Gesture/speech interaction in the perception of lexical units

Chloe Gonseth (chloe.gonseth@gipsa-lab.fr)
Anne Vilain (anne.vilain@gipsa-lab.fr)
Coriandre Vilain (coriandre.vilain@gipsa-lab.fr)
Grenoble University, Gipsa-Lab, Speech & Cognition Department (CNRS UMR 5216)
11 rue des Mathématiques, BP 46, F - 38402 Saint Martin d'Hères Cedex

Abstract

This paper explores gesture/speech interaction in language perception. An experimental study, based on an intermodal priming paradigm, required participants to make lexical judgments to deictic words, non-deictic words, or pseudo-words, after the production of a pointing or a grasping gesture. These two gestural priming conditions were compared to each other and to a baseline condition, where participants did not perform any gesture. This allowed us to characterize the influence of both gesture production and gesture type on word recognition. Our results reveal an interaction between the motor and the lexical representations of spatial deixis, that suggests that "*arm movement itself* [could] *be used as a linguistic signal*" (Gentilucci, Dalla Volta, & Gianelli, 2008). Communicative manual gestures appear to be involved in the production/perception mechanism associated with the semantic processing of language.

Keywords: Word recognition ; Spatial deixis; Gesture/speech interaction

Introduction

Several studies provide evidence for an integrated representation of speech and manual gestures during language production in the human linguistic brain (Kita & Ozyürek, 2003; Gonseth, Vilain, & Vilain, 2012a), even at a relatively early age in its development (Gonseth, Vilain, & Vilain, 2012b). However, the speech/gesture interaction in language perception remains relatively unknown. The nature of speech perception representations has been at the heart of strong debates for several years. These debates originally opposed auditory to motor theories, that respectively consider cognitive representations of speech as auditory or articulatory units. For a few years, new theories have emerged, according to which speech is of multimodal nature, involving both sensory and motor representations (Schwartz, Basirat, Ménard, & Sato, 2012; Hickok, Houde, & Rong, 2011). Many studies indeed showed that sensorimotor representations are involved in speech sounds perception (Sato et al., 2011; Fadiga, Craighero, Buccino, & Rizzolatti, 2002). Besides, the motor system appears to be involved in semantic units perception (Gentilucci, Benuzzi, Bertolani, Daprati, & Gangitano, 2000; Glenberg & Kaschak, 2002). For instance, Floël et al. demonstrated that perception of linguistic materials activates the motor cortices for both hands ; it thus seems that "*listening to "gestures" that compose spoken language [...] activate an extended articulatory/manual action-perception network*" (Flöel, Ellger, Breitenstein, & Knecht, 2003). In the same vein, Kelly et al. evidenced a mutual and obligatory interaction between speech and gesture in language comprehension, a tight relationship that "*may reflect the basic multimodal architecture of the human brain, which may be designed to optimally process and integrate information from across modalities*". The authors then proposed that gesture and speech form a single communication system, being "*simply two sides of the same coin : language*" (Kelly, Ozyürek, & Maris, 2010). In Gentilucci et al.'s terms, "*arm movement itself* [could] *be used as a linguistic signal*" (Gentilucci et al., 2008). In other words, the production of a communicative manual gesture should automatically activate the corresponding lexical representation. To test this hypothesis, we conducted an experimental study, where we tested whether gesture production could prime lexical access. We studied specifically the relation between deictic gestures and deictic words, in relation with the distance of the designated target.

Experimental study

Method

This experiment consisted in having participants make lexical judgments on auditorily presented items: deictic words vs non-deictic words vs pseudo-words. Item presentation could be preceded by different gestural priming conditions that may or may not facilitate word recognition: deictic gestures vs non deictic gestures vs no gesture. The distance of the designated target was also varied: close vs far. The two questions we were exploring were (i) whether gesture production induced a priming of related lexical items; (ii) whether this priming effect was distance specific.

