Not just form, not just meaning: Consistency in form-meaning mappings predicts age of acquisition beyond semantic and form neighborhood density

Giovanni Cassani and Niklas Limacher (Tilburg University)
g.cassani@tilburguniversity.edu

We extend previous work on the relation between word form and lexical semantics (Form-Semantic-Consistency, [1, 2]) in language processing by documenting a unique negative effect of FSC on English Age of Acquisition (AoA, [3]), with more systematic words being learned earlier. This effect holds after controlling for neighborhood density measured for word forms and lexical semantics [4], their interaction, and several other standard predictors of AoA including iconicity [5, 6], using linear regression and random forests.

First, we established a baseline statistical model where AoA is predicted as a linear combination of standard covariates, including frequency (derived from SUBTLEX, [7]), concreteness [8], morphological complexity [9], valence [10], and Semantic Neighborhood Density (SND). Moreover, when targeting orthographic forms we included word length in letters and OLD20 [11] among the covariates. When we targeted phonological representations, however, we included word length in phonemes and Coltheart's N [12], taken from the Massive Auditory Lexical Decision database [13]. The analysis focused on words for which all predictors were available: 8,566 words for the orthographic analysis and 6,855 for the phonological analysis. We then added an interaction term between form and semantic neighborhood density to the baseline model, which proved significant: words in denser form neighborhoods tend to have higher AoA when they are in sparser semantic neighborhoods.

In order to ensure the relation between FSC and AoA is not a simple by-product of an interaction between neighborhood density in form and semantic space (each shown to predict AoA across a number of typologically different languages [4]), we determined whether a model including FSC measures explained more variance in AoA than the baseline model enriched with the interaction term. Results indicated that FSC captures something more than the interaction between neighborhood density in form and semantics. This pattern was robust to the form encoding (orthography or phonology) as well as to the specific implementation of FSC [1, 14], although phonology had a generally higher predictive power. See the appendix for further details about the computational modelling framework.

Moreover, we ensure that the effect of FSC on AoA is not an epiphenomenon of other properties of the lexicon by running 1000 simulations on random pairings of word form and lexical meaning. We then measure the correlation between random FSC measures and AoA as well the difference in AIC between a model including all covariates and one also including random FSC measures. Across the board, we find that only true FSC measures explain unique variance in AoA, although simple pairwise correlations between AoA and random FSC measures are non-zero in certain configurations suggesting that systematicity is intertwined with other properties of the lexicon. A correlation matrix suggests that FSC is collinear with frequency, length, and measures of neighborhood density in form space.

We also show that FSC's effect is not reducible to iconicity [6]: when controlling for iconicity norms [15] in the linear regression, we still observe that FSC measures improve model fit, although their unique effect is reduced. We ran this analysis separately due to the limited coverage of the available iconicity norms, which resulted in 2,032 and 1,795 words for the orthographic and phonological analysis respectively. As a last sanity check, we re-ran all analyses using random forests to ensure that no documented effect was due to collinearity issues [16]: all patterns are confirmed, and we noticed that FSC measures tend to have rather high variable importance, suggesting they play a prominent role in predicting AoA.

To sum up, these analyses confirm that systematicity relates to language acquisition [17] while also showing that only true systematicity captures more than a simple interaction between network structure in form and meaning and is robust to encoding and implementation choices. Finally, while FSC and iconicity tap into similar aspects of AoA, they remain distinct, and both explain a significant portion of variance in acquisition patterns.

Computational Modeling Framework

FSC reflects the degree to which words that formally resemble a target word also share its semantics, such that the local similarity structure of the target word is similar in the space defined by word form as well as by semantics. Operationally, in order to compute FSC for target word $t \in V$ (V being the vocabulary), one starts by finding the most similar words $v_{1,\ldots,|V|} \in V$ in form space, then fetches their semantic representations and computes their similarity to the semantic representation of *t*.

The first operationalization of FSC [1] defined neighbors in form space as words which started with the target, e.g. *redemption* is a neighbor of *red* but *credibility* is not. This criterion was then relaxed to consider all words which embed the target as neighbors. An alternative way of determining neighbors in form-space is through more general distance functions, such as edit distance [18]. This approach was used by Hendrix and Sun [14], who considered the 5 words with smallest edit distance to the target as neighbors in form space. We take the same approach, with a slight modification in case more words share the same distance as the 5th closest word. Whereas Hendrix and Sun sampled at random among words at the same distance, we make the selection flexible and pick all words at the same distance as the 5th closest word. Results were robust to changes in this hyper-parameter.

Next to defining a way to find neighbors in form space, it is essential to decide on which semantic representation to adopt, and which metric to compute the similarity between target and form-based neighbors in semantic space. In this respect, we used word embeddings and cosine similarity [14], leveraging the space provided by Mandera and colleagues [19], who extracted 300-dimensional embeddings using the Continuous Bag of Word (CBoW) algorithm from *word2vec* [20] using a window of 6 words to the left and right of the target and a large corpus consisting of subtitles and internet pages in English.

