Functional and temporal relations between spoken and gestured components of language

A corpus-based inquiry

Kasper I. Kok
Vrije Universiteit Amsterdam

Based on the Bielefeld Speech and Gesture Alignment Corpus (Lücking et al. 2013), this paper presents a systematic comparison of the linguistic characteristics of unimodal (speech only) and multimodal (gesture-accompanied) forms of language use. The results suggest that each of these two modes of expression is characterized by statistical preferences for certain types of words and grammatical categories. The words that are most frequently accompanied by a manual gesture, when controlled for their total frequency, include unspecific spatial lexemes, various deictic words, and particles that express difficulty in word retrieval or formulation. Other linguistic items, including pronouns and verbs of cognition, show a strong dispreference for being gesture-accompanied. The second part of the paper shows that gestures do not occur within a fixed time window relative to the word(s) they relate to, but the preferred temporal distance varies with the type of functional relation that exists between the verbal and gestural channel.

Keywords: gesture, multimodal corpus, distributional analysis, relative frequency ratio

1. Introduction

Although the fields of corpus linguistics and gesture studies share an interest in the characteristics of situated language use, there is little convergence between them in terms of methodology. This can be explained by both theoretical and practical reasons. The most obvious theoretical reason is that the linguistic relevance of co-verbal behaviors has not always been recognized in the (corpus) linguistic literature. Corpus studies are generally biased toward written language, or they focus exclusively on the verbal component of face-to-face communication. Co-verbal

International Journal of Corpus Linguistics 22:1 (2017), **1–26**. DOI 10.1075/ijcl.22.1.01kok
ISSN 1384-6655 / E-ISSN 1569-9811

gestures are typically not acknowledged as linguistically relevant aspects of the data. However, accumulating evidence suggests that manual and facial gestures intersect with the structure of spoken language in numerous ways (for reviews, see Kok 2016, Müller et al. 2013). It can even be argued that a speech-only view on language is fundamentally incomplete. Consider for instance the following (fictive) transcript of a spoken discourse segment:

> When you see a sign shaped something like this, follow the road that curves around the hill in this way. At the end of the road you'll see a little guardhouse with some of those […] you know […] thingies painted on it. You'll just need to go like […] and the they'll open the gate for you.

Utterances like these are likely to be accompanied by manual, facial and other types of gestures. These gestures may have functions that are analogous to linguistic elements, such as verbs, adjectives and nouns (Enfield 2004, Fricke 2009, Ladewig 2012), modal particles (Schoonjans 2014b), markers of negation (Harrison 2009) or markers of illocutionary force (Kendon 1995). Most written transcripts of spoken corpora, however, leave the co-verbal behaviors of the speakers to the imagination of the analyst. This yields a representation of spoken language that does only partial justice to the data source. Moreover, the exclusion of gestural behaviors from spoken transcripts eliminates the opportunity to use these corpora for studying how gestural behaviors relate to the structure of spoken language.

A second, practical reason for the general neglect of gestures in corpus linguistics is the traditional paucity of large video corpora. Most video corpora used by gesture scholars are rather small, especially when compared to the multi-million word corpora available in (computational) corpus linguistics. Although corpus-based observations have been of immense value to the field of gesture studies, the moderate size of the existing corpora often limits the generalizability of the patterns observed. Moreover, the majority of corpus-based gesture research involves at least one round of subjective judgment by the analyst (e.g. interpreting the gestures as belonging to some functional category). Thus, some of the core strengths of corpus linguistic methodology – generalizability and objectivity – have not been widely exploited in the domain of gesture studies.

The latter factor no longer needs to be a hurdle. In recent years, there has been a rise of large-scale video corpora (Diemer et al. 2016, Knight et al. 2009, Lücking et al. 2013, Turner & Steen 2012, van Son et al. 2008), some of which contain thousands of gestures and detailed linguistic annotations. As discussed by Adolphs & Carter (2013), the study of multimodal corpora raises questions and opens avenues of inquiry that have received little attention in traditional, text-oriented research. For one, the question arises of whether spoken language is structured along the same principles as written text. Taking speech as data source revives the issue of

how the basic units and structures of spoken language can be defined, and whether these are appropriately captured by traditional linguistic models (cf. McCarthy & Carter 1996). Second, multimodal corpora can be used to study the functional and structural relations between the different semiotic channels through which people communicate. Depending on the nature of the corpus, one can address questions about the relationship between speech and gesture, prosody, body stance, eye gaze, and image. Third, audio-visual corpora provide better means than text-only corpora to examine the various ways in which speakers and listeners interactively structure the discourse through auditory and gestural cues (cf. Knight 2011).

The current paper is concerned with the second type of question. It aims to provide insights into the functional relation between elements of speech and manual gesture. In contrast to most previous studies that have addressed this relationship, this paper pursues a fully systematic, bottom-up approach. Based on one of the largest annotated multimodal corpora currently available – the Bielefeld Speech and Gesture Alignment corpus (Lücking et al. 2013) – the linguistic characteristics of gesture-accompanied speech are compared to those of gesture-unaccompanied speech. The second part of the paper extends this analysis by examining the relative timing between spoken and gestured elements of expression.

2. Previous research on gesture-accompanied linguistic structures

Various previous studies have pointed out that certain verbal patterns are associated with gestural expression. McNeill (1992), for instance, discusses 'speech-linked gestures', which are performed in syntactic slots marked by phrases such as *like* in English (e.g. *I was like* + [facial gesture]). Others have found that specific words and constructions are often accompanied by specific manual gestures. This holds for a number of German modal particles (Schoonjans 2014a) as well as for constructions like *all the way from X to Y* in English (Zima 2014), among other examples.

Whereas these studies have started out from specific linguistic patterns, others have pursued a bottom-up approach to gain insights into the linguistic contexts in which gestures tend to occur. Hadar & Krauss (1999) and Morrel-Samuels & Krauss (1992) examine the distribution of the words that were labeled as 'lexical affiliates' of the gestures in their corpus (i.e. the words to which the gestures were judged to relate most). The authors report a general preference for gestures to be co-expressive with nouns, verbs and prepositions, relative to other grammatical categories. A drawback of this approach is that, as a consequence of drawing on the notion of lexical affiliate, there is a predisposition towards semantically loaded words on the part of the coder. That is, with this method, one does not detect words or constructions that correlate with gesture performance for reasons other than

co-expressivity (e.g. words that can be used to allocate attention to a gesture). To obtain more comprehensive insights into the linguistic structures that characterize gesture-accompanied speech, an approach is needed that bypasses an intermediate level of human interpretation.

One way of avoiding subjectivity is to use the acoustic features of the speech channel as a basis for identifying the words to which gestures relate (e.g. the pitch accent; Alahverdzhieva 2013). This strategy can be motivated by the finding that hand movements tend to be coordinated in time with movements of the vocal tract (Loehr 2004). However, an acoustically based approach still risks being biased towards certain word groups (e.g. content words more often receive prosodic stress than articles) and it assumes gestures to be directly aligned in time with the words they relate to. The current paper pursues an alternative approach. It uses a method which aggregates all words that occur in the temporal proximity of the gestures in the corpus, and compares these to the set of words that occur in unimodal contexts. The rationale behind this method follows from the view that spoken-only expression constitutes a different 'linguistic mode' than spoken-gestured expression (Cienki 2012). Hence, the research question is whether the verbal structures used in unimodal and multimodal modes of linguistic expression are qualitatively and/or quantitatively different.

3. The corpus

The Bielefeld Speech and Gesture Alignment Corpus (SaGA; Lücking et al. 2010, Lücking et al. 2013) consists of 25 dialogues in German, spanning a total of 280 minutes of video, recorded from three camera positions. The dialogues are task-oriented. The task conducted by the participants consists of two parts. First, one of the participants sits in front of a large video screen and watches a virtual reality animation that makes it appear as if she or he is taking a tour through a fictive town (SaGA town). The tour passes five landmarks: a sculpture, a church, a town hall, a chapel and a fountain. In the subsequent phase of the task, the first participant is told to instruct the second participant to follow the same path through the town. No further constraints were imposed as to the type of questions that could be asked, and participants were allowed to converse as long as they liked.