Participants Thirty-three right-handed native French-speaking adults, 10 males and 23 females, aged 18-34 years, participated in the experiment. They all had normal or corrected-to-normal vision and they reported no auditory or motor disorders.

Speech material The stimuli set was composed of two targets, i.e. two French deictic words, either proximal or distal, *ici*/*here* and *là-bas*/*there*. We then selected two fillers via the Lexique database (New, Pallier, Brysbaert, & Ferrand, 2004), i.e. two French non-deictic words, *ami*/*friend* and *salut*/*hi*, and four French pseudo-words[1], [laɡa], [ini], [saʁy], [ati].
The stimuli were paired according to their first phoneme and chosen according to the following criteria: syllabic structure, mean word frequency (higher than 100 occurrences per million, i.e. high frequency), final phonological uniqueness point, neighborhood density (number of phonological neighbours higher than 10, i.e. high density).
Two trained native French speakers were chosen to produce

[1] French pseudo-words are meaningless strings of phonemes that respect the phonotactic rules of French.

the speech material. The stimuli were recorded in a soundproof room with a Marantz PMD 670 digital recorder in order to obtain wave files. Two versions of each stimulus were selected to avoid the negative effects of habituation. We thus obtained a sample of 32 stimuli (8 stimuli × 2 versions × 2 speakers), later approved by two naive individuals.

Procedure Participants were tested individually in a soundproof room. They had to decide as quickly as possible whether the stimulus, presented via AKG K58 headphones, was a French word or a pseudo-word. For that purpose, they were instructed to press, with their left fingers, a keyboard button corresponding to each of those answers (i.e. right/left arrow keys). This lexical decision task could be preceded by the following priming conditions:

- Related condition: Pointing. Participants were seated with two aligned light emitting diodes (LED), placed at 140cm (extra-personal close space) and 425cm (extra-personal far space) in front of them, both out of reach of the hand. They were required to point at the turned-on LED using a right index finger pointing until the diffusion of the stimuli.
- Unrelated condition: Grasping. Participants were seated with a wooden cylinder-shaped object (6cm high, with a diameter of 3cm) placed in front of them. They were required to grasp the object with their right hand until the diffusion of the stimuli. Note that none of the lexical stimuli was related with grasping.
- Baseline condition: No gesture. Participants did not perform any gesture before making their lexical judgments.

The length of the experiment was about 20 min, during which participants were presented with 288 trials (3 $Conditions \times$ 8 $Items \times$ 12 $Iterations$). We measured response times ($RT = T_{[button\ press\ initiation]} - T_{[stimulus\ presentation]}$) in milliseconds[2] and error rates (i.e. proportion incorrect).

Analyses and predictions

Only RTs for correct responses were analyzed, error rates being too weak to be analyzed. Data for five participants were excluded, due to high rates of failed trials (as failure to adhere to the instructions, e.g. grasping omissions). Ultimately, data for 28 participants were entered in the final analyses.

We conducted two repeated-measures ANalyses Of VAriance (ANOVA) on median values[3] of RTs, with a significance level fixed at p<.05. Data normality and data sphericity were measured, respectively with the Shapiro-Wilk test (Shapiro, 1965) and the Mauchly test (Mauchly, 1940). Data were adjusted if necessary, respectively with square-root transformations[4], the initial distribution following a Poisson distribution, and Greenhouse-Geisser corrections (Greenhouse & Geisser, 1959).

The first ANOVA was conducted with three fixed factors: $Speaker_{[S1\ -\ S2]} \times Condition_{[Pointing\ -\ Grasping\ -\ Baseline]} \times Item_{[Deictic\ -\ Non\ deictic-\ Pseudo\ word]}$. Our main hypothesis concerns the interaction between $Condition$ and $Item$. We expected a priming effect (i.e. faster RTs) only for deictic words in the pointing condition. In other words, we wanted to evidence an interaction between the motor and the lexical representations of spatial deixis.

