No Evidence of Theory of Mind Reasoning in the Human Language Network

C. Shain[1], A. Paunov[2], X. Chen[3], B. Lipkin[1] E. Fedorenko [1]MIT, [2]NeuroSpin, [3]Rice

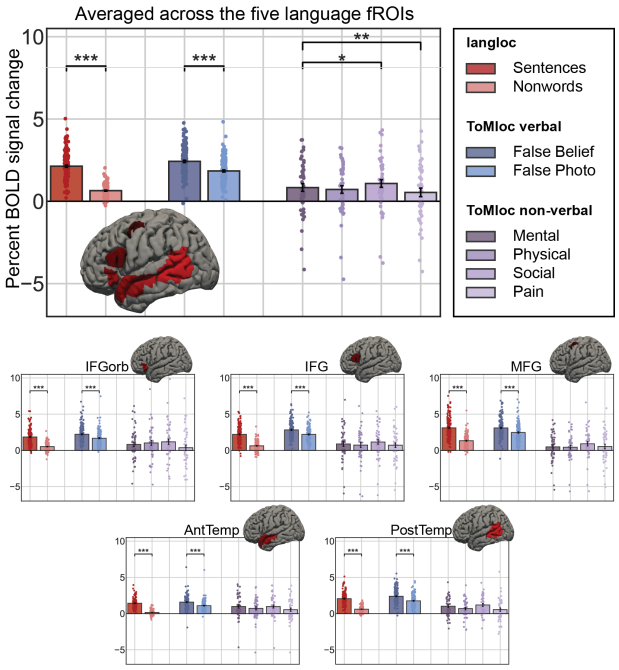
Language comprehension and the ability to infer others' thoughts (theory of mind, ToM) are core cognitive functions. Each of them is supported by a specialized brain network [7, 18], and they are closely related during both development [1, 3, 15] and language use [8, 20, 17]. This conceptual relationship between language and ToM abilities suggests the possibility of shared neural circuitry between them. However, neural evidence is mixed as to the relationship between language and ToM functions in the brain. Although robust dissociations between language and ToM have been reported in brain disorders [6, 5, 21], brain activations for contrasts that target language and ToM bear similarities, and there is direct evidence that language-responsive brain areas also engage in ToM reasoning [4].

Here we revisit the language-ToM relationship in a large-sample (151-participant) fMRI study by evaluating the response of the language network [7] during both a standard verbal ToM task [18] and a recently-developed non-verbal ToM task that has been validated in multiple prior studies [9, 16, 10]. In the verbal task, participants read short vignettes describing (i) characters with false beliefs (FB, +ToM) or (ii) false photographs depicting non-existent physical scenes (FP, –ToM). In the non-verbal task, participants watched a short video in which segments were coded for content: *mental* (segments likely to elicit mental state attribution; e.g., a character falsely believes they have been abandoned by a companion), *social* (segments depicting non-mental social interactions), *pain* (segments depicting characters in physical pain), and *physical* (segments depicting non-social physical events). In brain areas that support ToM, the mental condition should elicit more activation than the other three conditions. Five core left-hemisphere (LH) language responsive areas—inferior frontal gyrus (IFG) and its orbital part (IFGorb), middle frontal gyrus (MFG), anterior temporal lobe (AntTemp), and posterior temporal lobe (PostTemp)—are identified using a functional regions of interest (fROI) approach [7], which critically allows us to independently identify the language network in individual brains and then study its activity during the ToM tasks.

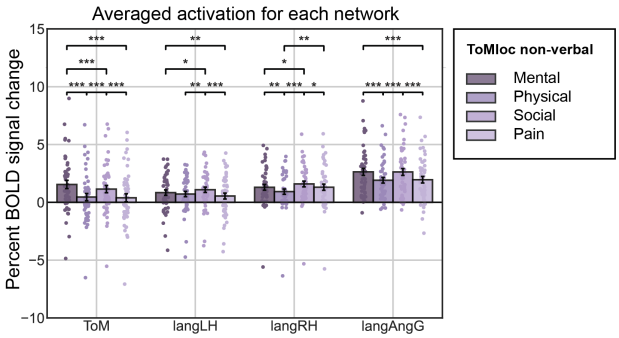
Analyses (**Fig 1a**) reveal that all core language regions respond more strongly when participants read vignettes about false beliefs compared to the control vignettes. However, no such effects appear in a non-verbal ToM task, and controlling for independently-supported linguistic confounds (e.g., surprisal [11] and dependency locality [19]) between the two conditions of the verbal ToM task greatly attenuates (by 84%) the language network's ToM response. Together, these results do not support the existence of ToM reasoning in the language network, and suggest that prior reports of ToM effects in the language network may have been driven by linguistic confounds like surprisal and dependency locality.