The total corpus contains 39,435 words and approximately six thousand gesture units.[1] The speakers' and listeners' expressions were heavily annotated during the

1. The SaGA documentation does not provide more specific numbers. It states that the corpus contains "4,961 iconic/deictic gestures [and] approximately 1,000 discourse gestures" (Lücking et al. 2013: 7). As explained in the methods section, the original distinction between iconic and

construction of the SaGA corpus, both on the level of speech (lemmatization, parts of speech, information structure) and the gestures (gesture type, form parameters, modes of representation) (see Lücking et al. 2013). Although some of these annotations can be used to address specific linguistic questions, the SaGA corpus was initially constructed for a different purpose. It was mostly oriented towards the design of virtual avatars and automated dialogue systems (Bergmann & Kopp 2009, Bergmann et al. 2010, Kopp et al. 2008, Lücking et al. 2010).

A feature of the corpus is that is useful for the present study is that it contains detailed information on the timing of the onsets and offsets of the words and gestures in the corpus. Word boundary segmentation was performed automatically by the WebMAUS plugin in ELAN (Kisler et al. 2012). The timing of the gestures has been annotated manually during the construction of the SaGA corpus, and has been validated through cross-coding (Lücking et al. 2013). The methods used in the current study for investigating the relationship between the spoken and gestured tiers of the corpus are described in the following section.

4. Methodology

Two linguistic data sets were abstracted from the SaGA corpus. The first will be called the lemma-corpus. It simply contains all lemma annotations from 23 videos in the SaGA corpus, ordered chronologically.[2] This corpus lends itself to addressing how gestures relate to the meanings and functions of individual words. The second is the POS-corpus, which consists of all part-of-speech tags that were assigned to the lemmas in the lemma-corpus. This level of analysis can provide an important addition, because connections between speech and gesture may exist on more abstract semantic levels than that of individual words (Kok & Cienki 2016).

For each unit of analysis, the corpus was divided into two sub-corpora: the gesture-accompanied sub-corpus contains all items that were uttered in the temporal proximity of the gestures in the corpus, whereas the speech-only sub-corpus contains all other items. As a definition of temporal proximity, the current analyses assume a time window of one second before and one second after the stroke phases

---

discourse gestures is not preserved in the current study, and a small part of the original corpus was excluded for the current analysis. Two videos were excluded from the analyses reported on in the current paper, because the relevant data were not available.

2. Note that Section 4 contains different uses of the word 'corpus'. In the context of the SaGA corpus, it refers to the entire, multi-tiered data set. The terms 'lemma-corpus' and 'POS-corpus' refer to single annotation tiers of the SaGA corpus. Hence, these terms essentially refer to different levels of representation of the same data, not to separate data sources.

of the gestures. That is, a word is considered to be gesture-accompanied if there is any temporal overlap between its articulation and the time frame that runs from one second prior to the onset of the gesture to one second after its offset (even if some part of the articulation falls outside this window). Previous literature has suggested that a time window of this size can be appropriate for capturing meaningful speech-gesture connections (Leonard & Cummins 2011, Loehr 2004, McNeill 1992). However, because this assumption cannot be taken for granted and has not been validated across different types of linguistic elements, the second part of this paper explores whether varying the operational definition of co-occurrence influences the results (e.g. assuming a temporal tolerance of 0, 2, or 3 seconds).

To assess the (dis)preference of linguistic structures for co-occurrence with manual gestures, the relative frequencies of all items in the speech-only sub-corpus are compared with those in the gesture-accompanied segments. The metric used for comparing these frequencies is the Relative Frequency Ratio (henceforth RFR; Damerau 1993). This is the ratio of the normalized frequencies of a linguistic item (a word or POS-tag) in the gesture-accompanied and the gesture-unaccompanied part of the corpus:

$$\mathrm{RFR}(i) = \frac{\left(\frac{\text{frequency of } i \text{ in gesture-accompanied sub-corpus}}{\text{number of items in gesture-accompanied sub-corpus}}\right)}{\left(\frac{\text{frequency of } i \text{ in speech-only sub-corpus}}{\text{number of items in speech-only sub-corpus}}\right)}$$

The RFR is a metric that serves to compare the linguistic characteristics of two or more corpora. Its application in the current context is consistent with Cienki's (2012) view of language as having different 'modes', discussed in Section 2. Provided that gesture-accompanied and gesture-unaccompanied speech constitute two different modes of expression, the RFR can be a helpful instrument to compare their linguistic characteristics. High values of the RFR indicate that the item occurs more often in the company of than in the absence of a gesture, taking into account the total size of each of the sub-corpora. In the plots below, the RFR values are mapped onto a natural logarithmic scale. Thus, positive numbers correspond to 'gesture-attracting' items, whereas negative numbers correspond to 'gesture-repelling' ones.

In order for the results to be meaningfully interpretable, it is important to take the role of chance into account. To assess which values of the (logged) RFR metric are different from what one might expect when comparing a random pair of sub-corpora, a confidence interval was estimated using a resampling method. The same analysis was applied five thousand times to pairs of randomly sampled sub-corpora of the same size as the two sub-corpora examined. This yields a distribution of the most likely values of the RFR for each of the items on the basis of

chance, which can be compared to the observed values. From the observed effect size and the confidence interval, a *p*-value was extracted (following Altman & Bland 2011), which can be interpreted as the likelihood that the observed RFR is a result of random variation. The following sections examine which lemmas and parts of speech have RFR values that exceed the 95% confidence interval. The findings are discussed in the light of the linguistic functions that gestures are capable of performing.

5. Analyses

This section presents the results of applying the procedures described above to the lemma-corpus and the POS-corpus. Subsequently, it examines whether these results are sensitive to the choice of the time-window that determines the corpus division.

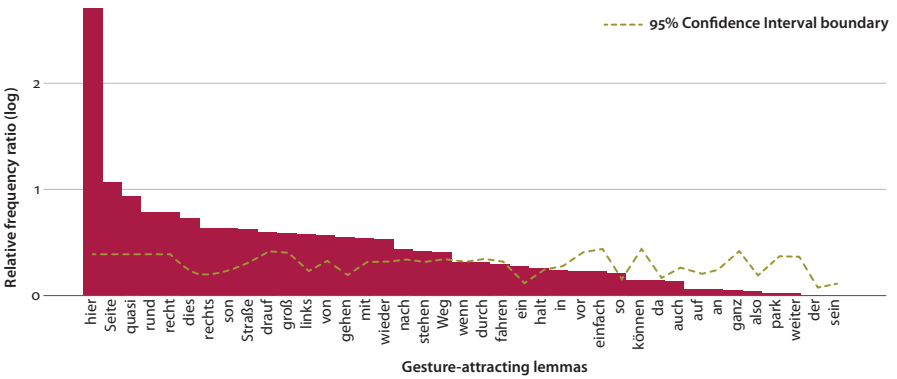
5.1 Lemma-level analyses

Assuming a one second tolerance, the total number of lemmas is 17,384 in the gesture-accompanied corpus and 13,986 in the speech-only corpus. To investigate the discrepancies between the relative frequencies in each of the sub-copora, Figure 1 plots the RFR values of all the lemmas in the corpus with at least 80 occurrences.[3] The dashed lines represent the outer borders of the 95% confidence intervals. Positive values, corresponding to gesture-attracting words, are shown in Figure 1a, while gesture-repelling words are displayed in Figure 1b.

