Relations temporelles entre parole et gestualité co-verbale en français spontané

Gaëlle Ferré

Relations temporelles entre parole et gestualité co-verbale en français spontané

Gaëlle Ferré

Laboratoire de Linguistique de Nantes (LLING)
Université de Nantes
Chemin de la Censive du Tertre, BP 81227
44312 Nantes cedex 3
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 [1]; Nobe [16]). Other studies have focused on the anticipation or delay of gestures regarding their co-occurrence with speech (Schegloff [18]; McNeill [15]; Chui [6]; Kida & Faraco [9]; Leonard and Cummins [13]) and we would like to take part in the debate in the present paper. We studied the timing relationships between iconic gestures and their lexical affiliates (Kipp, Neff et al. [11]) in a corpus of French conversational speech involving 6 speakers and annotated both in Praat (Boersma & Weenink [4]) and Anvil (Kipp [10]).

Keywords: Multimodality, co-verbal gestures, timing relationships, lexical affiliates

1. INTRODUCTION

Parmi les études toujours plus nombreuses en multimodalité qui s'intéressent à la gestualité co-verbale — dont le rôle communicationnel a été montré par McNeill [15] entre autres — un certain nombre s'est attaché à décrire les relations temporelles qui existent entre le geste et la parole. L'un des intérêts de ce type de recherche est de pouvoir comprendre les systèmes multimodaux et de pouvoir ainsi alimenter le développement d'avatars. Ainsi, par exemple, Beattie & Aboudan [1] et Nobe [16] se sont penchés sur la co-occurrence des gestes manuels et des pauses silencieuses ou du temps d'articulation. D'autres études (Schegloff [18], McNeill [15], Leonard & Cummins [13] sur l'anglais; Chui [6] sur le chinois; Kida & Faraco [9] sur le français) se sont concentrées sur l'anticipation ou le retard de la gestualité co-verbale par rapport à la parole. C'est sur ce point que portera le présent article, car avec le développement des corpus vidéos annotés, une plus grande précision peut être atteinte. Ainsi, nous avons travaillé sur le corpus CID (Bertrand, Blache et al. [2], Blache, Bertrand et al. [3]) et analysé les relations temporelles entre les gestes iconiques (décrits dans la section 2.2) et la parole. Pour ce faire, nous avons mis en relation les groupes intonatifs (IP) avec les phrases gestuelles (cf. section 2.2), et les affiliés lexicaux (cf. section 2.3) avec la phase de réalisation du geste (Gstroke, cf. section 2.2), car ces unités nous ont semblées comparables. Les résultats montrent une très nette anticipation de la gestualité par rapport à la parole. Ils montrent aussi que si une unité gestuelle peut être décomposée en plusieurs items comme l'unité intonative peut se décomposer en mots, les unités gestuelles sont également plus longues que les unités verbales.

2. CORPUS ET DONNÉES

Pour cette étude, nous avons travaillé sur une sous-partie du corpus vidéo CID (décrit dans Bertrand, Blache et al. [2]), soit 45 minutes de parole interactionnelle (3 dyades de 15 minutes chacune) impliquant 6 locuteurs. L'annotation et l'exploitation du corpus font l'objet actuellement d'un projet financé par l'ANR (ANR BLAN08-2_349062).

2.1. Transcription du corpus

Nous avons travaillé sur une transcription et un alignement semi-automatique du corpus dans Praat, corrigés manuellement. Les groupes intonatifs (Intonational Phrases, Selkirk [19]) ont également été annotés dans Praat. Nous avons en effet pensé que cette unité était beaucoup plus appropriée au découpage de l'oral que des unités comme la phrase syntaxique ou la proposition qui présentent certains inconvénients et ne correspondent pas toujours au découpage exprimé par les locuteurs : par exemple, il n'est pas rare qu'une conjonction soit insérée en fin de groupe intonatif et suivie d'une pause silencieuse. Si syntaxiquement, la conjonction fait partie du groupe syntaxique situé à sa droite, intonativement, elle est rattachée au groupe syntaxique gauche. Ceci a un impact pragmatique puisque cette stratégie permet au locuteur de conserver la parole (Ferré [7]). Le groupe intonatif nous semble pour cette raison plus approprié pour rendre compte du découpage de l'oral et peut plus facilement être mis en relation avec des unités gestuelles que nous allons décrire dans la section 2.2. Loehr [14] a d'ailleurs montré qu'il existe un lien entre groupes intonatifs et gestualité co-verbale.

