Evidence for a unitary semantic gateway for words and pseudowords across languages

Fabio Marson[1], Rolando Bonandrini[1], Simona Amenta[1], Marco Marelli[1]

[1] *Department of Psychology, University of Milano-Bicocca, Italy*

Background: We can have a sense of the meaning of an unknown word by the context in which it appears or by examining familiar elements in its orthographic form. In fact, recent studies showed that pseudowords (i.e., strings of letters not attested in any language) carry semantic information through their sublexical pattern [1–3]. Reaction times (RTs) during lexical decision were found to be modulated by pseudowords'estimated semantic load, with slower RTs signalling the interference of semantic information. However, it is unclear whether the activation of semantic information during pseudowords and words processing are supported by a shared mechanism: semantic effects in pseudowords might reflect strategic, metalinguistic processes that are inherently different from the automatic semantic access elicited by familiar stimuli. To test this hypothesis, we analyzed the RTs obtained from a series of crowdsourcing studies and tested whether individual differences in sensitivity to semantic information could modulate the activation of semantic information in pseudowords processing. Additionally, we further tested the robustness of such effect by analyzing megastudies from the same family of projects in different languages: English, Italian, Spanish, and Dutch [4–7].

Method: Reaction times (RTs) for words and pseudowords were extracted from the crowdsourcing projects, in which thousands of participants completed a vocabulary test. A distributional semantic model that capitalizes on sublexical elements (i.e., FastText) was used to extract semantic neighborhood density (SND; the number of semantically related words) and orthography-to-semantic consistency (OSC; the amount of semantic information shared by orthographically similar words) for all of the words and pseudowords included in the analyses. We first fit a linear model for each of the participants and extracted the beta values of OSC and SND as predictors of reaction times for words. These estimates were used as measures of sensitivity to semantics at the individual level. Then, we fit linear mixed models testing whether the individual measures of sensitivity to semantics (SemSens-OSC and SemSens-SND) would possibly interact with the effect of the estimated semantic load of pseudowords (OSC-pw and SND-pw) in predicting RTs for pseudowords processing.

Results: We observed significant interactions between individual sensitivity to semantics and semantic load of pseudowords in the English Crowdsourcing Project ($p < .001$). In detail, RTs resulted to be slower when pseudowords had higher SND-pw and OSC-pw, and these interference effects were even stronger in those participants that exhibited higher sensitivity to semantic measures for words. Similar results were observed for the crowdsourcing project in Italian, Spanish, and Dutch. The very same results were observed for SND across all languages. For what concerns OSC, Italian and Dutch showed a significant interaction that goes in the same direction of the English dataset, despite OSC-pw seems to have an overall facilitatory effect in the Dutch sample. OSC-pw by SemSens-OSC interaction in Spanish only approached significance ($p = .054$) even though the direction of the effect was consistent with that of the other languages.

Discussion: The results show that pseudowords processing is not only affected by their own semantic load, in line with current literature, but it is further modulated by individual measures of sensitivity to semantic information. Since these latter measures of sensitivity have been estimated from responses to words suggests that the mechanism that underlies attribution of semantic information to pseudowords and words could be the same. The interfering effect of semantic neighbors was found to be the most robust, while a some cross-linguistic variability was observed about OSC. This difference could be due to differences in orthographic depth between languages. Overall, we interpret these results as evidence of a shared mechanism of semantic attribution to strings of letters regardless of their lexicality status. This result further opens the discussion on what defines lexicality and the extent to which semantics could be a property emerging from distributional patterns of (sub)word usage.

