in German (the Figure is taken from (Mehler, Lücking, and Weiß 2010)).

complete arrangement is not known to the participants at the beginning. Interaction starts with the participant in the role of the instructor. He takes a card showing a part of the arrangement with exactly three objects: one new object (*focus object*) and two objects already present on the table (the very first two cards are an exception to this rule, since they clearly show just one, resp. two objects). Looking at his card, the instructor has to explain verbally to his partner which is the focus object. The partner then has to take that object from her basket (*object identification*), and to locate it on the map relative to the objects already laid out (*object location*). There are three special objects, which were selected because in German there are (at least) two possible names for them. These objects are called *critical objects* (see Figure 2). For instance, a spherical object might be referred to as *Ball* (ball) or *Kugel* (bowl). That both names of critical objects are usual and equally good means of referring to those objects has been verified in a pre-study (Schaffranietz, Weiß, Pfeiffer, and Rickheit 2007).

The variation of the critical object names is achieved based on the confederate priming paradigm (Branigan, Pickering, and Cleland 2000). The first participant plays the game with a confederate (confederate dialogue) who was instructed (depending on the experimental condition) to use only one of the names for each critical object so that these object names are primed in the current dialogue. After having played a JMG round with a confederate, this participant meets up with a second but this time naïve participant playing the game again (experimental dialogue). For data analysis regarding alignment on the lexical level, we can record how interlocutors name the objects and how this naming is taken up or not during the course of the dialogue.

The dialogues are videotaped and audio recorded. The recordings are used for annotating the data. Annotation is described in Section 3. In Section 4 quantitative characteristics of the JMG data are given in terms of descriptive statistics based on the annotation data.

3 Annotation

3.1 Transcription

Speech transcription comprises annotations of two tracks, a word track and an utterance track. The utterances produced by both the naïve participants and the confederates are transcribed at the level of words. The transcription procedure abstracts away from phonetic peculiarities of spoken dialogue. For instance, word blendings as typical for spoken language are separated and coded according to the standard German spelling rules. In addition to the written standard, however, the transcription rules acknowledge a set of particles and interjections that are quite regularly used in spoken interaction in order to, e.g., communicate acceptance or uncertainty. For instance, affirmation in spoken German is usually realised by "mhmh" (or variants thereof), a phonetic word that is not part of written German.

The words that manifest a sentence or a sub-sentential dialogue turn demarcate the boundaries for entries on the sentence annotation track.

3.2 Event Annotation

Each JMG dialogue is strictly divided into recurrent events that all have nearly the same three-part structure. An event corresponds to the act of completing the task of cooperatively placing a new object on the map (i.e. one instruction card). There are as many events as instruction cards (i.e. snippets of the total JMG map, see Fig. 1). We observe that events typically can be segmented into three functional phases that vary in duration from event to event, and from dialogue to dialogue. First, each event starts with the *identification* of the focus object. In this phase the participant that has the role of the instructor asks his partner to find the missing object. Then, a placing instruction (*object localisation*) follows. The third part of the event (*adjustment / confirmation*) usually contains clarification requests, adjustments and confirmations. The three parts of the event are annotated in a separate track. The following example[1] illustrates the three phases:

(1) A: So, äh, dann müsstest du einen roten Knopf haben, einen kleinen und äh, genau, [object identification] der müsste, ähm, von dir aus gesehen links etwas unterhalb der Spielfigur [object localisation].
(*Then, you should have a red button, a small one, exactly, [object identification] this one should be placed to the left from you, slightly below the token [object localisation].*)

B: So?
(*Like this?*)

A: Ja. Passt auch. [adjustment / confirmation].
(*Yes. It's ok. [adjustment / confirmation].*)

[1] The example is taken from the JMG dialogue `1_1` where it occurs at time `00:01:28.816`–`00:01:48.000`.

3.3 Lexical Operators

The lexical operators annotation corresponds to the morpho-syntactic and semantic analysis of the JMG data.[2] Since one of the goals of the project was to study lexical alignment that has to do with objects' naming, a more fine grained lexical annotation was performed. The reason is, that speakers often use many different terms to refer to the same objects. In order to nevertheless relate these terms to the objects used in the game, we annotated all nouns with the corresponding objects' reference ID's. Further, some terms are derivations, diminutives, compositions or specifications of the original term. There are also metaphors, hyponyms, ellipses etc. used to refer to the same object. In addition, not only words are used to refer to the objects, but also descriptions on the phrasal or sentential level. There are also repairs where a speaker first uses a wrong naming of the object, and then corrects herself (or is corrected by the partner).