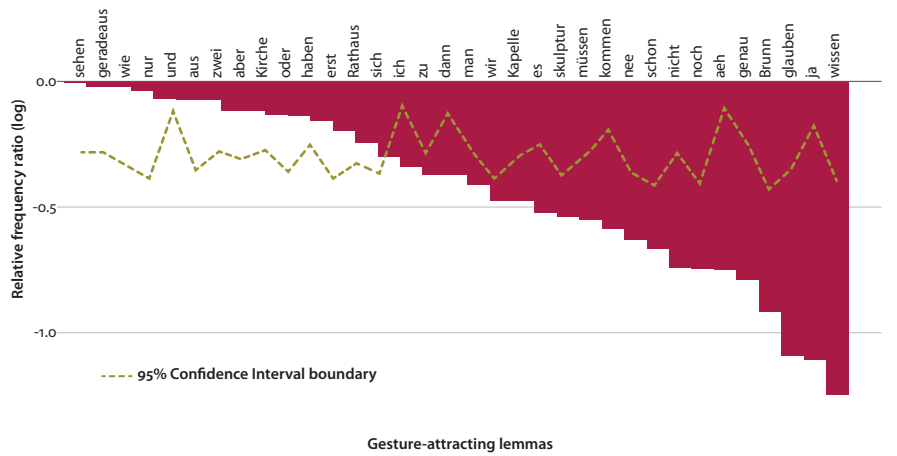
We see that 22 lemmas exceed the chance baseline on the gesture-attracting side, whereas 20 words have a RFR that is significantly lower than chance. All words with a RFR that exceeds chance level are listed in Table 1. The words are sorted according to their degree of gesture attraction. *p*-values are reported as a proxy for the statistical reliability of these results.

The list of gesture-attracting lemmas contains a variety of different word types. The proximal locative *hier* ("here") has by far the highest RFR score. A possible explanation for this finding is that (pointing) gestures are often used to restrict the reference domain of *hier*, which is otherwise somewhat indeterminate (see Fricke 2007 for a comprehensive discussion). In this light, it is surprising that its distal counterpart *da*/*dort* ("there") does not show up in this list of gesture-attracting

3. This threshold was chosen to allow for the inclusion of a large number of lemmas in the analysis, while maintaining sufficient statistical reliability.



1a.



1b.

Figure 1. Relative frequency ratios of most common lemmas (on a log scale); gesture-attractive words are shown in 1a, gesture-repelling ones in 1b (for translations, see Table 1)

words.[4] Distal locatives are more likely to be infelicitous when performed without some form of hand, head or eye movement (e.g. in utterances like *look over there*). The high RFR value for *hier* is also likely related to the fact that in the current discourse context (route direction), the speakers often refer to entities in fictive locations in front of their body. Since the participants speak about an environment

4. The distal deictic *da* is generally less marked than English *there*, however. The more emphatic version *dort* has too low a frequency in the corpus to be included here.

Table 1. Gesture-attracting lemmas

Lemma	English translation (most common senses)	N in gesture-accompanied sub-corpus	N in speech-only sub-corpus	Relative frequency ratio	*p*-value
hier	here	94	5	15.13	4.09e–38
Seite	side	87	24	2.92	1.07e–7
quasi	kinda, so to speak	76	24	2.55	3.45e–6
rund	(a)round	71	26	2.2	1.37e–4
recht	right	79	29	2.19	1.27e–4
dies	this, that	243	94	2.08	3.28e–11
rechts	(to the) right	288	122	1.9	2.1e–10
son	such a, a […] like this	202	86	1.89	1.42e–7
Straße	street	114	49	1.87	4.95e–5
drauf	on (top of) it/that	68	30	1.82	3.15e–3
groß	large	67	30	1.8	4.27e–3
links	(to the) left	239	108	1.78	1.34e–7
von	from, of	103	47	1.76	7.06e–4
gehen	go	304	141	1.73	3.28e–8
mit	with	111	52	1.72	7.52e–4
wieder	again	114	54	1.7	3.91e–4
nach	after	92	48	1.54	0.011
stehen	to stand	106	56	1.52	8.81e–3
Weg	street, road	86	46	1.5	0.019
ein	a(n)	594	361	1.32	1.41e–5
halt	well (discourse particle)	163	101	1.3	0.038
so	like this, in such a way	405	265	1.23	7.4e–3
Total N		**17,384**	**13,986**		

which they cannot perceive from the room in which they are located, they tend to set up fictive scenes in conversational space, so that they can display spatial relations between the objects referred to. The word *hier* in such cases can be used to establish a deictic center, with respect to which other entities can be referred to. This phenomenon is exemplified by the utterance in Example (1), taken from the SaGA corpus. The gestures in this utterance are both 'placing gestures', where the speaker moves his hands as if positioning some object in front of him. The temporal structure of the gesture is represented on a separate tier, following the conventions used by Kendon (2004). The label "prep" stands for the preparation phase of the gesture, "stroke" stands for the most effortful part, and "hold" is the phase following the stroke, where the hands are typically held in place for some time before they are retracted to rest position.

(1) also das *hier* ist das ganze U und dann ist das *hier* die Vorderseite

|~~~ |********** |~~~~~~~~~~ |*** |**********|

|prep stroke prep stroke hold

("so this *here* is the entire U and then this *here* is the front")

The high rating for the nouns *Seite* ("side"), *Straße* ("street") and *Weg* ("street") are also no doubt related to the specifics of the discourse situation. Route directions often involve reference to particular sides of the road or of the referenced objects (*on the left side you see* […]). In addition, phrases like *der straße folgen* ("follow the street") are particularly common in this discourse type. The data suggest that such phrases are comparatively often accompanied by manual gestures. A plausible reason is that, by virtue of their iconic potential, gestures can be more parsimonious than words when specifying the spatial relations between objects.

The high RFR for the discourse particle *quasi* ("kinda/so to speak") plausibly has different underlying reasons. Since the meaning of *quasi* is interpersonal in nature, typically expressing approximation or indeterminacy, the correlation with gesture performance cannot be indicative of a shared referent between the channels. Instead, it suggests that the use of gestures is generally linked to situations where speakers are not fully able to express themselves verbally, or not fully committed to the accuracy of their utterance. In some of these cases, the lack of verbal specificity might be compensated for by manual expression. In Example (2), for instance, the speaker expresses a lack of commitment to the accuracy of the word *Torbogen* for describing the object she refers to, and concurrently uses her hands to display its physical contours.

(2) wenn du rechts von dir halt sonen Torbogen siehst quasi ähm dann musst du da rein

|~~~~~~~~~~~~~~ |********** |******|~~~~~ |******* |*******|

|prep stroke hold prep stroke hold

("when you see on your right well one of those arches so to speak uhm then you have to get in there")

The discourse particle *halt*, which also shows up as gesture-attracting, is not semantically loaded either. *Halt* often marks the content of an utterance as plausible or indicates that something is pragmatically given or predefined in the communicative context (Schoonjans 2014b, Thurmair 1989). However, *halt* can also be used as a placeholder, i.e. as a way of delaying the discourse in order to plan an upcoming utterance. The latter use could be related to the tendency for it to co-occur with gestural expression. When *halt* is used to delay the upcoming speech in case of difficulty in lexical retrieval, gestures might be used to aid this retrieval process or to compensate for unspecific lexical content. An additional possibility follows from

the observation by Thurmair (1989) that *halt*, and other particles related to obviousness, can be used by speakers as a way of "masking" their uncertainty. Provided that this phenomenon is consistent in the current corpus, the gesture-attracting nature of *halt* is closely related to the high correlation of gestural expression with *quasi*.

We see two adjectives listed in Table 1: *rund* ("round") and *groß* ("large"). These have at least two features in common: their basic meaning is spatial and they are relatively unspecific. When performed in concurrence with these adjectives, gestures may function to further qualify their meaning, for instance indicating how large an object is or what dimension of it is round. The existence of a general relation between gesture and space-related words is also evident from the fact that several prepositions are found in Table 1: *von, mit* and *nach*. Although the spatial meaning of these and other prepositions is of course somewhat bleached in many cases, the current corpus does contain various instances of these words where they describe spatial relations that are simultaneously depicted using the hands.

