Timing Relationships between Speech and Co-Verbal Gestures in Spontaneous French

Gaëlle Ferré

To cite this version:

Gaëlle Ferré. Timing Relationships between Speech and Co-Verbal Gestures in Spontaneous French. Language Resources and Evaluation, Workshop on Multimodal Corpora, May 2010, Malta. pp.86-91. hal-00485797

HAL Id: hal-00485797

https://hal.archives-ouvertes.fr/hal-00485797

Submitted on 21 May 2010

Timing Relationships between Speech and Co-Verbal Gestures in Spontaneous French

Gaëlle Ferré
Laboratoire de Linguistique de Nantes (LLING)
Université de Nantes
Chemin de la Censive du Tertre, BP 81227
44312 Nantes cedex 3 -- FRANCE
E-mail: Gaelle.Ferre@univ-nantes.fr

Abstract

Several studies have described the links between gesture and speech in terms of timing, most of them concentrating on the production of hand gestures during speech or during pauses (Beattie & Aboudan, 1994; Nobe, 2000). Other studies have focused on the anticipation, synchronization or delay of gestures regarding their co-occurrence with speech (Schegloff, 1984; McNeill, 1992, 2005; Kipp, 2003; Loehr, 2004; Chui, 2005; Kida & Faraco, 2008; Leonard and Cummins, 2009) and we would like to participate in the debate in the present paper. We studied the timing relationships between iconic gestures and their lexical affiliates (Kipp, Neff et al., 2001) in a corpus of French conversational speech involving 6 speakers and annotated both in Praat (Boersma & Weenink, 2009) and Anvil (Kipp, 2001).
The timing relationships we observed concerned the position of the gesture stroke as compared to that of the lexical affiliate and the Intonation Phrase, as well as the position of the gesture Phrase as regards that of the Intonation Phrase. The main results show that although gesture and speech are co-occurring, gestures generally start before the related speech segment.

1. Introduction

These last years, a major effort has been made by the international community to extend the number, variety and size of annotated multimodal corpora in several languages, especially since it has been shown by McNeill (1992) among others that gestures play a role in communication. Nowadays, some tools even allow automatic recognition of body movements (Campbell, 2009) which will save time in the annotation of the less interpretative gesture configurations (eyebrow rise, hand trajectory, for instance). Yet, manual annotation is still needed for features that involve interpretation (type of hand gesture, etc). Manual annotation is also needed for corpora recorded before the development of special tools.

This is why the OTIM project (Bertrand et al., 2008; Blache et al., 2009) is based on the annotation of several hours of conversational speech in French. Part of the annotation process is automatic (transcription and alignment of words and phonemes, annotation of syntactic clauses and morphological categories), but the rest is manual (gesture and body movements and postures, prosodic phenomena, discourse units). These annotations (whether automatic or manual) are made with the annotation tool Praat (Boersma & Weenink, 2009) for speech and Anvil (Kipp, 2001) for gestures, which is not the case of every study concerning gesture-speech relationships (for instance, the studies of Chui, 2005, and Kida & Faraco, 2008, were not based on alignment of speech transcription and gesture annotation). This does not mean that linguistic studies which are not based on time-aligned annotations are of no value, but simply that temporal alignment of annotations adds precision to otherwise more intuitive observations.

Among the studies concerned with co-verbal gestures, a few of them described the timing relationships between gesture and speech. Beattie & Aboudan (1994) and Nobe (2000), for instance, analyzed gesture production co-occurring with speech or with silent pauses. Others like Schegloff (1984), McNeill (2001, 2005), Kipp (2003), Loehr (2004), Kranstedt et al. (2006) and Leonard & Cummins (2009) for English, Chui (2005) for Chinese, Kida & Faraco (2008) for French, and Rochet-Capellan (2008) for French and Portuguese, concentrated on the timing of gesture with regards accompanying speech or parts of speech. These studies all show the interest of developing annotated corpora and timing relationships will also be the object of the present paper, in which we will compare the timing of the gesture stroke with regard to the lexical affiliate, and the gesture phrase with regard to the Intonation Phrase, after having briefly presented the corpus and the annotations made.