Lexical operators (LexOps) are used to annotate the kind of information described above. LexOps are needed to resolve, for example, the diminutive or to reveal the meaning of a metaphor in its local JMG context. This is done by means of four attributes: *start-* and *endpoint* of the utterance referring to the object, the lexical (morphological, syntactic, semantic) *operation*, and its *value*. Annotating the stem form of the diminutive, for example, would result in the start- and endpoint of the word, the name of the operation (i.e. stemming), and its value (i.e. the stem form). The operators are organized in a sort of hierarchy that predefines the order in which they are applied (phrase ops > morphological ops > lexical ops > semantic ops). That is, if there are many operators for one utterance, morphological operators precede the semantic ones. For example, the German word *Kästchen* is combined of two operators: *morphology* : *de-derivation* and *semantics* : *substitution*. The annotation of the word *Kästchen* (little box) would look like as shown in the following pseudo-code example:

```
<start time="17:23:00.000" />
<end time="17:23:05.000" />
<operator Ddv(de-derivation)="Kasten" />
<operator Sub(substitution)="object 3" />
...
```

The Ddv-operator returns the stem form of *Kästchen* (i.e. *Kasten*), then, Sub indicates the semantics of the de-derived word *Kasten*. Altogether, there are 22 possible LexOps subordinated to the four categories: phrase, morphology, lexis and semantics. LexOps of each speaker are annotated in a separate track.

3.4 Data Evaluation

Since disagreements in the highly regulated data (the data are of Type I according to the qualitative data distinction of Gwet (2001)) can only occur as a result of a spelling error, we let correct the annotations of an annotator by an "annotation proof reader"

[2]Detailed descriptions and annotation guidelines are listed in the annotation manual (Mehler, Weiß, Abramov, and Hellmann 2010).

LEVEL	GLOBAL #	AVG # PER DIALOGUE
words	93,120	1,501.935
utterances	28,380	457.742
events	1,731	27.919
event phases	5,153	83.113
lexical operators	4,415	71.210
repairs	3,327	53.661

Table 1: Quantitative characteristics of the JMG Corpus, measured by the number of annotations on various levels.

(annotator and proof reader are different persons). Eventual mistakes are corrected in this process.

In addition to manual annotation, part of the automatic preprocessing of our data is part-of-speech (POS) tagging. Since POS tagging is error-prone, the output of the POS tagger has been controlled by a human annotator. Controlling has been carried out in three steps:

- The human annotator checks all word form tokens that have not been recognised by the POS tagger.
- Heuristic selection processes has been employed in order to generate sample sets of POS taggings to be reviewed by human annotator. A heuristics, for instance, that draws on German spelling conventions says to collect all those entries which has been tagged as a noun but are written with an initial lower letter.
- The human annotator randomly draws some tagging results from the complete tagging.

Data analyses has been carried out on the verified data.

4 Quantitative Characteristics

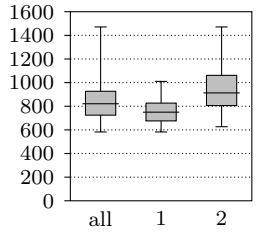
4.1 Corpus size

The Jigsaw Map Game Corpus consists of 64 dialogues, with a duration of approximately 20 hours, totalled up.

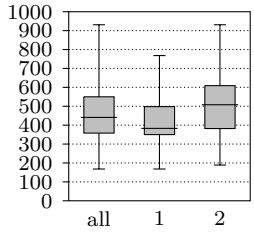
The overall number of annotations for words and utterances, event-related annotations, lexical operators and repairs can be found in Table 1. That table contains also average values for each of these dimensions.

We regularly observed significant differences between confederate dialogues and experimental dialogues (cf. the past paragraphs of section 2).