We additionally explore ToM responses in the "periphery" of the language network [2], namely language-responsive areas in the right hemisphere (RH) homotopes of the core LH language regions, and in the bilateral angular gyri, since prior work has shown that these areas are functionally distinct from the core LH language network but closely related to it [13, 14, 12]. We find evidence (**Fig 1b**) of social processing in this periphery (FB $>$ FP, mental $>$ physical, social $>$ physical), but no evidence of ToM-specificity in particular (mental $\not>$ social).

These results argue against cognitive and neural overlap between language processing and ToM, and clarify plausible sources (linguistic confounds) of previous reports to the contrary. Nevertheless, the language and ToM networks likely communicate with each other, as suggested both by the likely importance of ToM for pragmatic inference in language processing [8] and by evidence of significantly greater synchrony between the language and ToM networks than between either network and domain-general executive areas [13]. Analyses of the language network's periphery suggest that it might play a role in this communication pathway through its engagement in both linguistic and social processing, which may be a promising direction for future research.



(a) Responses to the conditions of the language localizer task, and verbal and non-verbal ToM tasks in the language network.



(b) Responses by network to the four conditions of the non-verbal ToM localizer (Mental, Physical, Social, and Pain).

References

[1] Astington, J. W. and Jenkins, J. M. *Developmental psychology*, 1999.

[2] Chai, L. R., Mattar, M. G., Blank, I. A., Fedorenko, E., and Bassett, D. S. *Cerebral Cortex*, 2016.

[3] de Villiers, J. G. and de Villiers, P. A. *Topics in Language Disorders*, 2014.

[4] Deen, B., Koldewyn, K., Kanwisher, N., and Saxe, R. *Cerebral cortex*, 2015.

[5] Diehl, J. J., Bennetto, L., and Young, E. C. *Journal of abnormal child psychology*, 2006.

[6] Dronkers, N. F., Ludy, C. A., and Redfern, B. B. *Journal of Neurolinguistics*, 1998.

[7] Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., and Kanwisher, N. *Journal of Neurophysiology*, 2010.

[8] Grice, H. P. Logic and conversation. 1975.

[9] Jacoby, N., Bruneau, E., Koster-Hale, J., and Saxe, R. *Neuroimage*, 2016.

[10] Kamps, F. S., Richardson, H., Murty, N. A. R., Kanwisher, N., and Saxe, R. *Human Brain Mapping*, 2022.

[11] Lopopolo, A., Frank, S. L., den Bosch, A., and Willems, R. M. *PloS one*, 2017.

[12] Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffman, M., Mineroff, Z., Jouravlev, O., and Fedorenko, E. *bioRxiv*, 2022.

[13] Paunov, A., Blank, I., and Fedorenko, E. *Journal of neurophysiology*, 2019.

[14] Paunov, A., Blank, I. A., Jouravlev, O., Mineroff, Z., Gallée, J., and Fedorenko, E. *bioRxiv*, 2022.

[15] Richardson, H., Koster-Hale, J., Caselli, N., Magid, R., Benedict, R., Olson, H., Pyers, J., and Saxe, R. *Nature communications*, 2020.

[16] Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., and Saxe, R. *Nature communications*, 2018.

[17] Roberts, C. *Semantics and pragmatics*, 2012.

[18] Saxe, R. and Kanwisher, N. *Neuroimage*, 2003.

[19] Shain, C., Blank, I. A., Fedorenko, E., Gibson, E., and Schuler, W. *Journal of Neuroscience*, 2022.

[20] Sperber, D. and Wilson, D. *Behavioral and brain sciences*, 1987.

[21] Willems, R. M., Benn, Y., Hagoort, P., Toni, I., and Varley, R. *Neuropsychologia*, 2011.