2. Corpus and data

This study is based on analysis of the data in a subset of the CID video corpus, fully described in Bertrand et al. (2008) and Blache et al. (2009). This corpus is still under the annotation process (OTIM project), but we were able to work on the hand gestures annotated in 75 minutes of speech, involving 6 speakers in dialogues of spontaneous French.

2.1 Speech transcription and annotations used to establish timing relationships

The paper is based on a semi-automatic transcription and its alignment with the sound file in Praat, which were then corrected manually. Intonation Phrases (IPs) as defined by Selkirk (1978 and later works) have also been annotated in Praat: whereas syntactic units such as clauses or sentences would be relevant for written texts, we considered that Intonation Phrases are quite appropriate for the chunking of speech recordings since prosody (including pauses, different degrees of stress and boundaries) gives some clue on information structure. If we consider the following example from the corpus:

/ y avait un espèce d'écran géant donc un matériel d'enfer /

The utterance could have two possible interpretations due to the structure of spoken French and the possible placement of the conjunction *donc* ("so") which can be placed before, in the middle or after the clause in its syntactic domain: (a) so there was some sort of huge screen, high tech resources, or (b) there was some sort of huge screen so they had high tech resources. Now if we consider prosody, the ambiguity is not present anymore and the two Intonation Phrases are in fact:

/ y avait un espèce d'écran géant / donc un matériel d'enfer /

This is not determined by the presence of a pause as there is none in the example but rather by the fact that there is a pitch rise on "géant" and a reset on "donc". If "donc had been part of the first Intonation Phrase, it would still have been low in pitch but the pitch reset would have occurred on "un", so that there would have been a prosodic break between "donc" and "un".

Other studies have previously established a relationship between prosodic units and gestures. For instance, Loehr (2004) has shown that there is a timing relationship between Intermediate Phrases and gesture phrases (described in the next section). His study is in the framework of J. Pierrehumbert's autosegmental theory of intonation and what he terms Intermediate Phrases corresponds to Intonation Phrases in Selkirk's metrical theory so the prosodic units we are considering are the same. Below is represented the metrical analysis of the Intonation Phrase "tu signes le papier" ("you sign the paper", the example is also given in section 2.3).

```
(                       x)  Intonation Phrase
(     x    )  (         x)  Accentual Phrases
(     x    )  (         x)  ω
     (x    )        (   x)  Σ
 x    x  x     x    x   x   σ
 tu  signes    le  papier
```

In this study, the distinction between Major and Minor accentual phrases (Kratzer and Selkirk, 2007) was not relevant since only Intonation Phrases were annotated, but it is important to understand exactly what elements they comprise. The stresses in Accentual Phrases correspond to pitch accents in the autosegmental theory, whereas stresses at Intonation Phrase level correspond to phrase tones in the autosegmental theory. Selkirk analyses a further level, the sentence level, which is not relevant here. Stresses at this level would correspond to edge tones.

For all these reasons, IPs seemed to be quite an appropriate unit in a comparison between speech and gesture, and their timing has been directly compared to the timing of gesture phrases as described in the next paragraph, whereas gesture strokes have been linked to lexical affiliates. All speech transcriptions and annotations made in Praat were then imported in Anvil which was used for the annotation of gestural phenomena.

