Modeling Morphological Processing in Human Magnetoencephalography

Yohei Oseki
Faculty of Science & Engineering
Waseda University
oseki@aoni.waseda.jp

Alec Marantz
Department of Linguistics & Psychology
New York University
marantz@nyu.edu

Abstract

In this paper, we conduct a magnetoencephalography (MEG) lexical decision experiment and computationally model morphological processing in the human brain, especially the Visual Word Form Area (VWFA) in the visual ventral stream. Five neurocomputational models of morphological processing are constructed and evaluated against human neural activities: Character Markov Model and Syllable Markov Model as "amorphous" models without morpheme units, and Morpheme Markov Model, Hidden Markov Model (HMM), and Probabilistic Context-Free Grammar (PCFG) as "morphous" models with morpheme units structured linearly or hierarchically. Our MEG experiment and computational modeling demonstrate that "morphous" models outperformed "amorphous" models, PCFG was most neurologically accurate among "morphous" models, and PCFG better explained nested words with non-local dependencies between prefixes and suffixes. These results strongly suggest that morphemes are represented in the human brain and parsed into hierarchical morphological structures.

1 Introduction

Under the single-route decomposition model of morphologically complex visual word recognition (Taft, 1979, 2004; Taft and Forster, 1975), there are three functionally different stages of morphological processing: morphological decomposition, lexical access, and morphological recombination. In the first stage of morphological decomposition, morphologically complex words are visually decomposed into component morphemes. In the second stage of lexical access, meanings of decomposed morphemes are lexically retrieved from the mental lexicon. In the last stage of morphological recombination, retrieved meanings of decomposed morphemes are semantically composed.

In the cognitive neuroscience literature, Fruchter and Marantz (2015) employed magnetoencephalography (MEG) to spatiotemporally dissociate those stages of morphological processing. Specifically, the first stage of morphological decomposition has been indexed by evoked response components such as M170 (Zweig and Pylkkänen, 2009; Solomyak and Marantz, 2010; Lewis et al., 2011; Fruchter et al., 2013; Gwilliams et al., 2016) or Type II (Tarkiainen et al., 1999; Helenius et al., 1999) in the visual ventral stream of the human brain (Pylkkänen and Marantz, 2003; Hickok and Poeppel, 2007). Moreover, Dehaene et al. (2005) proposed local combination detectors (LCDs) where linguistic units such as characters, syllables, and morphemes are convolutionally represented and processed in the visual ventral stream from posterior occipital to anterior temporal cortices and, importantly, morphemes have been localized to the left fusiform gyrus known as the Visual Word Form Area (VWFA; Cohen et al., 2000, 2002; Dehaene et al., 2001, 2002). For example, Solomyak and Marantz (2010) and Lewis et al. (2011) computed transition probabilities from stems to suffixes (e.g. $P(\text{Suffix}|\text{Stem})$) to successfully predict neural responses to real (e.g. *teach-er*) and pseudo (e.g. *corn-er*) bimorphemic words, respectively. These results have suggested that morphemes may be neurologically real in the human brain.

However, "amorphous" models without morpheme units have recently been proposed in the morphological processing literature (Baayen et al., 2011; Virpioja et al., 2017). For instance, Baayen et al. (2011) and Milin et al. (2017) proposed Naive Discriminative Learning (NDL), a connectionist model with direct mappings from forms to meanings, to explain morphological processing without morpheme units. In addition, Virpioja et al. (2017) and Hakala et al. (2018) employed

Proceedings of the Society for Computation in Linguistics (SCiL) 2020, pages 209-219.
New Orleans, Louisiana, January 2-5, 2020

Morfessor, an unsupervised finite-state model with statistically induced "morphs" (Creutz and Lagus, 2007), to predict human reaction times and neural responses without linguistically defined morphemes. Furthermore, as correctly pointed out by Libben (2003, 2006), bimorphemic words exclusively tested in the previous literature (Zweig and Pylkkänen, 2009; Solomyak and Marantz, 2010; Lewis et al., 2011) cannot distinguish linear morphological decomposition from hierarchical morphological parsing (cf. Song et al., 2019; Oseki et al., 2019). Therefore, whether morphemes are represented in the human brain and, if so, processed linearly or hierarchically remains to be empirically investigated.

In this paper, we conduct an magnetoencephalography (MEG) experiment where participants perform visual lexical decision on morphologically complex words and, generalizing the computational modeling technique developed in the sentence processing literature (Frank et al., 2015; Brennan et al., 2016), computationally model morphological processing in the human brain, with special focus on the VWFA in the visual ventral stream. Specifically, five neurocomputational models of morphological processing are constructed and evaluated against human neural activities: Character Markov Model and Syllable Markov Model as "amorphous" models without morpheme units, and Morpheme Markov Model, Hidden Markov Model (HMM), and Probabilistic Context-Free Grammar (PCFG) as "morphous" models with morpheme units structured linearly or hierarchically.