The second ANOVA, dealing with the pointing condition, was conducted with three fixed factors: $Speaker_{[S1\ -\ S2]} \times Item'_{[Proximal\ deictic\ -\ Distal\ deictic\ -\ Non\ deictic-\ Pseudo\ word]} \times Distance_{[140cm\ -\ 425cm]}$. Our main hypothesis concerns the interaction between $Distance$ and $Item'$. We expected RTs to be faster when gesture and speech conveyed congruent information, in terms of spatial content. Thus, a pointing gesture directed to a close (resp. distant) object should activate preferentially the congruent lexical representation, i.e. the proximal deictic term (resp. distal), as compared with the incongruent ones, i.e. distal deictic terms (resp. proximal), non-deictic words, and pseudo-words. In other words, we wanted to evidence a specific rather than a generic interaction between the motor and the lexical representations of distance encoding.

Results

For both analyses, the results depend on the nature of the stimuli rather than on the specificity of the speaker, since none of the effects involving this factor (i.e. main effect and interaction effects) is significant ($p > .05$).

ANOVA 1 : RTs as a function of *Item* **and** *Condition* We found no significant effect of the $Condition$ on RTs ($F(2,52) = 2.37$, $p = .10$, NS). Answers are not delayed in the gestural priming conditions compared to the baseline condition despite the higher cognitive load they must require. We found a significant effect of the $Item$ ($F(2,52) = 8.58$, $p < .01$), illustrated Figure 1. Post hoc comparisons indicate that words in general (i.e. deictic words and non-deictic words) are correctly classified faster than pseudo-words. This result could be interpreted as a Word Superiority Effect, i.e. WSE (Cutler, Mehler, Norris, & Segui, 1987).

More importantly, the interaction between Condition and Item (see Figure 2) reaches the significant threshold ($F(4,104) = 2.42$, $p = .05$).

Post hoc comparisons do not indicate any inter-item difference in the baseline condition but faster RTs for targets in the pointing condition and faster RTs for fillers in the grasping condition. In other words, the WSE seems to be related to the pronounceability of the stimuli rather than to their lexical nature; it only occurs when the task calls up a larger part of cognitive resources (i.e. only in grasping and pointing conditions). Furthermore, the production of a deictic gesture preferentially activates deictic words, as compared with pseudo-words. In other words, the activation of a motor deictic repre-

[2] Response times higher than 2000 ms or lower than 100 ms were removed.

[3] We choose median values instead of mean values as a way to exclude outliers. The median is a robust value, less affected by outliers and skewed data than the mean is.

[4] $Y' = \sqrt{Y}$, where Y' is the final distribution and Y the initial one.