We furthermore see three determiners in the list: *dies, son* and *ein*. As discussed by Hole & Klumpp (2000), *son* is a fully grammaticalized article (derived from *so ein*, "such a") that is used to refer to an indefinite token of a definite type. Fricke (2012) characterizes the function of the *son* + gesture combination as a "turning point" between characterizing a semantic category and singling out a specific token. When *son* is combined with a pointing gesture directed at an extralinguistic object, it marks the semantic type of the referent as identifiable, while the token remains indefinite (that is, pointing gestures combined with *son* do not designate a specific object, but a type or class of entities for which the referenced object is typical). *Son* combined with a depictive gesture (e.g. tracing the outline of an object), also narrows down the conceptual category referred to, but typically achieves a lower degree of type-definiteness. In either case, the gesture-attraction of *son* plausibly derives from its close relation to the potential of gestures to contribute to semantic specification. The finding that the indefinite article *ein* is substantially more gesture-attracting than the definite article *der* (in lemma-form, including other genders) further corroborates the finding that the gestures in the corpus more often support indefinite than definite reference. The high RFR for the demonstrative *dies* ("this/that"), however, suggests that demonstrative reference is an exception to this trend.

Finally, we see a high RFR for the qualitative deictic adverb *so*. *So* is a likely candidate for being accompanied by a depictive gesture, as it is generally associated with manner or quality expression (Fricke 2012, Streeck 2002). Streeck (2002: 582) claims that *so* can serve as "a 'flag' that alerts the interlocutor that there is extralinguistic meaning to be found and taken into account in making sense of what is being said." Note that although *so* is indeed found to be gesture-attracting, it has the lowest RFR of all words that exceed chance level. Given the raw frequencies, the current data compromise Streeck's (2002: 581) intuition that "when Germans

depict the world with their hands as they talk […] they almost always utter *so* in the process."

Table 2 displays the words for which the RFR is lower than chance level.

Table 2. Gesture-repelling lemmas

Lemma	English translation (most common senses)	N in gesture-accompanied sub-corpus	N in speech-only sub-corpus	Relative frequency ratio	*p*-value
wissen	to know	25	70	.29	5.42e–9
ja	Yes, [modal particle]	141	348	.33	2.3e–33
glauben	to believe	35	85	.33	3.23e–9
Brunnen	fountain	28	57	.40	3.48e–5
genau	exactly, I agree	94	168	.45	2.24e–10
aeh	uhm [filled pause]	329	567	.47	2.1e–28
noch	still, yet	35	60	.47	3.28e–4
nicht	not	78	133	.47	8.79e–8
schon	Already [discourse particle]	31	49	.51	.0025
nee	no	42	64	.53	.0013
kommen	to come, to arrive	161	235	.55	1.06e–8
müssen	must	67	94	.57	3.62e–4
Skulptur	sculpture	41	57	.58	.0071
es	it	88	120	.59	1.79e–4
wir	we	40	52	.62	.019
Kapelle	chapel	69	90	.62	.0033
man	one [indef. pers. prn.]	82	100	.66	.0052
zu	to, for, at	81	95	.68	.0107
dann	then	445	522	.68	1.44e–8
ich	I	678	771	.71	4.99e–11
Total N		**17,384**	**13,986**		

On the gesture-repelling side of the spectrum, there are twenty lemmas for which the RFR exceeds the baseline. The verbs *glauben* ("to believe") and *wissen* ("to know") are among the lemmas with the strongest tendency to occur without a manual gesture. These are both verbs of cognition that do not have clear spatial-perceptual features associated with them. Moreover, because these words take propositional complements, one can expect some degree of structural distance to the elements of the utterance that are more prone to gestural co-expression. The low ranking of the deontic modal *müssen* ("must") can also be accounted for by the first mentioned explanation, as it does not have clear spatial properties either. More remarkable is the fact that *kommen* ("to come"/"to arrive") shows up as gesture-repelling. Like the gesture-attracting verb *gehen* ("to go"), *kommen* expresses directed movement. The most salient semantic difference between these two verbs

is that *kommen* is associated with motion from a distal source to a proximal goal, whereas *gehen* refers to motion in the reverse direction. The finding that the former is substantially less often gesture-accompanied than the latter could be related to the fact that outward movement of the hands – congruent with the semantics of *gehen* – is more natural than movements that start from a distal location and are directed toward the body. In addition, the verb *kommen* may not bring the entire path of movement, but only the final segment of it into focus (cf. Langacker 1987: 69). It should also be noticed that there are many instances of *kommen* in the corpus where the deictic center is not the speaker, but a location in the town (e.g. *you come/arrive at a square*). Performing a gesture that parallels the path described by this use of *kommen* would entail a viewpoint shift from the perspective of the route-follower to that of a, presumably inanimate, point of arrival.

According to the current data, the word *ja* is also unlikely to be gesture-accompanied. This finding is remarkable in light of Schoonjans' (2014) finding that the modal particle *ja* correlates with a specific head gesture (the 'pragmatic headshake'). These findings are not incompatible, however, as the current data set does not carefully distinguish between the uses of *ja* as a responsive particle (translating into "yes") or as a modal particle (roughly translating into "simply"; indicating that no contradiction is expected). Moreover, head gestures are not taken into consideration in the present analysis. The placeholder *aeh* ("uhm") also occurs significantly more often in speech-only than in gesture-accompanied conditions. This appears at odds with the idea that gesture plays an important role in word retrieval (Krauss et al. 2000). Again, however, these findings cannot be compared directly. Filled pauses can be used for a range of interactional functions other than concept search, for instance allowing the speaker to plan upcoming sentences, and these are not systematically discerned in the transcription of the corpus.

The adverbs found among the gesture-repelling lemmas are semantically distinct from the ones we have seen in the list of gesture-attracting words. None of the adverbs found to be gesture-repelling – *dann* ("then"), *nicht* ("not"), *noch* ("still") and *schon* ("already"/"just") – have clear visual-spatial characteristics (in contrast to the gesture-attracting adverbs *rechts* and *links*). With respect to *nicht*, there is again an ostensible conflict with the previous literature, which has pointed out a close link between certain gestures and the verbal expression of negation (Harrison 2008, Kendon 2004). Although the current data do not contest the existence of an association between certain gestural forms and the verbal expression of negation, as these studies have suggested, they show that the German negation particle *nicht* is considerably more often expressed in unimodal than in multimodal contexts.

Finally, the list contains three personal pronouns: *es* ("it"), *ich* ("I"), and *wir* ("we"). These are typically unstressed words that occur in topic position. As gestures tend to occur together with newsworthy information (Levy & McNeill 1992,

McNeill 1992) pronouns are unlikely candidates for gestural co-expression. In addition, since *ich* and *wir* are self-referencing words, no depictive or indexical specification of their semantics is to be expected.

5.2 POS-based analyses

To gain deeper insight into the relation between gesture performance and the grammatical categories of the co-expressed words, the analytical procedures were repeated as applied to the Part-Of-Speech (POS) tags in the corpus. That is, instead of looking at lemma frequencies, the current section focuses on the distribution of the 22 different POS labels in the speech-only and gesture-accompanied sections of the corpus. The POS-tags were automatically assigned during the construction of the SaGA corpus by the Weblicht plugin in ELAN (Hinrichs et al. 2010) and are based on the Stuttgart-Tübingen-tagset (STTS; Schiller et al. 1995).[5] Figure 2 shows the RFR values for each of the grammatical categories, with gesture-attracting POS-labels on the left, and gesture-repelling ones on the right.