2 Methods

2.1 Participants

The participants were 26 native English speakers recruited at New York University. All participants were right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971) and with normal or corrected-to-normal vision. They provided written informed consent and were paid $15/hour for their participation. We excluded 6 participants based on their behavioral performance: 3 participants excluded due to low accuracy ($< 75\%$) and 3 participants excluded due to slow (> 2000 ms) or fast mean reaction times (< 500 ms). Thus, 20 participants were included in the statistical analyses (10 males and 10 females, $M = 28.4$, $SD = 9.27$).

2.2 Stimuli

The stimuli were 800 morphologically complex trimorphemic words and nonwords. The stimuli creation procedure consisted of several steps. First, 600 trimorphemic words were created based on the CELEX database (Baayen et al., 1995) in accordance with syntactic (syntactic categories), morphological (affix combinations), and phonological (orthographic adjustments) selectional restrictions of derivational affixes, but without semantic selectional restrictions explicitly taken into consideration. In this sense, these trimorphemic words are grammatical ("possible") but not necessarily acceptable ("actual") words (cf. Halle, 1973; Bauer, 2014). These 600 trimorphemic words were subcategorized into 300 linear words [$_X$ [$_Y$ [$_Z$ $\sqrt{\text{Root}}$] Suffix] Suffix] with productive derivational suffixes (Plag and Baayen, 2009) and 300 nested words [$_X$ Prefix [$_Y$ [$_Z$ $\sqrt{\text{Root}}$] Suffix]] with productive derivational prefixes (Zirkel, 2010). Furthermore, these trimorphemic words have zero surface frequencies in the CELEX database, thereby enhancing the possibility that those words have never been encountered by participants and stored in the mental lexicon (Hay, 2003). Second, in order to weed out semantically implausible words, 600 trimorphemic words were normed with crowdsourced acceptability judgment experiments, where participants judged them on 1$\sim$7 Likert scale. Third, 500 trimorphemic words (250 linear and 250 nested) with higher acceptability judgments (> 3.5) and lower standard deviations (< 2.5) were selected and, correspondingly, 500 trimorphemic nonwords (250 linear and 250 nested) were also created based on the CELEX database in violation of syntactic selectional restrictions of inner derivational suffixes, resulting in 1000 trimorphemic words and nonwords. Fourth, in order to ensure that words and nonwords are correctly judged as such, 1000 trimorphemic stimuli were further normed with crowdsourced lexical decision experiments, where participants decided whether presented stimuli were possible English words or not as quickly and accurately as possible. Finally, 400 trimorphemic words (200 linear and 200 nested) and 400 trimorphemic nonwords (200 linear and 200 nested) with higher accuracies ($> 75\%$) were selected, resulting in the balanced and extensively normed set of 800 trimorphemic stimuli to be tested in this experiment. The stimuli are summarized in Table 1:

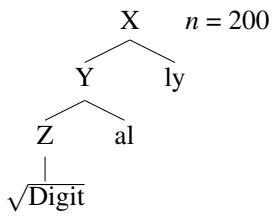
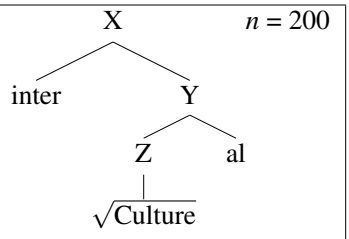
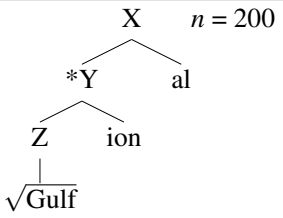
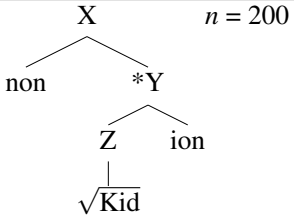
	Linear	Nested
Word		
Nonword		

Table 1: Summary of stimuli. The horizontal dimension is morphological structure: linear vs. nested. The vertical dimension is lexicality status: word vs. nonword. The asterisk (*) on subtrees (Y) of nonwords indicates that inner derivational suffixes violate syntactic selectional restrictions on syntactic categories of roots.

2.3 Procedure

The experiment was conducted in the Neuroscience of Language Lab at New York University, New York. Before MEG recording, each participant's head shape was digitized with a Polhemus FastSCAN laser scanner (Polhemus, Vermont, USA) and five fiducial points were marked on his/her forehead, onto which marker coils were attached during the recording. In order to familiarize the participants with visual lexical decision, the participants completed one practice block with 16 practice stimuli, 4 stimuli per each stimulus type, that do not overlap with the target stimuli. The task instructions were exactly the same as the main experiment, but the participants received feedback after each trial ("CORRECT" or "INCORRECT") during the practice block.

A 157-channel axial gradiometer whole-head MEG system (Kanazawa Institute of Technology, Kanazawa, Japan) recorded the MEG data continuously at a sampling rate of 1000 Hz (1 datapoint per each millisecond), while the participants lay in a dimly lit magnetically shielded room (MSR) and performed visual lexical decision. The MEG data were filtered online between DC and 200 Hz with a notch filter at 60 Hz. Five marker coils were attached to the corresponding fiducial points marked on the forehead and their positions were measured before and after the main experiment, in order to align the MEG data and head shapes and estimate how much the participants moved during the MEG recording. The main experiment itself lasted for about 35 minutes.