Ces annotations sur la parole ont ensuite été importées dans Anvil (logiciel d'annotation des fichiers vidéos, Kipp [10]) afin de pouvoir comparer les données verbales et les données gestuelles. Anvil présente également l'avantage d'imposer une structuration hiérarchique des données de type XML ce qui a un impact sur l'annotation des gestes présentée ci-dessous.

2.2. Annotations gestuelles

L'ensemble des gestes manuels des 6 locuteurs a été transcrit manuellement sur les 45 minutes de corpus (l'annotation de 3 heures de corpus est actuellement en cours). Outre la configuration de la main, le type de mouvement, etc, qui ne nous sont pas directement utiles ici, nous avons annoté le type de geste (d'après la typologie de McNeill [15], décrite plus bas) dans ce qui constitue la Phrase gestuelle (Kendon [8]), c'est-à-dire le geste dans sa globalité, depuis la mise en place des articulateurs (bras, mains, doigts) jusqu'au repos final ou jusqu'au début du geste suivant lorsque deux gestes sont enchaînés sans retrait des articulateurs (en ayant cependant à l'esprit que l'annotation gestuelle est moins précise que l'annotation de la parole puisqu'elle est basée sur un enregistrement comptant 24 images/seconde).

Toujours selon Kendon [8], la Phrase gestuelle se décompose en différentes phases que sont la préparation (mise en place des articulateurs), la réalisation du geste (stroke), une éventuelle tenue du geste, et la rétraction. Seule la phase de réalisation est nécessaire pour former une phrase gestuelle, les autres phases étant facultatives. Ces différentes phases ont également été annotées sur 45 minutes d'enregistrement.

La typologie des gestes manuels employée pour l'annotation se compose des types de geste suivants : les iconiques représentent une caractéristique physique d'un objet de discours ou miment des actions, les métaphoriques représentent des idées abstraites, les déictiques pointent vers un référent (spatial ou énonciatif), les emblèmes sont des gestes conventionnels, les battements des gestes de scansion du discours, et enfin, les adaptateurs des gestes d'auto-contact.

Parmi ces gestes, nous avons retenu les gestes iconiques uniquement, plus nombreux que les autres, soit 107 occurrences (nous avons écarté 18 gestes iconiques pour lesquels il n'était pas possible de déterminer un affilié lexical).

2.3. Affiliés lexicaux

En effet, s'il s'agit de mettre en relation les gestes manuels co-verbaux et la parole, il faut pouvoir être certain de mettre en relation des unités de nature comparable, d'où la notion d'affiliation lexicale sur laquelle repose l'article de Schegloff [18] et définie par Kipp et al. [11] comme : « The word or words deemed to correspond most closely to a gesture in meaning ». Si l'on considère les gestes iconiques, il apparaît que dans 85.6% des occurrences, il est possible de déterminer un affilié lexical dans une relation de redondance par rapport à la parole et correspondant à un mot comme dans les Figures 1 et 2.

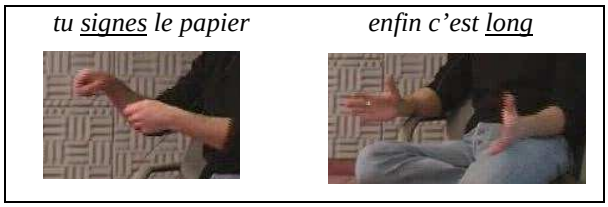

Figure 1 : Gestes iconiques correspondant aux affiliés lexicaux « signes » et « long ».