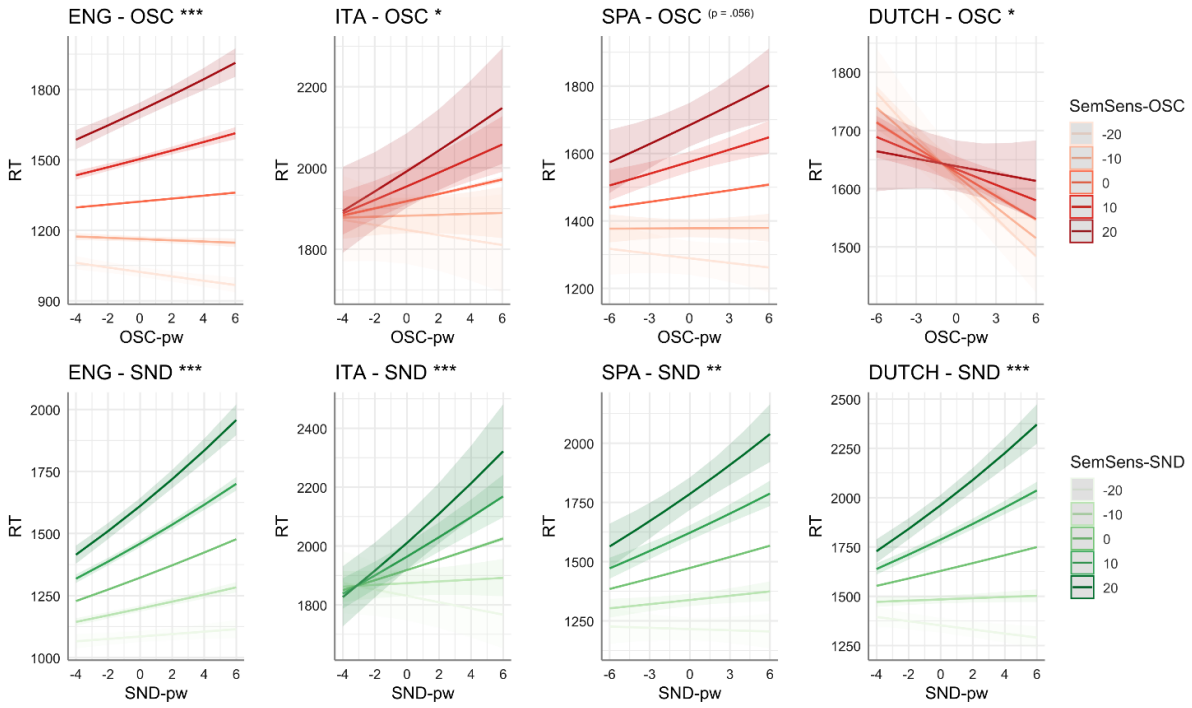

Figure 1: Interaction between semantic sensitivity to semantics and semantic load measures during pseudowords processing for OSC (upper panels) and SND (lower panels) across different languages (from left to right: English, Italian, Spanish, Dutch). Abbreviations: SemSens-OSC = individual sensitivity to semantics in relation to OSC; SemSens-SND = individual sensitivity to semantics in relation to SND; OSC-pw = orthography-to-semantic consistency for pseudowords; SND-pw = semantic neighborhood density for pseudowords. (Significance levels: *** = p < .001; ** = p < .01; * = p < .05.)

References

[1] Hendrix, P., & Sun, C. (2021). A word or two about nonwords: Frequency, semantic neighborhood density, and orthography-tosemantics consistency effects for nonwords in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(1), 157. https://doi.org/10.1037/xlm0000819

[2] Gatti, D., Marelli, M., & Rinaldi, L. (2023). Out-of-vocabulary but not meaningless: Evidence for semantic-priming effects in pseudoword processing. *Journal of Experimental Psychology: General*, *152*(3), 851–863. https://doi.org/10.1037/xge0001304

[3] Bonandrini, R., Amenta, S., Sulpizio, S., Tettamanti, M., Mazzucchelli, A., & Marelli, M. (2023). Form to meaning mapping and the impact of explicit morpheme combination in novel word processing. *Cognitive Psychology*, *145*, 101594. https://doi.org/10.1016/j.cogpsych.2023.101594

[4] Mandera, P., Keuleers, E., & Brysbaert, M. (2020). Recognition times for 62 thousand English words: Data from the English Crowdsourcing Project. *Behavior Research Methods*, *52*(2), 741–760. https://doi.org/10.3758/s13428-019-01272-8

[5] Amenta, S., Varda, A. D., Mandera, P., Keuleers, E., Brysbaert, M., & Marelli, M. (2025). The Italian Crowdsourcing Project: Visual word recognition times for 130,495 Italian words. *Behavior Research Methods*, *57*(1), 1–18. https://doi.org/10.3758/s13428-024-02548-4

[6] Aguasvivas, J., Carreiras, M., Brysbaert, M., Mandera, P., Keuleers, E., & Duñabeitia, J. (2020). How do Spanish speakers read words? Insights from a crowdsourced lexical decision megastudy. *Behavior Research Methods*, *52*(5), 1867–1882. https://doi.org/10.3758/s13428-020-01357-9

[7] Brysbaert, M., Keuleers, E., & Mandera, P. (2019). Recognition times for 54 thousand Dutch words: Data from the Dutch Crowdsourcing Project. *Psychologica Belgica*, *59*(1), 281. https://doi.org/10.31234/osf.io/f8cj3