2.2 Gesture annotations

Although the general OTIM project has started the annotation of various gestures, movements and postures that include all body parts, the author of the present paper has been more particularly concerned with the annotation of hand gestures. So far, 1477 gestures have been annotated on 75 minutes of speech (the ultimate aim being to annotate all the gestures during 3 hours of corpus). Each hand gesture was described in terms of symmetry (single-handed vs. two-handed gesture, symmetric vs. asymmetric hand configuration). We then annotated each gesture's phases (Kendon, 1980): preparation – stroke – hold – retraction – recoil, as well as the gesture's phrase (Kendon, op. cit.), that is the gesture in its whole, to which we assigned a dimension (McNeill, 2005) or a function regarding co-occurring speech (we retained the semiotic types proposed by Kipp, 2003). Each gesture was also described in terms of hand shape, gesture trajectory, space, velocity and amplitude, although these descriptions which were useful in determining lexical affiliates, were not used *per se* in the present study. The gesture onset corresponds to the first frame in which the hand(s) moves from its rest position whereas the offset corresponds to the first frame in which the hand returns to its rest position when the gesture is produced in isolation. When the gesture is produced in between two other gestures without any return to rest position, its onset corresponds to the first frame in which the hand changes trajectory from the previous gesture (initiates the preparation or stroke of the gesture). Its offset corresponds to the last frame before the hand changes trajectory for the preparation or stroke of the next gesture. One has to keep in mind that due to the granularity of the videos (24 frames per second), the onset and offset of hand gestures are defined less precisely than the onset and offset of speech shown in Figure 1 below.

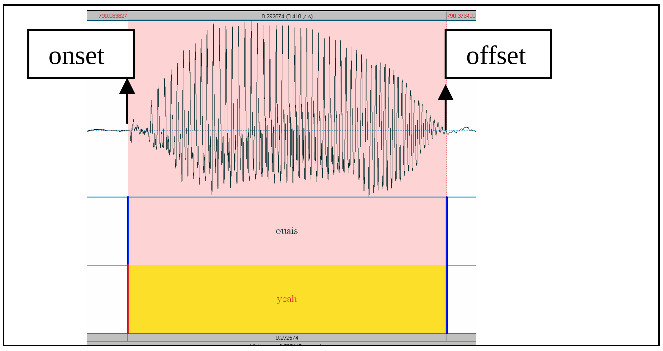
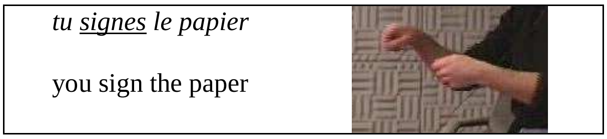


Figure 1: Speech onset and offset.

Among the annotations, we retained only the iconic hand gestures for the study, which was based on a number of 244 gestures out of a total of 286 iconics (42 were discarded, either because it was not possible to determine a lexical affiliate in speech due to the absence of a verbal affiliate or due to the fact that it was not possible to determine precisely a word in speech which would have a close meaning to the one conveyed by the gesture; some of the gestures were also interrupted and not taken into consideration.

2.3 Lexical affiliate

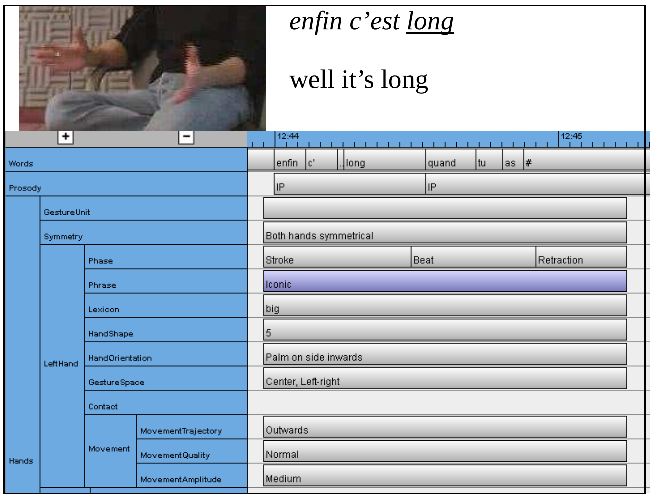
In order to establish a relationship between gesture and speech in terms of timing, it is necessary that the link between them be of an explicit nature. Schegloff (1984) described lexical affiliates as "the word or words deemed to correspond most closely to a gesture in meaning." In the case of iconics, it appears that in 85.3 % of the occurrences, it is possible to determine a lexical affiliate that actually corresponds to a word in the co-occurring speech, which is quite a high rate. This close correspondence between lexical affiliates and what Kipp calls 'redundant iconics' (2003:153) is shown in Figures 2 and 3.



tu signes le papier

you sign the paper

Figure 2: Iconic gesture corresponding to the lexical affiliate "sign" in terms of hand configuration and movement.