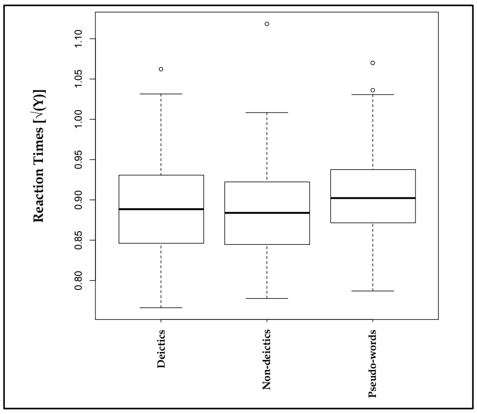

Figure 1: Square-root transformed RTs as a function of *Item*

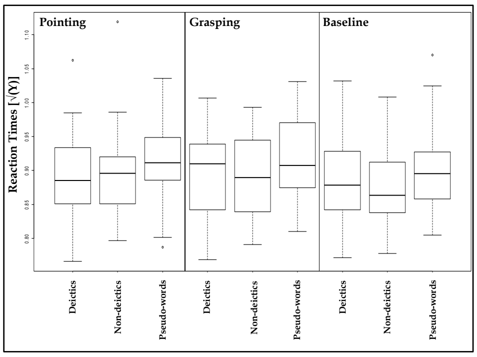

Figure 2: Square-root transformed RTs as a function of $Condition * Item$

sentation automatically induces that of the corresponding lexical representation. Interestingly, the production of a grasping gesture preferentially activates non-deictic words, as compared with pseudo-words. This result could be explained by an inhibition mechanism, or Stroop effect (Stroop, 1935): the production of a grasping gesture inhibits the recognition of unrelated items (i.e. deictic words), thus resulting in slower RTs.

ANOVA 2 : RTs as a function of $Item'$ **and** $Distance$ We found a significant but unexpected main effect of $Distance$ ($F(1,26) = 15.42$, $p < .01$), illustrated Figure 3. RTs are indeed faster when pointing gestures are directed to close objects, as compared to distant ones. This result could be explained by the spatial arrangement of the two LEDs. It is potentially more difficult to designate the distant LED, aligned with the close one, which induces a decrease in the cognitive resources allocated to the lexical judgment.

We also found a significant effect of $Item'$ ($F(3,78) = 6.25$, $p < .01$), illustrated Figure 4. Post hoc comparisons show that both proximal and distal deictics are identified faster than pseudo-words. This result confirms the previous one, according to which the production of a pointing gesture automatically activates the lexical representation of pointing. However, proximal deictics are identified faster than distal deictics. It is possible that the instructions given to the participants did not indicate clearly enough that a compound word could be considered as a single lexical unit, resulting in delayed RTs to accept *là-bas* as a French word.

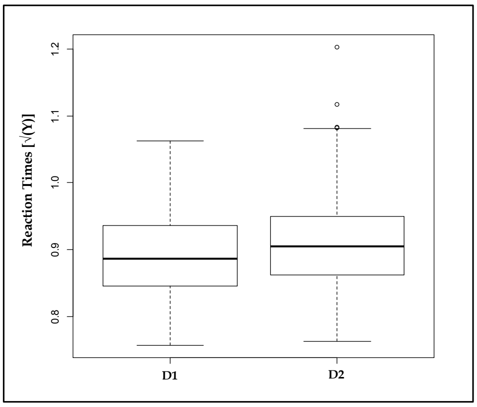

Figure 3: Square-root transformed RTs as a function of $Distance$

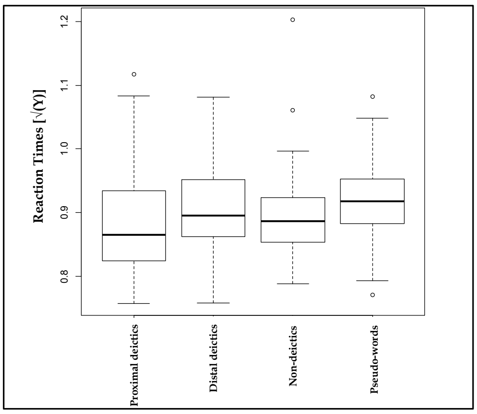

Figure 4: Square-root transformed RTs as a function of $Item'$

Finally, the interaction between $Distance$ and $Item'$ is not significant ($F(3,78) = 1.04$, $p = .38$, NS). Proximal deictics tend to be identified faster when the pointing gesture is directed to the close LED compared to the distant one, but this tendency does not reach the significant threshold. This result might be explained by the weaker proportion of related targets in this analysis, as compared with the first analysis (12.5% of related targets instead of 25%).