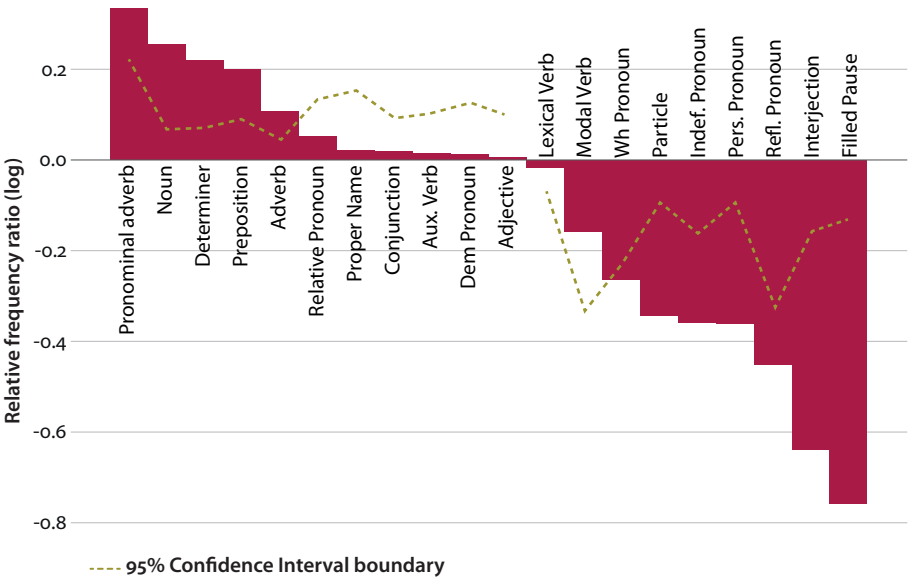


Figure 2. Relative frequency ratios for the POS-corpus

5. The labels used in the current analysis are based on the broader categories used by the STTS to group the more fine-grained labels.

From the visualized distribution, it appears that five parts of speech exceed the baseline on the positive side, whereas seven parts of speech were found to be significantly gesture-repelling. The gesture-attracting parts of speech are listed in Table 3.

Table 3. Gesture-attracting parts of speech

Lemma	N in gesture-accompanied subcorpus	N in speech-only subcorpus	Relative frequency ratio	*p*-value
Pronominal adverb	193	111	1.40	0.0039
Noun	2,036	1,266	1.29	7.91e–14
Determiner	1,740	1,122	1.25	4.41e–10
Preposition	1,161	763	1.22	5.73e–06
Adverb	3,712	2,677	1.12	1.88e–06
Total N	**17,384**	**13,986**		

The gesture-attraction values are highest for pronominal adverbs. These are words that refer to a spatial relationship with respect to a previously specified entity or location (e.g. *drauf,* "on (top of) it/that"; *darin,* "in it/that"). The high RFR suggests that the multimodal expression of such spatial relations is a common phenomenon. Simple prepositions are also found in the list of gesture-attracting parts of speech, but with a lower RFR. As seen above, the prepositions *auf, von, nach* and *mit* in particular have a tendency to be accompanied by manual gestures.

The finding that gestures are likely to occur in the company of nouns and determiners is in line with the idea that the hands can function as an attribute of a noun phrase (Fricke 2009, Ladewig 2012). Many gestures in the corpus serve to depict the size or shape of the landmarks that the speakers refer to. In one of the videos, for instance, the speaker refers to *ein großes rundes Fenster* ("a large round window") and traces the outline of the window with his index finger. This gesture can be interpreted as co-expressing (or specifying) the semantic content of the noun phrase. In this light, the finding that adjectives show up as 'gesture-neutral' (RFR = 1.01, $p = .95$) is rather striking. A possible explanation is that adjectives and gestures fulfill similar roles and therefore generally cancel out each other's necessity: when a depictive gesture is performed together with a noun phrase, an adjective with (roughly) the same meaning is no longer needed, and vice versa.

For verb phrases, the observed pattern is remarkably different from what we see for noun phrases. Gestures do have a significant tendency to co-occur with adverbs, but they are not correlated with any type of verb (for lexical verbs: RFR = 1.01, $p = .74$). This finding could allude to a differential contribution of gestures to noun-phrases and verb phrases. Provided that immediate temporal coincidence between a gesture and a word is indicative of a functional analogy, it follows that when gestures co-occur with a verb phrase, they will tend to take a role analogous

to a modifier, not to the verb itself. For gestures performed in concomitance with a noun phrase, by contrast, the closest functional analog of the gesture seems to be the head noun. Given the limits of the current data set and discourse genre, claims like these of course remain somewhat speculative, but the statistical trends observed appear rather robust. Table 4 shows the parts of speech that occupy the lower end of the spectrum. These correspond to some of the linguistic categories that gestures are unlikely to co-occur with.

Table 4. Gesture-repelling parts of speech

Lemma	N in gesture-accompanied subcorpus	N in speech-only subcorpus	Relative frequency ratio	*p*-value
Filled Pause	330	567	.47	2.1e–28
Interjection	250	381	.53	3.77e–15
Refl. Pronoun	57	72	.64	.012
Pers. Pronoun	841	971	.69	1.6e–14
Indef. Pronoun	269	310	.70	2.29e–5
Particle	802	910	.71	1.23e–12
Wh-Pronoun	123	129	.77	.038
Total N	**17,384**	**13,986**		

The most obvious common denominator in the list of gesture-repelling parts of speech is that all are generally short words: we see four different types of pronouns, filled pauses, interjections and particles. As mentioned before, it can be assumed that pronouns are gesture-repelling because they are likely to occur in positions with given, rather than new information. The low gesture-attraction for the other word types – filled pauses, interjections and discourse particles – are understandable for the same reason. An additional explanation might be that the latter set of words do not have clear iconic or indexical properties. Some of their pragmatic functions could be co-expressed by facial gestures (e.g. nodding, shoulder shrugging), but the current data suggest that the functions of interjections and particles are not systematically associated with hand movements.

5.3 The effect of the choice of time window

The findings presented so far are based on a somewhat arbitrarily chosen definition of temporal co-occurrence – linguistic units were considered to be gesture-accompanied if they were performed no more than one second before or after the stroke phase of a gesture. Although this decision was motivated by previous literature (see Section 4), it is imaginable that the results vary when a time window of a different sized is used. The current section explores whether and how modifying

the operational definition of co-occurrence influences the results of the analysis presented above, and discusses how this informs the temporal dynamics of spoken-gestured expression.

A relevant finding in the light of the current research interest is that the relative timing between speech and gesture varies along with certain semantic factors. Morrel-Samuels & Krauss (1992) find that the onset latency between gestures and their lexical affiliates is inversely correlated with the familiarity of these words; less familiar words occur with more temporal distance to the co-expressed gestures than more familiar words. Bergmann et al. (2011) also report on an interaction between timing and semantics. They show that speech and gesture are produced in closer temporal proximity when they are semantically redundant (i.e. when they express more or less the same information) than when they are complementary in meaning.

This section examines timing effects in the current data. The above procedures are repeated with amended criteria for dividing the corpus into speech-only and gesture-accompanied parts. That is, the RFR values are compared under a range of different operational criteria for considering a word as gesture-accompanied. These include a zero-lag condition – where only those words are regarded as gesture-accompanied that overlap directly with a gesture stroke – as well as conditions with a temporal tolerance of up to four seconds (three of these divisions are sketched in Figure 3).

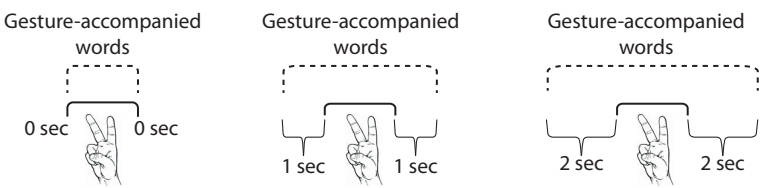


Figure 3. Different ways of operationalizing the notion of speech-gesture co-occurrence. All intervals between 0 and 4 seconds, with 0.5 second increments, are taken into consideration

Apart from the modification of the time window, the analyses carried out here follow the exact same procedures as above. To avoid data abundance, a set of eight lemmas and a set of seven parts of speech were selected to serve as point in case. The selection was based on the RFR scores (all significantly above chance according to the previous analysis) and on functional diversity. Figure 4 shows the RFR scores of the eight selected lemmas as a function of the size of the time window. The dashed lines represent the chance baseline (upper limits of the 95% confidence intervals), computed separately for each corpus division. Note that the plots are scaled to fit the window, so that the contours are most visible. As a consequence, different scales are used on the y-axes for each of the lemmas.