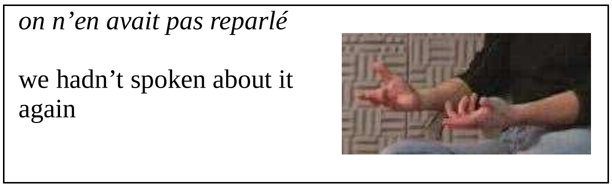
enfin c'est long

well it's long

```
Words                     enfin | c' | ..long | quand | tu | as | #
Prosody                   IP | IP
Hands  GestureUnit
       Symmetry           Both hands symmetrical
       LeftHand  Phase              Stroke | Beat | Retraction
                 Phrase             Iconic
                 Lexicon            big
                 HandShape          5
                 HandOrientation    Palm on side inwards
                 GestureSpace       Center, Left-right
                 Contact
                 Movement  MovementTrajectory  Outwards
                           MovementQuality     Normal
                           MovementAmplitude   Medium
```

Figure 3: Iconic gesture and Anvil annotation corresponding to the lexical affiliate "long".

This close correspondence between gesture feature and meaning in speech is not so explicit with other gesture dimensions such as metaphorics, for instance, which may be used to add a modality to the entire IP as presented in Figure 4:



on n'en avait pas reparlé

we hadn't spoken about it again

Figure 4: Metaphoric gesture adding a modality to the IP with no precise lexical affiliate.

The explicit affiliation is the reason why we chose to study the timing relationships between iconics and co-occurring speech which was also the choice made by Chui (2005), whereas Schegloff (1984) for similar reasons, chose deictics (the number of deictics in our corpus was too small with only 137 occurrences to motivate such a choice, this depending much on the type of corpus), and Leonard & Cummins (2009) chose beats. Other studies had a larger understanding of lexical affiliation and did not restrict their observations to a particular dimension (Loehr, 2004).

3. Results

The first observation that needs to be made concerning the timing of gestures and co-occurring speech is that gesture units (Gstroke at lexical level and Gphrase at phrase level) are generally longer than corresponding verbal units (word and IP), as shown in Figure 5 below, and that the difference between the mean duration of phrasal units is smaller than the difference of the mean duration at word level.

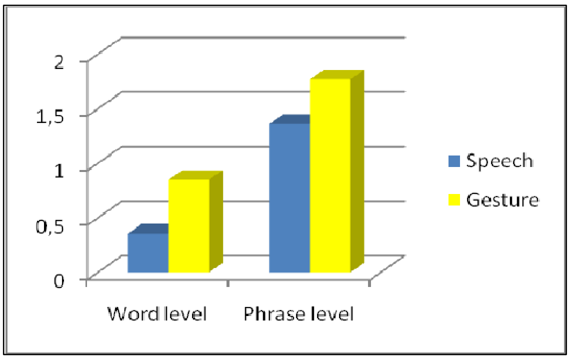


Figure 5: Mean duration (in seconds) of lexical units (affiliate/Gstroke) and phrasal ones (IP/Gphrase).