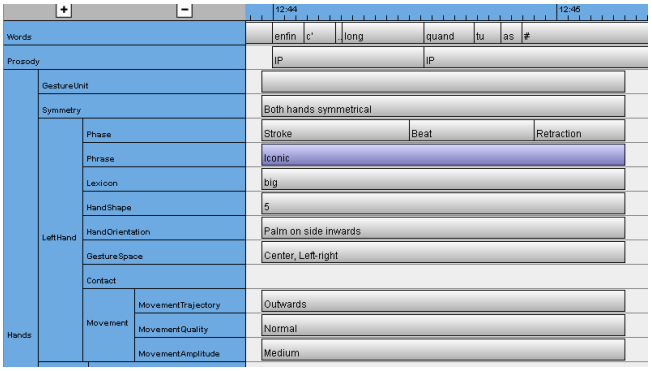

Figure 2 : Annotation Anvil correspondant à l'affilié lexical « long » dans « enfin c'est long quand tu as # tout le le la période de travail et tout ».

En ce qui concerne les autres catégories de gestes, soit elles contiennent très peu d'occurrences comme le cas des emblèmes, soit il est impossible de déterminer un affilié lexical précis comme dans le cas de nombreux métaphoriques qui apportent une modalité à tout l'énoncé comme dans la Figure 3. La relation sémantique entre geste et parole est alors implicite, or, pour pouvoir effectuer une comparaison en termes de temporalité, il faut pouvoir se baser sur une relation sémantique explicite entre geste et parole. C'est pour cette raison que nous avons choisi pour cette étude de ne retenir que les gestes iconiques, choix qui était également celui de Chui [6], alors que Kida & Faraco [9] et Loehr [14] ont travaillé sur différentes catégories de geste.

Figure 3 : Geste métaphorique produit sur « on n'en avait pas reparlé », qui permet d'apporter une modalité à l'énoncé mais pour lequel il est difficile de déterminer un affilié lexical précis.

3. RÉSULTATS

En ce qui concerne les résultats de cette étude, la première remarque que l'on peut faire est que les unités gestuelles (au niveau lexical, « Gstroke », et au niveau phrastique,

« Gphrase ») sont plus longues que les unités verbales correspondantes (mot et IP), même si l'écart entre unités phrastiques est moins important que l'écart entre unités lexicales.

En termes de relations temporelles (cf. pourcentages et écart moyen dans la Table 1), si l'on se place au niveau des unités lexicales en comparant le début et la fin du geste lui-même (« stroke ») au début et à la fin des tokens qui constituent les affiliés lexicaux, on constate qu'une large majorité de gestes (81.3%) commencent largement avant la production de l'affilié lexical en parole, et une proportion plus importante de gestes se terminent après la production de l'affilié lexical (61.7%) qu'avant celle-ci. Un Test T apparié montre que l'anticipation de la phase de réalisation sur l'affilié lexical en parole est significative (t=-7.85 ; p=1.73E-12). En revanche, on ne peut pas dire que le geste se termine avant ou après la parole de manière significative au niveau lexical (t=1.14 ; p=0.12). Ces statistiques ne s'expliquent donc pas uniquement par la durée plus importante du geste par rapport à la parole puisque le décalage temporel moyen est plus important dans le cas de l'anticipation du geste.