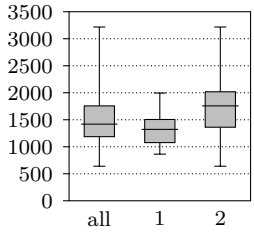
For the sake of brevity, during statistical analysis we referred to confederate dialogues as type 1 dialogues, and to experimental dialogues as type 2 dialogues, especially in figures and diagrams.



(a) length



(b) extension (utterances)



(c) extension (words)

SAMPLE	LENGTH	EXT. UTTERANCES	EXT. WORDS
all	821.170	441	1 419
1	750.000	383	1 322
2	912.958	508	1 757

(d) numeric data

Figure 3: Boxplots of the distribution of the three measures of dialogue size/extension, shown for three samples: **all** – complete corpus; **1** – for type 1 / confederate dialogues; **2** – for type 2 / experimental dialogues.

4.2 Dialogue size

We examined the size of a dialogue with three different measures: We measured

1. its duration in seconds (called its *length*),
2. its number of utterances (called its *utterance extension*),
3. and its number of words (called its *word extension*).

Dialogues have an average length of 13:41 minutes, with a standard deviation of 03:07 minutes. In the shortest dialogue, participants needed only 09:43 minutes to complete the game. Another pair of participants needed 24:31 minutes, this being the longest dialogue of the corpus.

If measured by the extension in utterances or words, we observed an average size of 441 utterances and 1,419 words per dialogue.

For all these three measures of dialogue size, we could find two differences between confederate (type 1) and experimental dialogues (type 2) that can be captured by the following rules:

Central tendency: We found a significant difference in average dialogue size (with all three measures) for both dialogue types.

Dispersion: We found that type 1 dialogues have a significantly lower inter-quartile range than type 2 dialogues.

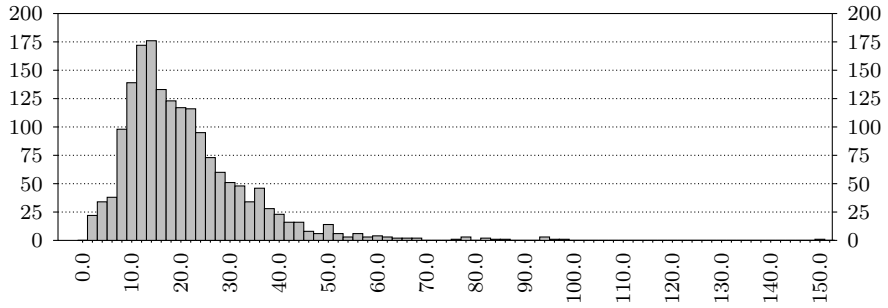


Figure 4: Length distribution of events, measured in seconds.

Discussion. It should be safe to assume that these differences are caused by the routine of the confederate. Since she knows the experimental setup and is very familiar with the objects and their positions, her effort in the game is considerably lower than that of the participant she is playing with. In addition, since she has the task of producing certain critical names, she is particularly involved in the production of a correct (thus efficient, fluent and shorter) utterance. The lower number of words could be explained by a lower number of hesitations, repairs and clarifications, at least on behalf of the participant.

4.3 Event size

Figure 4 shows a histogram for the distribution of event lengths over the whole corpus. This distribution has a positive skew which makes it jeopardous to describe the distribution by means of a central tendency or dispersion.

With this in mind, we can observe a median of 20.846 seconds. The distribution peaks at the range of 12–14 seconds. There are also much shorter events, even some that only take 4 seconds or less. On the other hand, the right tail of the distribution reaches the area of 60 seconds and more, and there are even outliers with events with a length of 90–100 seconds, and, in one case, 150 seconds.

Discussion. It seems that there is a range of typical durations for an event, since a majority of events lies within that duration range. This is apparent, since the task given in each event does not change in principle, but only by the particular object constellation given on each instruction card. As a consequence, participants have the opportunity to develop a routine during the game, which leads to a rather standardised length of events. Moreover, in several cases participants converge to some extent to one or more strategies during playing. These strategies can be syntactical patterns as well as particular techniques of how to describe spatial constellations on the table. This is an interesting parallel to the findings in the maze game experiments (cf. (Garrod and Anderson 1987), and (Pickering and Garrod 2004, pp. 171f)), where the participants aligned on a descriptive strategies for navigating through the maze.