In terms of timing relationships (*cf.* percentages given in Table 1 below), at word level, when comparing the onset and offset of gesture stroke and lexical affiliate, we observe that a large majority of strokes (72 %) start before the onset of the lexical affiliates and an even greater proportion of strokes (87 %) end after the offset of lexical affiliates. A paired T-Test showed that the anticipation of strokes on lexical affiliates is highly significant: the mean difference (M=-0.454, SD=0.692, N=244) is significantly greater than zero (t(243)=-10.2, two-tail p=1.07e-20). A 95 % C.I. about stroke/affiliate onset is (-0.54, -0.36). However, although a higher proportion of gesture strokes ended after the offset of the lexical affiliate, the mean difference (M=0.03, SD=0.76, N=244) is not significant (t(243)=0.80, two-tail p=0.42). A 95 % C.I. about stroke/affiliate offset is (-0.05, 0.13). These statistics show that the difference in timing of the onset and offset of gesture and speech are not only due to the fact that gesture strokes are generally longer than lexical affiliates but also that there is a marked preference for anticipation in the gesture production. It is also quite important to say that in accordance with McNeill's remarks on the question of co-occurrence of gesture and speech (2005), gesture strokes and lexical affiliates are generally produced in overlap in our corpus with only 22 % of strokes being completed before the production of the corresponding speech affiliate.

At phrase level (GPhrase vs. IP), the tendency is exactly similar with an anticipation of Gphrase of the same order as the one at word level (70 % of Gphrases start before the onset of IPs). A paired T-Test showed that the anticipation of GPhrases on IPs is significant: the mean difference (M=-0.19, SD=0.79, N=244) is significantly greater than zero (t(243)=-3.8, two-tail p=0.0001). A 95 % C.I. about GPhrase/IP onset is (-0.29, -0.09). Although the percentages are not clearly cut for the offset (61 % of gesture offset occurring after IP offset), the paired T-Test showed a mean difference (M=-0.22, SD=0.86, N=244) significantly greater than zero (t(243)=-3.96, two-tail p=9.7e-05). A 95 % C.I. about GPhrase/IP offset is (-0.32, -0.11). In all the occurrences, as opposed to the production of lexical affiliates and gesture strokes, an overlap between the production of Gphrases and IPs was observed. There was no occurrence of a Gphrase completed before its corresponding IP.

Lastly, comparing the production of gesture phrases and lexical affiliates, we only found 21 cases (8.6 %) where the gesture phrase was completed before the production of the lexical affiliate (although it was overlapping the IP containing the affiliate). Most of the cases contained some verbal hesitation as in the following example:

le village / il fait une espèce de / il est sur une espèce de colline [The village makes some sort of / is on some sort of hill.]

Where the speaker produces two identical iconic semi-spherical gestures representing a 'hill'. What is apparent here is that the idea of a hill had already formed in the speaker's mind but due to the false start, the first gesture is not synchronized with the lexical affiliate 'hill'.

Concerning the comparison of gesture and speech production of Gphrases vs. affiliates, we note that in 95 % of the cases, the Gphrase onset starts before the onset of the affiliate. The paired T-Test showed a mean difference (M=-0.82, SD=0.76, N=244) significantly greater than zero (t(243)=-16.90, two-tail p=6.50e-43). A 95 % C.I. about GPhrase/affiliate onset is (-0.92, -0.72). A high proportion (75 %) of Gphrase offsets occur after the offset of affiliates. Once again, the paired t-test provided evidence that the mean difference (M=0.595, SD=0.92, N=244) is significantly greater than zero (t(243)=10.05, two-tail p=4.21e-20). A 95 % C.I. about GPhrase/affiliate offset is (0.47, 0.71).

	% of gestures starting		**% of gestures ending**	
	Before speech	**After speech**	**Before speech**	**After speech**
Gstroke/ Affiliate	72	28	12	87
Gphrase/IP	70	30	39	61
Gphrase /Affiliate	95	5	25	75

Table 1: Percentage of gestures starting/ending before or after speech.

The results concerning gesture-speech timing relationships may be summarized in the following figure:

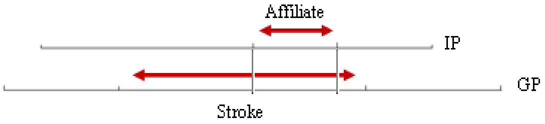


Figure 6: Timing relationships between gesture and Intonation Phrase, and between gesture stroke and lexical affiliate.