En ce qui concerne la relation temporelle entre les unités phrastiques (phrase gestuelle vs. IP), la tendance est la même, à savoir une anticipation de la gestualité sur la parole (60.75% des phrases gestuelles commencent avant les IP), mais le pourcentage est moins élevé que pour les unités lexicales. Egalement, 64.5% des phrases gestuelles se terminent après les IP avec une différence temporelle moyenne plus importante pour les gestes qui anticipent sur la parole. Pour le temps de début comme pour le temps de fin, le Test T apparié montre que le geste commence significativement avant la parole au niveau phrastique (t=-2.92 ; p=0.002) et se termine après la parole ((t=2.90 ; p=0.002). On notera toutefois qu'au niveau phrastique, la différence est moins importante qu'au niveau lexical. Si dans tous les cas rencontrés, il y avait nécessairement chevauchement entre la production des phrases gestuelles et la production des IP, lorsque le verbal anticipe sur le geste, le décalage temporel est plus important encore, de l'ordre d'une demi-seconde. Il faudrait donc regarder si dans ces cas précis, il n'y aurait pas d'hésitation au niveau de la production verbale.

Enfin, en ce qui concerne la comparaison entre la phrase gestuelle et l'affilié lexical, il est apparu que sur 107 iconiques, nous n'avons trouvé que 8 cas où la phrase gestuelle dans sa globalité était terminée avant la production de l'affilié lexical (nombre de ces cas contenaient des marques d'hésitation) alors que dans tous les autres cas, la phrase gestuelle et l'affilié lexical sont co-occurrents. Quant à la relation temporelle, 99% des phrases gestuelles commencent avant la production de l'affilié lexical (avec une différence hautement significative : t=-13.02 ; p=4.85E-24) et 85% d'entre elles se terminent après la production de l'affilié lexical (également de manière très significative : t=6.79 ; p=3.21E-10).

Table 1 : Pourcentage de gestes qui commencent ou finissent avant/après le verbal

	% de gestes qui commencent		**% de gestes qui se terminent**	
	avant la parole	**après la parole**	**avant la parole**	**après la parole**
Gstroke/ Affilié	**81,3**	**18,7**	**38,3**	**61,7**
Différence moyenne	0,566 s	0,14 s	0,44 s	0,391 s
Gphrase/IP	**60,75**	**39,25**	**35,5**	**64,5**
Différence moyenne	0,271 s	0,412 s	0,413 s	0,191 s
Gphrase /Affilié	**99**	**1**	**15**	**85**
Différence moyenne	0,76 s	0,098 s	0,578 s	0,804 s

4. DISCUSSION

Dans cette étude, nous avons présenté les résultats d'une des premières études portant spécifiquement sur le geste réalisées à partir du corpus CID. En effet, les récentes annotations gestuelles sur ce corpus nous ont permis de tester les relations temporelles entre gestualité co-verbale et parole dans le cas des gestes iconiques. Le choix de la catégorie gestuelle est justifié par la possibilité de déterminer pour ce type de geste un affilié lexical explicitement mentionné par les locuteurs.

Les résultats présentés dans ce travail – portant sur 107 gestes iconiques produits par 6 locuteurs pendant 45 minutes de français spontané – montrent que les relations temporelles qui existent entre la gestualité co-verbale et la parole, vont clairement dans le sens d'une anticipation du geste sur la parole. Mais si l'on considère les différentes études réalisées dans ce domaine, celles-ci affichent des résultats opposés. En effet, pour le chinois, Chui ([6]:878) a trouvé une plus grande proportion de gestes synchronisés avec la parole que de gestes anticipant la parole (60.1% vs 35.6%), avec des résultats semblables pour Loehr [14] sur l'anglais, toutes catégories de gestes confondues. En revanche, Schegloff [18], qui a travaillé sur les gestes déictiques en anglais, constate que les réalisations gestuelles (« strokes ») sont produites généralement de manière anticipée par rapport à leur affilié lexical. Leonard & Cummins [13], dans une récente étude, trouvent également une anticipation du geste sur la parole dans cette langue. Leur travail concernait plus précisément l'alignement des battements, décomposés en leurs différentes phases, avec l'affilié lexical. Ils ont montré – sur un corpus très réduit – que la phase de réalisation du battement anticipait sur l'onset de la voyelle dans l'affilié lexical correspondant. Ils ont aussi montré que le geste s'achève après la parole, comme dans notre corpus. Bourguet & Ando [5], sur les gestes déictiques en japonais, insistent plutôt sur la variabilité des relations temporelles entre geste et voix, et montrent qu'en fonction du type de déictique produit, le geste peut anticiper la

parole ou au contraire être produit après la parole. Enfin Kranstedt et al. [12], également sur les déictiques en anglais, montrent que le geste est produit avec un retard par rapport à la parole.