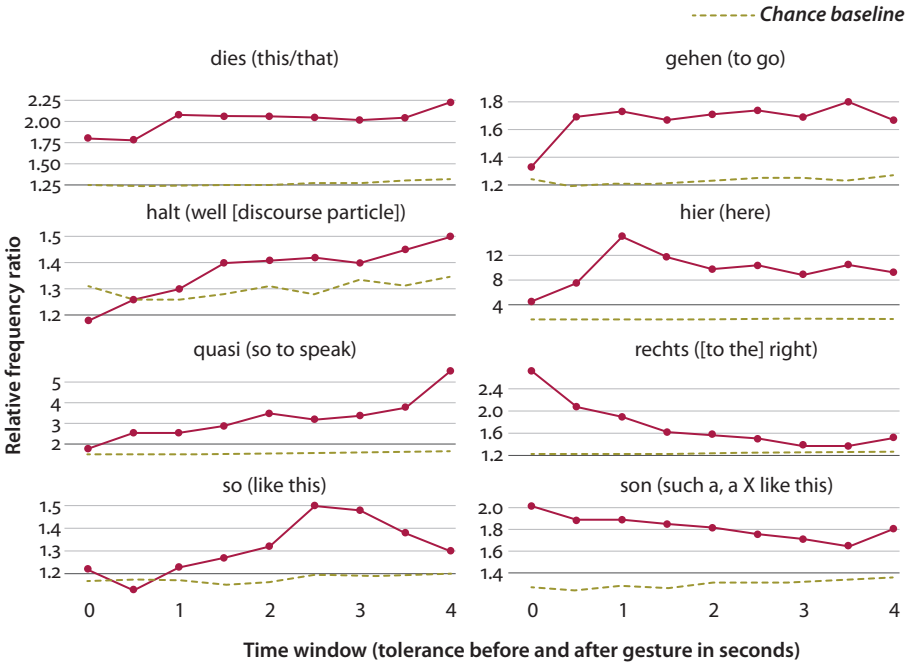


Figure 4. Relative frequency ratios of gesture-attracting words for different time windows

From visual inspection of the plot, it is evident that the choice of time window has a differential impact for the various lemmas examined. Some contours, in particular those for *dies* and *son*, are relatively flat. Both of these words are determiners with a deictic component, which can allocate the interlocutor's attention to some quality depicted gesturally. The correlation of these words with gestural expression, however, appears not to be limited to direct co-occurrence; RFR scores are of the same order of magnitude when considering wider time windows. This suggests that *son* and *dies* potentially (re)allocate the interlocutor's attention not just directly after these words are uttered, but possibly up until multiple seconds thereafter (or before). The relationship between the word *so* and gestural expression also plays out on a rather wide time scale, but the preferred time window appears to be more restricted. The RFR is at chance for the zero-lag condition, and peaks around 2.5–3 seconds. A similar type of pattern is found for the locative adverb *hier*. Its gesture-attractiveness holds for any time window, but the signal-to-noise ratio appears highest when the temporal tolerance is defined at one second.

One of the words in Figure 4 has a maximum RFR for a time window of zero seconds: *rechts* ("to the right"). This suggests that when *rechts* is expressed in concurrence with a manual gesture, there tends to be a very short lag or no lag at all. The opposite is true for the word *gehen* ("to go"). We see that *gehen* has a RFR that

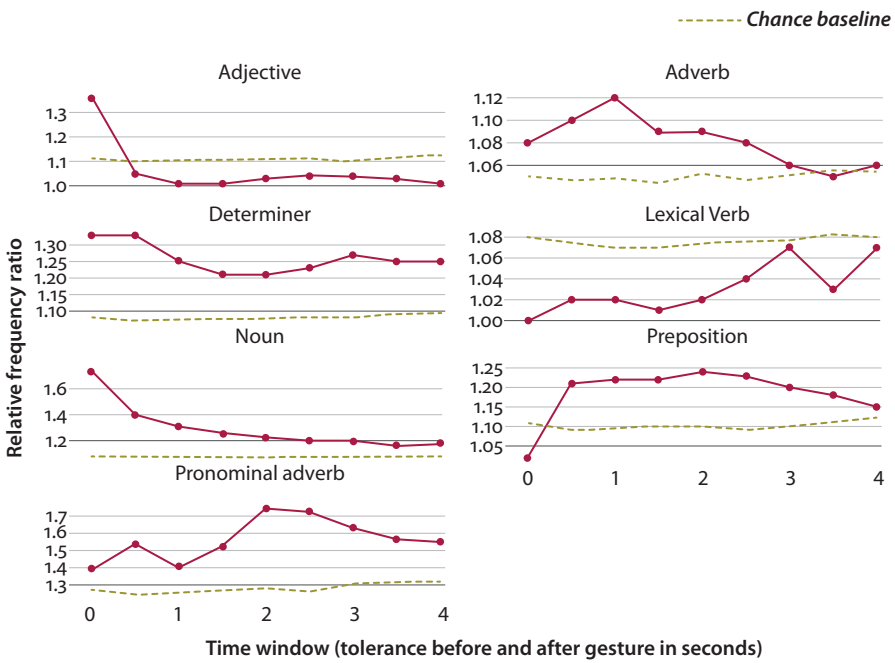


Figure 5. Relative frequency ratios of gesture-attracting parts of speech for different time windows

is close to the chance baseline when looking only at immediate temporal overlap between the verbal and the gestural channel. For all larger time windows, however, the gesture-attraction value remains well above chance. The inverse relationship between the contours of *rechts* and *gehen* is striking, as it may be expected that these words often go together in route directions, for instance in phrases like *du gehst rechts* ("you go to the right"). The current data suggest that when such phrases are accompanied by a gesture, the gesture is more likely to temporally coincide with the adverb than with the verb.

For two of the words inspected, we see a steady increase of the RFR as a function of the temporal tolerance: *quasi* and *halt*. The observed correlation of these words with gesture occurrence becomes stronger when larger time windows are taken into account. As discussed above, *halt* and *quasi* are both discourse particles that have a relatively indirect relation to gestural expression. The current data suggest that the relation between gestures and linguistic elements which perform a meta-discursive function (e.g. hedging, holding the floor) is characterized by relatively large onset latencies.

The final analysis addresses how the choice of time window impacts the results on the level of grammatical categories (Figure 5). The set of POS-tags taken into

account includes all parts of speech that were found to be gesture-attracting, plus adjectives and lexical verbs (which have a high frequency, but were not found to be gesture-attracting in the previous analysis).

Like in the previous analysis, we see a diversity of patterns. For determiners, the RFR values are quite stable, with slightly higher scores for smaller time windows. This contour plausibly results from collapsing over definite articles, indefinite articles and demonstratives, which have somewhat diverse dynamics, as seen above. Nouns and adjectives have homologous patterns, with higher RFR values for direct co-occurrence than for larger time windows. The line for adjectives takes the steepest descent, dropping below the chance baseline for all other time windows than the zero-lag one. This presents an important qualification to the findings presented above, where no positive values for adjectives were reported; adjectives are apparently correlated with gesture use only when looking at immediate coincidence. By contrast, the RFR values for nouns stay above chance for all time windows.