4. Discussion and conclusion

In this paper, we presented the results of one of the few studies on gesture-speech synchrony in the case of iconics. The choice of the iconic dimension was justified by the explicit relationship which exists between redundant iconics and lexical affiliates (namely words).

The results in this study – obtained from 244 iconics produced by 6 speakers during 75 minutes of spontaneous French – show that the timing relationships between gesture and speech are much alike in French and in English, as opposed to Chinese. Indeed, in Chinese, Chui (2005:878) found a higher proportion of gestures synchronized with speech than gestures anticipating speech (60.1 % vs. 35.6 %). In English, on the contrary, Schegloff (1984), who worked on deictics, observed that gesture strokes are generally produced in anticipation to the lexical affiliate. In a recent study, Leonard & Cummins (2009) also find an anticipation of gesture in English. Their work was more precisely based on the temporal alignment of beats' phases with lexical affiliates. They showed – on a very restricted corpus though – that the onset of the gesture stroke anticipated on the vowel onset in the corresponding affiliate. They also found, like in the present study, that the gesture offset occurred after speech. Although we did not have any movement detection device during the recording of the corpus[1] (and would therefore not reach the same degree of precision as Leonard & Cummins), the corpus has also been transcribed into phonemes so we should be able to go into finer detail in the future. More refinement will also be needed concerning the relationship between

[1] This type of recording would not be quite possible with spontaneous interactions.

gesture stroke and other speech and gesture dimensions. For instance, Rochet-Capellan et al. (2008) showed that in French and Portuguese, deictics' alignment with speech was very much dependent on the number of syllables in the co-occurring speech: although they found general gesture-speech synchrony, they also observed a 'tendency to delay pointing events with the increase of *n-syl*' which could result 'from the interaction between the two systems' (p. 160). Working on deictics as well, Kranstedt et al. (2006) found that the initiation of the gesture generally starts slightly after the co-occurring speech and that the stroke generally ends before the affiliate (p. 145). The difference between these last two studies and our results may very well lie in the fact that they are based on experimental corpora, whereas the present study is based on uncontrolled speech, but it would be interesting in a future study to investigate whether the variability in gesture-speech timing can be explained by different gesture amplitude. In their experimental setting Kranstedt et al. (2006) the participants were pointing to objects on a table, some of which were quite near the participants, others being quite distant. Our complex annotation on the CID which codes gesture amplitude would make such an enquiry possible.

What we added to the studies on timing relationships was the fact that not only gesture strokes (i.e. the relevant part of the gesture) and lexical affiliates could be compared, but that we can also compare the timing relationships between gesture and intonation phrases, since the lexical affiliate is in the same type of relationship with the entire IP as the stroke is with the gesture phrase, which means that a correspondence can be established between stroke and lexical affiliate as between gesture phrase and IP. At phrase level, the timing relationships are of the same order as those at word level. This corroborates what Loehr (2004) found for English, although his results showed higher gesture / speech synchrony (but he mentions variability). This difference can be explained by the fact that Loehr considered all gesture types.

At last, this study shows one of the many interesting aspects of the annotation of multimodal corpora, since only this type of annotation allows a comparison between units from different modalities, such as co-verbal gestures and speech: some of the studies quoted in this paper have been produced without any systematic annotation of either gestures or speech units. They certainly helped in formulating hypotheses on timing relationships, but it is extremely difficult to obtain precise results in terms of temporal alignment, even when one watches a video frame by frame, whereas greater precision can be attained when using annotation software like Praat for speech and Anvil for gesture phenomena. The results in such studies can be used to improve the generation of animated agents.

5. Acknowledgements

This research is supported by the French National Research Agency (Project number: ANR BLAN08-2_349062) and is based on annotations made by various team members besides the author of the current paper. The OTIM project is referenced on the following webpage: http://aune.lpl.univ-aix.fr/~otim/. The author thanks the reviewers for their valuable comments and suggestions on a previous version of this paper.