Devant une telle variabilité des résultats obtenus dans les différentes études, il convient de s'interroger sur les raisons de cette variabilité. Nous nous sommes tournés vers la réalisation du geste et notamment son amplitude qui pourrait agir sur les relations temporelles entre geste et parole. La tendance observée est une anticipation plus grande pour les gestes de grande amplitude et moins grande pour les gestes de petite amplitude. Mais ces observations de moyennes ne sont pas statistiquement significatives. Rochet-Capellan, et al. [17] remarquent que sur les déictiques produits en français et en portugais, la parole et la gestualité co-verbale tendent à l'isochronie, avec un décalage du geste afin que son apogée corresponde à la syllabe accentuée de l'affilié lexical. Nous n'avons pas pu vérifier cette tendance sur notre corpus, mais il est possible que la variabilité observée dans les études tienne à la nature de la relation geste / parole. En effet, la relation étudiée ici sur les gestes iconiques était une relation de redondance alors que les gestes déictiques dans d'autres travaux sont dans une relation de complémentarité avec la parole. On pourrait penser que dans le cas de la redondance, la synchronisation entre geste et parole est moins nécessaire que dans le cas de la complémentarité, et tend à être moins précise. Enfin, le type de corpus annoté (conversationnel vs. expérimental) pourrait également avoir un impact sur les relations temporelles entre gestes et parole.

BIBLIOGRAPHIE

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

[2] R. Bertrand, P. Blache, et al. Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *TAL*, 49(3): 105-133, 2008.

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

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

[5] M.-L. Bourguet and A. Ando. Synchronization of Speech and Hand Gestures during Multimodal Human-Computer Interaction. In Karat, C.-M., et al. (eds.), *Human Factors in Computing Systems, CHI 98.* Los Angeles, CA, ACM Press, 241-242, 1998.

[6] K. Chui. Temporal Patterning of Speech and Iconic Gestures in Conversational Discourse. *Journal of Pragmatics*, 37:871-887, 2005.

[7] G. Ferré. Les pauses démarcatives déplacées en anglais spontané : marquage prosodique et kinésique. *Lidil*, 26:155-169, 2002.

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

[9] T. Kida and M. Faraco. Prédication gestuelle. *Faits de Langues*, 31-32:217-226, 2008.

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

[11] M. Kipp, M. Neff, et al. An annotation Scheme for Conversational Gestures: How to Economically Capture Timing and Form. *Language Resources and Evaluation*, 41:325-339, 2007.

[12] A. Kranstedt, P. Kühnlein and I. Wachsmuth. Deixis in Multimodal Human Computer Interaction: An Interdisciplinary Approach. In A. C. Volpe (ed.) *Gesture-Based Communication in Human-Computer Interaction.* Berlin, Heidelberg, Springer-Verlag, 112-123, 2004.

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

[14] D. Loehr. *Gesture and Intonation.* PhD Thesis. Georgetown University, 2004.

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

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

[17] A. Rochet-Capellan, C. Vilain et al. Does the Number of Syllables Affect the Finger Pointing Movement in a Pointing-naming Task? *8th International Seminar on Speech Production.* Strasbourg, 257-260, 2008.

[18] E. A. Schegloff. On Some Gestures' Relation to Talk. In J. M. Atkinson & J. Heritage (eds.), *Structures of Social Action*. Cambridge, CUP, 266-298, 1984.

[19] E. Selkirk. On Prosodic Structure and its Relation to Syntactic Structure. In T. Fretheim (ed.), *Nordic Prosody II*. Trondheim, Tapir, 111-140, 1978.