Some of the outliers seem to be caused by a related phenomenon: In these cases, the participants noticed problems or inefficiency in their joint action. Then, they would

insert a negotiation sequence in which they perform a meta-communicative evaluation and improvement of the situation.

For example, in one case, both participants started the game with a lengthy discussion of from whose perspective to use the terms of "left" and "right", and whether or not they should switch that perspective according to who was in the role of the instructor.

In addition to the afore-given descriptive insights, the JMG data has been used in computational corpus linguistic applications, which are briefly reviewd in the following section.

5 Applications

The JMG corpus has been a starting point and first test scenario for a structural model of alignment. This model is a network model and has been formalised in Mehler, Lücking, and Weiß (2010). It is called the *Two-Level Time-Aligned Networks* (TiTAN) model. In its earliest stage it has been specifically adapted to the JMG data and the lexical, referential research topic tested therewith (cf. Section 2 above). The early TiTAN model represents lemmas from *dialogue lexica* (Pickering and Garrod 2004). It consists of a time-adjusted graph that was partitioned into two layers. The encompassing graph model captures the whole dialogue while interlocutors' lexica are captured as subgraphs. The graph is time-adjusted because of its dynamic update mechanism: driven by turn-taking the dialogue lexica are networked turn-wise. The result is a TiTAN series, where each step in this series captures the state of the dialogue lexicon after the turn-taking sequence leading to this state. Thus, TiTAN allows to track the formation of dialogue lexica.

The notion of alignment is operationalised within a quantitative model of structure formation based on the mutual information of the subgraphs that represent the interlocutor's dialogue lexica. This information-theoretic notion is adapted to graph structures by means of neighbourhood circles. Two corresponding nodes from the interlocutors' sub-lexica are compared in terms of their neighbours. The neighbours are ordered according to their distance (i.e., number of edges) measured from the reference nodes (Mehler, Lücking, and Weiß 2010). Thus, TiTAN contributes a similarity measure to the arsenal of network indices. For these reasons, the TiTAN model makes the structure and the formation of dialogue lexica accessible for machine learning applications. In Mehler, Lücking, and Weiß (2010), the TiTAN model has been evaluated against a subset of JMG dialogues. Based on TiTAN series modelling, JMG dialogues the manifest alignment has been shown to be distinguishable from JMG dialogues that do not show alignment.

The TiTAN model has been extended in several respects. In Mehler, Weiß, Menke, and Lücking (2010) it has been applied as a reference model of lexical alignment in the evolutionary framework of naming games.

The dynamic, dialogue formation aspect has been provided a neural network interpretation in Mehler, Lücking, and Menke (2011a). Here it is argued how spreading activation on the cortical level within interlocutors leads to observable alignment on the symbolic level of linguistic behaviour. In addition, the TiTAN model has proven to be able to

distinguish aligned from not-aligned dialogues, but this time taken all JMG dialogues into account.

In contrast to the strictly task-oriented JMG dialogues, TiTAN has also been applied to more free and spontaneous direction dialogues (Mehler, Lücking, and Menke 2011b). The classification task has been the separation of correct vs. false direction, making productive the claim of the alignment theory that aligned dyads are more effective than not aligned ones (Pickering and Garrod 2004). The classification task has been achieved with an F-score of 0.97

In ongoing work, the TiTAN model is extended towards multilogs and towards bilingualism.

6 Conclusion

We have presented the *Jigsaw Mapgame* (JMG) corpus, which is a dialogically and semantically annotated corpus on alignment of referential expressions based on the JMG experiment of Weiß, Pfeiffer, Schaffranietz, and Rickheit (2008). Due to the Elan XML-like format, it can be used in evaluation and machine learning applications. Due to its controlled, semi-spontaneous dialogues it can be used as a test scenario for studies on referential behaviour over the time-course of dialogues. The corpus already has been used in evaluating structural, network-based accounts on assessing lexical alignment. Thus, the JMG corpus is a useful tool for psycholinguistic as well as computational linguistics research that deals with referential behaviour and alignment.

Acknowledgements

The JMG corpus is part of the work of the CRC 673 "Alignment in Communication" funded by the German Research Foundation. We thank Petra Weiß and Sara Maria Hellmann for their support in designing annotation guidelines and helping annotating the data.