6. References

Beattie, G. and R. Aboudan (1994). Gestures, pauses and speech - an experimental investigation of the effects of changing social-context on their precise temporal relationships. *Semiotica*, 99, pp. 3--4.

Bertrand, R., Blache, P., et al. (2008). Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *Traitement Automatique des Langues*, 49(3), pp. 105--133.

Blache, P., Bertrand, R., et al. (2009). Creating and Exploiting Multimodal Annotated Corpora: The ToMA Project. In M. Kipp (ed.), *Multimodal Corpora*. Berlin, Heidelberg: Springer-Verlag, pp. 38--53.

Boersma, P. & Weenink, D. (2009). *Praat: doing phonetics by computer (Version 5.1.05)* [Computer program]. Retrieved May 1, 2009, from http://www.praat.org/

Campbell, N. (2009). Tools and Resources for Visualising Conversational-Speech Interaction. In M. Kipp et al. (Eds), *Multimodal Corpora. From Models of Natural Interaction to Systems and Applications.* Berlin, Heidelberg: Springer-Verlag, pp. 176--188.

Chui, K. (2005). Temporal Patterning of Speech and Iconic Gestures in Conversational Discourse. *Journal of Pragmatics*, 37, pp. 871--887.

Ferré, G. (2002). Les pauses démarcatives déplacées en anglais spontané : marquage prosodique et kinésique. *Lidil*, 26, *Gestualité et syntaxe*, pp. 155--169.

Kendon, A. (1980). Gesture and speech: two aspects of the process of utterance. In M.R. Key (ed.), *Nonverbal Communication and Language*, The Hague: Mouton, pp. 207--227.

Kida, T. & Faraco, M. (2008). Prédication gestuelle. *Faits de Langues*, 31-32 (La prédication), pp. 217--226.

Kipp, M. (2001). Anvil - A Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1367--1370.

Kipp, M. (2003). Gesture Generation by Imitation. From Human Behavior to Computer Character Animation. PhD Thesis, Saarbrucken: Saarland University.

Kipp, M., Neff, M., et al. (2007). An annotation Scheme for Conversational Gestures: How to Economically Capture Timing and Form. In *Proceedings of Language Resources and Evaluation*, 41, pp. 325--339.

Kranstedt A., A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth. (2006). Deictic object reference in task-oriented dialogue. In G. Rickheit and I. Wachsmuth (eds.), *Situated Communication*, Berlin: Mouton de Gruiter, pp. 155–208.

Kratzer, A. & selkirk, E. (2007). Phase theory and prosodic spellout: The case of verbs. *The Linguistic Review*, 24**,** pp. 95--135.

Leonard, T. and Cummins, F. (2009). Temporal Alignment of Gesture and Speech. In *Proceedings of Gespin*, Poznan, Pologne (24-26 septembre). [CD-Rom].

Loehr, D. (2004). *Gesture and Intonation.* PhD Thesis. Georgetown University.

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

McNeill, D. (2005). *Gesture and Thought*. Chicago and London: The University of Chicago Press.

Nobe, S. (2000). Where do *most* spontaneous representational gestures actually occur with respect to speech? In D. McNeill (Ed.), *Language and Gesture*. Cambridge: CUP, pp. 186--198.

Rochet-Capellan, A., Vilain, C., Dohen, M., Laboissière, R. & Schwartz, J.-L. (2008). Does the Number of Syllables Affect the Finger Pointing Movement in a Pointing-naming Task? *8th International Seminar on Speech Production (ISSP 2008).* Strasbourg, pp. 257--260.

Schegloff, E. A. (1984). On Some Gestures' Relation to Talk. In J. M. Atkinson and J. Heritage (Eds.), *Structures of Social Action*. Cambridge: CUP, pp. 266--298.

Selkirk, E. (1978). On Prosodic Structure and its Relation to Syntactic Structure. In T. Fretheim (Ed.), *Nordic Prosody II*. Trondheim: Tapir, pp 111--140.