Formally, the first implementation of FSC [1], defined target-embedded (FSC_{te}), is computed following equation (1), where *t* is the target word, $n \in N$ is a neighbor from the set of neighbors in form space which embed the target word, *f* is the corpus frequency, *cos*(•,•) is the cosine similarity function, while $\boldsymbol{t}$ and $\boldsymbol{n_i}$ are the embeddings of the target and neighbors. The semantic similarity between target and neighbor is thus weighted by the frequency of the neighbor, such that more frequent neighbors weigh more.

$$\mathrm{FSC_{te}(t)} = \frac{\sum_{i=1}^{N} \cos(\mathbf{t},\mathbf{n_i}) * \mathrm{f_n}}{\sum_{i=1}^{N} \mathrm{f_n}} \quad (1)$$

The second implementation of FSC [14], defined levenshtein-distance (FSC_{ld}), is computed following equation (2). Most symbols are shared with equation (1) and retain the meaning they had there. The new symbol, d_i, indicates the Levenshtein distance between the target *t* and neighbor n_i. This time, however, the semantic similarity is weighted by the distance in form space, such that closer neighbors weigh more on the measure of coherence. The resulting score is finally normalized by the number of neighbors.

$$\mathrm{FSC_{ld}(t)} = \frac{\sum_{i=1}^{N} \cos(\mathbf{t},\mathbf{n_i}) * 1/\mathrm{d_i}}{N} \quad (2)$$

References:

1. Marelli, M., S. Amenta, and D. Crepaldi, *Semantic transparency in free stems: The effect of Orthography-Semantics Consistency on word recognition.* Quarterly Journal of Experimental Psychology, 2015. **68**(8): p. 1571-83.
2. Amenta, S., M. Marelli, and S. Sulpizio, *From sound to meaning: Phonology-to-Semantics mapping in visual word recognition.* Psychonomic Bulletin Review, 2017. **24**(3): p. 887-893.
3. Kuperman, V., H. Stadthagen-Gonzalez, and M. Brysbaert, *Age-of-acquisition ratings for 30 thousand English words.* Behav Res Methods, 2012. **44**(4): p. 978-90.
4. Fourtassi, A., Y. Bian, and M.C. Frank, *The Growth of Children's Semantic and Phonological Networks: Insight From 10 Languages.* Cognitive Science, 2020. **44**(7): p. e12847.
5. Braginsky, M., et al., *Consistency and Variability in Children's Word Learning Across Languages.* Open Mind (Camb), 2019. **3**: p. 52-67.
6. Nielsen, A. and M. Dingemanse, *Iconicity in Word Learning and Beyond: A Critical Review.* Language and Speech, 2020: p. 23830920914339.
7. van Heuven, W.J.B., et al., *Subtlex-UK: A New and Improved Word Frequency Database for British English.* Quarterly journal of experimental psychology, 2014. **67**(6): p. 1176-1190.
8. Brysbaert, M., A.B. Warriner, and V. Kuperman, *Concreteness ratings for 40 thousand generally known English word lemmas.* Behavior Research Methods, 2014. **46**(3): p. 904-11.
9. Sanchez-Gutierrez, C.H., et al., *MorphoLex: A derivational morphological database for 70,000 English words.* Behavior Research Methods, 2018. **50**(4): p. 1568-1580.
10. Warriner, A.B., V. Kuperman, and M. Brysbaert, *Norms of valence, arousal, and dominance for 13,915 English lemmas.* Behavior Research Methods, 2013. **45**(4): p. 1191-207.
11. Yarkoni, T., D.A. Balota, and M. Yap, *Moving beyond Coltheart's N: a new measure of orthographic similarity.* Psychonomic Bulletin and Review, 2008. **15**(5): p. 971-9.
12. Coltheart, M., et al., *Phonological Encoding in the Lexical Decision Task.* Quarterly Journal of Experimental Psychology, 2018. **31**(3): p. 489-507.
13. Tucker, B.V., et al., *The Massive Auditory Lexical Decision (MALD) database.* Behavior research methods, 2018.
14. Hendrix, P. and C.C. Sun, *A word or two about nonwords: Frequency, semantic neighborhood density, and orthography-to-semantics consistency effects for nonwords in the lexical decision task.* Journal of Experimental Psychology: Learning, Memory, and Cognition, 2020.
15. Perry, L.K., M. Perlman, and G. Lupyan, *Iconicity in English and Spanish and Its Relation to Lexical Category and Age of Acquisition.* PLoS One, 2015. **10**(9): p. e0137147.
16. Tomaschek, F., P. Hendrix, and R.H. Baayen, *Strategies for addressing collinearity in multivariate linguistic data.* Journal of Phonetics, 2018. **71**: p. 249-267.
17. Monaghan, P., et al., *How arbitrary is language?* Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 2014. **369**(1651): p. 20130299.
18. Levenshtein, V.I., *Binary codes capable of correcting deletions, insertions, and reversals.* Soviet physics. Doklady, 1965. **10**: p. 707-710.
19. Mandera, P., E. Keuleers, and M. Brysbaert, *Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation.* Journal of Memory and Language, 2017. **92**: p. 57-78.
20. Mikolov, T., et al., *Distributed representations of words and phrases and their compositionality*, in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. 2013, Curran Associates Inc.: Lake Tahoe, Nevada. p. 3111–3119.