For prepositions and pronominal adverbs, the observed relation is somewhat inverse to what we see for nouns. In the zero-lag condition, no significant gesture-attraction is observed for prepositions, and the RFR only marginally exceeds the baseline for pronominal adverbs. The gesture-attraction of these grammatical classes shows up much more clearly, however, with any larger time window. A possible interpretation of this finding is that the semantic relationship between gestures and the meanings of prepositions and pronominal adverbs is indirect in nature: given that these word types typically denote spatial relationships, they relate more strongly to the relative temporal and spatial positioning of successively performed gestures than to the individual gesture strokes.

Regarding lexical verbs and adverbs, the current data show that the findings reported in the previous section are relatively independent of timing. For almost all time windows, there is a discrepancy between the high RFR rates for adverbs and the low ones for verbs. Unlike what we have seen for adjectives, the low values for verbs hold for any choice of time window, although the chance baseline is approximated for larger windows. As far as adverbs are concerned, we see that the RFR values peak at one second and drop below chance level at three seconds. The hypothesis can be derived that gestures that modify verb phrases are typically performed within three seconds of the articulation of the adverb(s) they relate to. When interpreting these results, however, it should again be borne in mind that subtle patterns in the data could be masked as a result of averaging across all adverbs in the corpus, including words such as *links* and *rechts* (with short articulation-gesture lags) and *so* (with larger lags).

6. Discussion and conclusion

The application of corpus linguistic methods to multimodal data can yield insights into the functional and temporal relations between spoken and gestured components of linguistic expression. Through a bottom-up method, the current paper has revealed tendencies for manual gestures to be co-expressed with particular words and grammatical categories. A small set of words was found to be positively correlated with gesture performance, including several lexemes with perceptual-spatial meanings, deictic terms and discourse particles. Other lemmas were less often gesture-accompanied than expected by chance. These included words without clear spatial features (e.g. verbs of cognition) and words that typically have topical status in an utterance (e.g. pronouns). A comparable analysis, applied to parts of speech tags, corroborated that certain word classes are more "gesture-friendly" than others. Pronominal adverbs, nouns, determiners, prepositions and adverbs were found to occur significantly more often in multimodal than in unimodal contexts. This is in line with the view that gestures can perform some of the functions that these words have, such as making reference to entities and ascribing static and dynamic properties to them. However, the interaction between gestures and grammatical structure appears more complex than this. Neither adjectives nor verbs were found to be on the gesture-attracting side of the spectrum, suggesting a differential role of co-speech gestures in noun phrases and verb phrases.

A subsequent analysis examined the relative timing of some of the most gesture-attracting words relative to gesture performance. The degree of gesture-attraction of the linguistic units inspected was found to vary substantially with the choice of time window that was used to define co-occurrence. Some linguistic items are most strongly gesture-attracting when looking only at direct coincidence (e.g. *gehen*, "to go"), whereas other words seem to be in a much looser temporal connection with gesture performance (e.g. *quasi*, "so to speak"). An examination of the temporal aspects of the POS-tier yielded similar results: the preferred gesture-articulation interval is variable among the different grammatical categories examined. This finding has clear methodological implications for studies that investigate patterns in speech-gesture co-occurrence. It shows that the correlational results one obtains are strongly dependent on one's criteria for regarding a word as gesture-accompanied. The preferred relative timing between speech and gesture is by no means stable, but varies with the linguistic functions gestures serve in the context of the utterance.

For the current approach to be viable, it was necessary to aggregate across all speakers in the corpus and across all types of manual gesture – token frequencies were not sufficient to allow for more fine-grained analyses. As a result, the patterns observed are somewhat rudimentary and a number of possible limitations

are worth mentioning. For one, it remains unclear whether the tendencies observed apply equally to each individual speaker. Previous research, conducted on the basis of the same corpus, has shown that gesture styles differ substantially among participants, and that patterns on the individual level do not necessarily reflect those on the aggregate level (Bergmann et al. 2010). Another possible drawback of the current approach is that it treats manual gesture as isolated from other bodily behaviors. In reality, strong relations exist between manual behaviors and movements of the body, eye gaze and intonation (Loehr 2004, Streeck 1993). As mentioned before, different results may have been obtained if different bodily articulators were taken into account – verbal expressions that are negatively correlated with manual gesture may be positively correlated with head or shoulder gestures. Furthermore, the current approach treats the gestures in the corpus as independent from each other and as products of the speakers only. However, dialogue participants are known to adapt their gestures to their own and each other's behaviors in previous stages of the discourse (e.g. McNeill 2000). These dynamics are not captured by the current methodology.

Given that other, possibly larger corpora can be studied using the procedures introduced in this paper, several extensions of this research are imaginable. One of the most urgent ones is to validate the methods and the results across different types of corpus (e.g. elicited versus free conversations) and different discourse contexts. This can reveal to what extent the outcomes can be generalized across other settings than the route direction discourse examined here. Another avenue of future research is to take more complex verbal units into account, such as bigrams and semi-filled word sequences (e.g. *VP + like* in English). A further refined categorization of the gestural behaviors could also be valuable; separate analyses could for instance be conducted for iconic, indexical and discourse-related gestures. However, the pervasive multifunctionality of gestural expression renders the notion of gestural category somewhat problematic (Kok et al. 2016). A perhaps more fruitful direction of future research is to focus on specific gestural patterns and the linguistic characteristics of the verbal contexts in which they are performed. With a few modifications, the current method can be applied to arrive at a detailed characterization of the 'linguistic profiles' of recurrent gestural units. These profiles would not only include the lexical-grammatical characteristics of the contexts in which they tend to occur, but also a representation of their preferred timing relative to the spoken tier. Given the numerous potential ways for validating and extending the results obtained, the contents of this paper are surely just the tip of the iceberg when it comes to seeking convergence between gesture studies and corpus linguistics.

Acknowledgements

The author is grateful to the Netherlands Scientific Organization (NWO; PGW-12–39) and the German Academic Exchange Service (DAAD; 91526618–50015537) for their support. He would also like to thank Kirsten Bergmann and Stefan Kopp for granting him access to the data, and Alan Cienki, Mike Hannay and Lachlan Mackenzie for valuable comments on an earlier version of the manuscript.

References

Adolphs, S., & Carter, R. (2013). *Spoken corpus linguistics: From monomodal to multimodal*. New York, NY: Routledge.

Alahverdzhieva, K. (2013). *Alignment of speech and co-speech gesture in a constraint-based grammar*. (Unpublished doctoral dissertation), University of Edinburgh, Edinburgh.

Altman, D. G., & Bland, J. M. (2011). How to obtain the P value from a confidence interval. *British Medical Journal*, *343*(2304). Retrieved from http://www.bmj.com/content/343/bmj.d2304 (last accessed February 2017).

Bergmann, K., Aksu, V., & Kopp, S. (2011). *The relation of speech and gestures: Temporal synchrony follows semantic synchrony*. Paper presented at the 2nd Workshop on Gesture and Speech in Interaction, Bielefeld, Germany.

Bergmann, K., & Kopp, S. (2009). GNetIc – Using Bayesian decision networks for iconic gesture generation. In Z. Ruttkay, M. Kipp, A. Nijholt & H. H. Vilhjálmsson (Eds.), *Proceedings of the 9th International Conference on Virtual Agents* (pp. 76–89). Amsterdam: Springer.

Bergmann, K., Kopp, S., & Eyssel, F. (2010). Individualized gesturing outperforms average gesturing – evaluating gesture production in virtual humans. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud & A. Safonova (Eds.), *Proceedings of the 10th Conference on Intelligent Virtual Agents* (pp. 104–117). Philadelphia, PA: Springer. doi:10.1007/978-3-642-15892-6_11

Cienki, A. (2012). Usage events of spoken language and the symbolic units we (may) abstract from them. In K. Kosecki & J. Badio (Eds.), *Cognitive Processes in Language* (pp. 149–158). Frankfurt am Main: Peter Lang.