Conclusion

Our results are in favour of an interaction between manual and vocal systems in the perception of lexical units: motor representations of pointing automatically activate the corresponding linguistic representations, regardless of the spatial content of each modality. Speech perception, phonologically constrained by articulatory gestures production as proposed by Schwartz et al. (Schwartz et al., 2012), appears to be lexically constrained by manual gesture production. Speech perception thus relies on the activation of sensorimotor units of multimodal nature. The speech/gesture interaction then appears to take place at a high level in language processing. It would be interesting to confirm these data by comparing structurally similar gestures, in terms of motor configuration, that carry different functions, that is of different nature, i.e. communicative or not (for instance, a pointing and a grasping gesture directed to a reachable object). A similar priming effect for both gestures would be in favour of a low level mechanism whereas a more important priming effect for the communicative gesture would indicate a high level (i.e. lexical) mechanism.

To date, the exact role of motor representations in language processing is still highly controversial: are the vocal and the manual systems systematically and functionally linked in acquisition, comprehension, and production of language, as proposed by Glenberg and Gallese (Glenberg & Gallese, 2012), or rather structurally linked, the motor system not being directly involved in language processing (Hickok, 2010; Mahon & Caramazza, 2008; Meteyard, Cuadrado, Bahrami, & Vigliocco, 2012)? As Coello and Bartolo put it, "*these issues represent original and fascinating challenges in the field of cognitive neurosciences and would obviously viewed as important avenues for the research on action and language in the future*"' (Coello & Bartolo, 2012).

References

Coello, Y., & Bartolo, A. (2012). Contribution of the action system to language perception and comprehension : Evidence and controversies. In Y. Coello & A. Bartolo (Eds.), *Language and action in cognitive neuroscience* (p. 321-342). London: Psychology Press.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology*, *19*, 141-177.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience*, *15*, 399-402.

Flöel, A., Ellger, T., Breitenstein, C., & Knecht, S. (2003). Language perception activates the hand motor cortex: Implications for motor theories of speech perception. *European Journal of Neuroscience*, *18(3)*, 704-8.

Gentilucci, M., Benuzzi, F., Bertolani, L., Daprati, E., & Gangitano, M. (2000). Language and motor control. *Experimental Brain Research*, *133*, 468-490.

Gentilucci, M., Dalla Volta, R., & Gianelli, C. (2008). When the hands speak. *Journal of Physiology*, *102*(1-3), 21–30.

Glenberg, A., & Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex*, *48(7)*, 905-922.

Glenberg, A., & Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, *9*, p. 558-565.

Gonseth, C., Vilain, A., & Vilain, C. (2012a). An experimental study of speech/gesture interactions and distance encoding. *Speech Communication*, In Press.

Gonseth, C., Vilain, A., & Vilain, C. (2012b). Ontogeny of two communicative tools: Distance encoding and multimodality in deictic pointing. In J. Hurford, M. Tamariz, E. Cartmill, & T. Scott-phillips (Eds.), *The Evolution of Language: Proceedings of the 9þ International Conference (EVOLANG9)* (p. 150-157). Kyoto, Japan: World Scientific.

Greenhouse, S., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*, 95-112.

Hickok, G. (2010). The role of mirror neurons in speech perception and action word semantics. *Language and Cognitive Processes*, *25*, 749-776.

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: Computational basis and neural organization. *Neuron*, *69(3)*, 407-422.

Kelly, S., Ozyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21(2)*, 260-267.

Kita, S., & Ozyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*(1), 16–32.

Mahon, B., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology*, *102*, 59-70.

Mauchly, J. (1940). Significance test for sphericity of a normal *n*-variate distribution. *The Annals of Mathematical Statistics*, *11*, 204-209.

Meteyard, L., Cuadrado, S., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, *48(7)*, 788-804.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments & Computers: A Journal of the Psychonomic Society Inc.*, *36*, 516-524.

Sato, M., Grabski, K., Glenberg, A., Brisebois, A., Basirat, A., Ménard, L., et al. (2011). Articulatory bias in speech categorization: Evidence from use-induced motor plasticity. *Cortex*, *47(8)*, 1001-1003.

Schwartz, J., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, *25(5)*, 336-354.

Shapiro, S. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591-611.

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643-662.