References

Anderson, A. H. et al. (1991). "The HCRC Map Task Corpus". In: *Language and Speech* 34, pp. 351–366.

Branigan, Holly P., Martin J. Pickering, and Alexandra A. Cleland (2000). "Syntactic Coordination in Dialogue". In: *Cognition* 25, B13–B25.

Clark, Herbert H. (1996). *Using Language*. Cambridge MA: Cambridge University Press.

Garrod, Siman and Martin J. Pickering (2004). "Why is conversation so easy?" In: *Trends in Cognitive Sciences* 8.1, pp. 8–11.

Garrod, Simon (1999). "The challenge of dialogue for theories of language processing". In: *Language processing*. Hove: Psychology Press, pp. 389–415.

Garrod, Simon and Anthony Anderson (1987). "Saying what you mean in dialogue: a study in conceptual and semantic co-ordination". In: *Cognition* 27.2, pp. 181–218.

Gwet, Kilem (2001). *Handbook of Inter-Rater Reliability*. Gaithersburg, MD: STATAXIS Publishing Company.

Krauss, R. M. and S. Weinheimer (1966). "Concurrent feedback, confirmation, and the encoding of referents in verbal communication". In: *Journal of Personality and Social Psychology* 4, pp. 343–346.

Mehler, Alexander, Andy Lücking, and Peter Menke (Aug. 2011a). "From Neural Activation to Symbolic Alignment: A Network-Based Approach to the Formation of Dialogue Lexica". In: *Proceedings of the 2011 International Joint Conference on Neural Networks*. IJCNN 2011. San Jose, California, pp. 527–536.

– (Feb. 2011b). "Modelling Lexical Alignment in Spontaneous Direction Dialogue Data by Means of a Lexicon Network Model". In: *12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2011)*. CICLing. Tokyo, Japan.

Mehler, Alexander, Andy Lücking, and Petra Weiß (2010). "A Network Model of Interpersonal Alignment in Dialogue". In: *Entropy* 12.6, pp. 1440–1483. DOI: 10.3390/e12061440.

Mehler, Alexander, Petra Weiß, Olga Abramov, and Sara-Maria Hellmann (2010). "Annotationshandbuch 1.1". Annotation Manual of the Jigsaw Map Game Corpus.

Mehler, Alexander, Petra Weiß, Peter Menke, and Andy Lücking (2010). "Towards a Simulation Model of Dialogical Alignment". In: *The Evolution of Language. Proceedings of the 8th International Conference*. Ed. by Andrew D. M. Smith, Marieke Schoustra, Bart de Boer, and Kenny Smith. EVOLANG8. Singapore: World Scientific, pp. 238–245.

Pickering, Martin J. and Simon Garrod (2004). "Toward a mechanistic psychology of dialogue". In: *Behavioral and Brain Sciences* 27, pp. 169–226.

Schaffranietz, Gesche, Petra Weiß, Thies Pfeiffer, and Gert Rickheit (2007). "Ein Experiment zur Koordination von Objektbezeichnungen im Dialog". In: *Kognitionsforschung 2007: Beiträge zur 8. Jahrestagung der Gesellschaft für Kognitionswissenschaft*, pp. 41–42.

Schiller, Niels O. and Jan Peer de Ruiter (2004). "Some notes on priming, alignment, and self-monitoring". In: *Behavioral and Brain Sciences* 27, pp. 208–209.

Schober, M. F. and S. E. Brennan (2003). "Processes of interactive spoken discourse: the role of the partner". In: *Handbook of discourse processes*. Ed. by A. C. Graesser, M. A. Gernsbacher, and S. R. Goldmann. Mahwah: Erlbaum, pp. 123–164.

Weiß, Petra, Thies Pfeiffer, Gesche Schaffranietz, and Gert Rickheit (Apr. 2008). "Coordination in dialog: Alignment of object naming in the Jigsaw Map Game". In: *Proceedings of the 8th Annual Meeting of the Cognitive Science Society of Germany*, pp. 1–17.

Weiß, Petra, Olga Pustylnikov, Alexander Mehler, and Sara Maria Hellmann (July 2009). "Patterns of alignment in dialogue: Conversational partners do not always stay aligned on common object names". In: *Proc. of the Conference on Embodied and Situated Language Processing*, p. 17.