Damerau, F. J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management*, *29*(4), 433–447. doi:10.1016/0306-4573(93)90039-G

Diemer, S., Brunner, M. L., & Schmidt, S. (2016). Compiling computer-mediated spoken language corpora. *International Journal of Corpus Linguistics*, *21*(3), 348–371.

Enfield, N. J. (2004). On linear segmentation and combinatorics in co-speech gesture: A symmetry-dominance construction in Lao fish trap descriptions. *Semiotica*, *149*(1–4), 57–124.

Fricke, E. (2007). *Origo, Geste und Raum: Lokaldeixis im Deutschen*. Berlin: Walter de Gruyter. doi:10.1515/9783110897746

Fricke, E. (2009). *Multimodal attribution: How gestures are syntactically integrated into spoken language*. Paper presented at the first Gesture and Speech in Interaction conference (GeSpIn), Poznań, Poland.

Fricke, E. (2012). *Grammatik multimodal: Wie Wörter und Gesten zusammenwirken*. Berlin: Walter de Gruyter. doi:10.1515/9783110218893

Hadar, U., & Krauss, R. K. (1999). Iconic gestures: The grammatical categories of lexical affiliates. *Journal of Neurolinguistics*, *12*(1), 1–12. doi: 10.1016/S0911-6044(99)00001-9

Harrison, S. (2008). The expression of negation through grammar and gesture. In J. Zlatev, M. Andrén, M. J. Falck & C. Lundmark (Eds.), *Studies in Language and Cognition* (pp. 405–419). Cambridge: Cambridge Scholars Press.

Harrison, S. (2009). *Grammar, gesture, and cognition: The case of negation in English* (Unpublished doctoral dissertation). University of Bordeaux, Bordeaux, France.

Hinrichs, E., Hinrichs, M., & Zastrow, T. (2010). WebLicht: Web-based LRT services for German. In S. Kübler (Ed.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics 2010 System Demonstrations* (pp. 25–29). Uppsala, Sweden.

Hole, D., & Klumpp, G. (2000). Definite type and indefinite token: The article son in colloquial German. *Linguistische Berichte*, *182*(1), 231–244.

Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *Journal of Pragmatics*, *23*(3), 247–279. doi: 10.1016/0378-2166(94)00037-F

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511807572

Knight, D. (2011). *Multimodality and Active Listenership: A Corpus Approach*. London: Continuum Books.

Knight, D., Evans, D., Carter, R., & Adolphs, S. (2009). HeadTalk, HandTalk and the corpus: Towards a framework for multi-modal, multi-media corpus development. *Corpora*, *4*(1), 1–32. doi: 10.3366/E1749503209000203

Kisler, T., Schiel, F., & Sloetjes, H. (2012). Signal processing via web services: The use case WebMAUS. In E. Hinrichs, H. Neuroth & P. Wittenburg (Eds.), *Proceedings of the Service-oriented Architectures (SOAs) workshop at the Digital Humanities Conference 2012* (pp. 30–34). Hamburg, Germany.

Kok, K. I. (2016). The grammatical potential of co-speech gesture: A Functional Discourse Grammar perspective. *Functions of Language*, *23*(2), 149–178.

Kok, K. I., Bergmann, K., Cienki, A., & Kopp, S. (2016). Mapping out the multifunctionality of speakers' gestures. *Gesture*, *15*(1), 37–59.

Kok, K. I., & Cienki, A. (2016). Cognitive Grammar and gesture: Points of convergence, advances and challenges. *Cognitive Linguistics*, *27*(1), 67–100.

Kopp, S., Bergmann, K., & Wachsmuth, I. (2008). Multimodal communication from multimodal thinking – towards an integrated model of speech and gesture production. *International Journal of Semantic Computing*, *2*(1), 115–136. doi: 10.1142/S1793351X08000361

Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), *Language and Gesture* (pp. 261–283). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511620850.017

Ladewig, S. H. (2012). *Syntactic and semantic integration of gestures into speech: Structural, cognitive, and conceptual aspects* (Unpublished doctoral dissertation). European University Viadrina, Frankfurt (Oder).

Langacker, R. W. (1987). *Foundations of Cognitive Grammar, Volume I: Theoretical Prerequisites*. Stanford: Stanford University Press.

Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, *26*(10), 1457–1471. doi: 10.1080/01690965.2010.500218

Levy, E. T., & McNeill, D. (1992). Speech, gesture, and discourse. *Discourse Processes*, *15*(3), 277–301. doi: 10.1080/01638539209544813

Loehr, D. P. (2004). *Gesture and intonation* (Unpublished doctoral dissertation). Georgetown University, Washington D.C.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H. (2010). The Bielefeld speech and gesture alignment corpus (SaGA). In M. Kipp, J. C. Martin, P. Paggio & D. Heylen (Eds.), *Proceedings of the 7th International Conference for Language Resources and Evaluation* (pp. 92–98). Valetta, Malta.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H. (2013). Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces*, *7*(1–2), 5–18. doi: 10.1007/s12193-012-0106-8

McCarthy, M., & Carter, R. (1996). Ten criteria for a spoken grammar. In E. Hinkel & S. Fotos (Eds.), *New Perspectives in Grammar Teaching in Second Language Classrooms* (pp. 51–75). Mahwah, NJ: Lawrence Erlbaum.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL: University of Chicago Press.

McNeill, D. (2000). Catchments and contexts: Non-modular factors in speech and gesture production. In D. McNeill (Ed.), *Language and Gesture* (pp. 312–328). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511620850.019

Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 615–622.

Müller, C., Ladewig, S. H., & Bressem, J. (2013). Gestures and speech from a linguistic perspective: A new field and its history. In C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill & J. Bressem (Eds.), *Body-Language-Communication: An International Handbook on Multimodality in Human Interaction* (Vol. *1*, pp. 55–81). Berlin/Boston: De Gruyter Mouton.

Schiller, A., Teufel, S., & Thielen, C. (1995). *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Unpublished report. University of Stuttgart.

Schoonjans, S. (2014a). Is gesture subject to grammaticalization? *Papers of the Linguistic Society of Belgium*, *8*. Retrieved from http://uahost.uantwerpen.be/linguist/SBKL/Vol8.htm (last accessed May 2016).

Schoonjans, S. (2014b). *Modalpartikeln als multimodale Konstruktionen. Eine korpusbasierte Kookkurrenzanalyse von Modalpartikeln und Gestik im Deutschen* (Unpublished doctoral dissertation). University of Leuven, Leuven, Belgium.

Streeck, J. (1993). Gesture as communication: Its coordination with gaze and speech. *Communications Monographs*, *60*(4), 275–299. doi: 10.1080/03637759309376314

Streeck, J. (2002). Grammars, words, and embodied meanings: On the uses and evolution of so and like. *Journal of Communication*, *52*(3), 581–596. doi: 10.1111/j.1460-2466.2002.tb02563.x

Thurmair, M. (1989). *Modalpartikeln und ihre Kombinationen*. Tübingen: Niemeyer. doi: 10.1515/9783111354569

Turner, M., & Steen, F. (2012). Multimodal Construction Grammar. In M. Borkent, B. Dancygier & J. A. J. Hinnell (Eds.), *Language and the Creative Mind* (pp. 255–274). Stanford, CA: CSLI Publications.

van Son, R., Wesseling, W., Sanders, E., & van den Heuvel, H. (2008). The IFADV corpus: A free dialog video corpus. *Proceedings of the sixth international conference on Language Resources and Evaluation (LREC)* (pp. 501–508). Marrakech: European Language Resources Association (ELRA).

Zima, E. (2014). English multimodal motion constructions. A construction grammar perspective. *Papers of the Linguistic Society of Belgium*, *8*. Retrieved from http://uahost.uantwerpen.be/linguist/SBKL/sbkl2013/Zim2013.pdf (last accessed May 2016).

Author's address

Kasper I. Kok
Department of language, literature and communication
Vrije Universiteit Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
k.i.kok@vu.nl