The stimuli were presented with PsychoPy package (Peirce, 2007, 2009) in Python. They were projected on the screen approximately 50 cm away from the participants and presented in white 30 lowercase Courier New font on a grey background. The 800 stimuli were randomly distributed into 8 blocks of 100 stimuli with 25 stimuli from each stimulus type. First, the explanation appeared on the screen: "In this experiment, you will read English words and determine whether you think they are possible English words. We are not concerned with whether or not these words are actual English words already listed in a dictionary. Instead, we are interested in whether or not these words could be used by a native speaker of English". Then, the task instruction appeared on the screen: "The experiment is about to begin. Please fixate on the cross in the center of the screen. Respond with your index finger if the string is word. Respond with your middle finger if it is not a word". Each trial consisted of the fixation cross (+) for 500 ms, the blank for 300 ms, and the stimulus until the participants respond with their index finger (YES) or middle finger (NO) of their left hand. The inter-stimulus interval (ISI) followed the standard normal distribution with the mean of 400 ms and the standard deviation of 100 ms.

2.4 Computational models

Five computational models were implemented with Natural Language Tool Kit package (Bird et al., 2009) in Python: Character Markov Models (Character), Syllable Markov Models (Syllables), Morpheme Markov Models (Markov), Hidden Markov Model (HMM), and Probabilistic Context-Free Grammar (PCFG). Those models were trained on the entire CELEX database via Maximum Likelihood Estimation with token weighting and Lidstone smoothing at $\alpha = 0.1$. The architectures of Markov Model, HMM, and PCFG are summarized below.

2.4.1 Markov Model

Markov Models (also called *n*-gram models) are defined by *n*-order Markov processes that compute transition probabilities of linguistic units (e.g. characters, syllables, morphemes) at position *i* given *i–n* context (e.g. $P(x_i|x_{i-n}, x_{i-1})$). Since the length of morphologically complex words is inherently limited relative to syntactically complex sentences, Markov Models were defined with $n = 1$ (i.e. bigram models), which compute transition probabilities of linguistic units at position *i* given the immediately preceding unit (e.g. $P(x_i|x_{i-1})$). For training, Markov Models were trained on character strings (Character Markov Model), syllable strings (Syllable Markov Model), and morpheme strings (Morpheme Markov Model), respectively, where character and morpheme strings were available from the CELEX database, while syllable strings were generated with `syllabify` module implemented in Python by Kyle Gorman through ARPABET transcriptions assigned by LOGIOS Lexicon Tool in the Carnegie Mellon University Pronouncing Dictionary. For testing, those trained Markov Models then computed morpheme probabilities of morphologically complex words equivalent to their transition probabilities given the Markov assumption. Markov Models are linear models, which should accurately predict local dependencies of linear words (e.g. *digitally*), but not non-local dependencies of nested words (e.g. *unpredictable*) because local dependencies (e.g. **unpredict*) are unattested in the training data.

2.4.2 Hidden Markov Model

HMMs generalize Markov Models with *n*-order Markov processes defined over "hidden" linear strings. HMMs compute transition probabilities of part-of-speech (POS) tags at position *i* given *i–n* context (e.g. $P(t_i|t_{i-n}, t_{i-1})$), and emission probabilities of morphemes at position *i* given POS tags at the same position *i* (e.g. $P(m_i|t_i)$). Like Markov Models, HMMs were also defined with *n* = 1, which compute transition probabilities of POS tags at position *i* given the immediately preceding POS tag (e.g. $P(t_i|t_{i-1})$). For training, HMMs were supervisedly trained on tagged morpheme strings generated from morphological structures available from the CELEX database (e.g. [(*digit*, N), (*al*, A), (*ly*, B)]). For testing, those trained HMMs then computed morpheme probabilities of morphologically complex words as the ratio of prefix probabilities at position *k* to position *k*–1, where prefix probabilities are the sum of path probabilities compatible with morphemes until position *k* (Rabinar, 1989). HMMs are linear models, which should accurately predict local dependencies of linear words (e.g. N-A-B for *digitally*), but also non-local dependencies of nested words (e.g. *unpredictable*) if component local dependencies (e.g. A-V for **unpredict*) are attested in the training data.

2.4.3 Probabilistic Context-Free Grammar

PCFGs generalize Context-Free Grammars (CFGs) with probability distributions defined over hierarchical structures. PCFGs compute nonterminal probabilities of right-hand sides given left-hand sides of nonterminal production rules (e.g. $P(rhs|lhs)$), and terminal probabilities of right-hand side terminals given left-hand side nonterminals of terminal production rules (e.g. $P(m_i|t_i)$), equivalent to HMM emission probabilities. Nonterminal production rules are head-lexicalized, which model syntactic selectional restrictions of derivational affixes (e.g. N → A *ness*). For training, PCFGs were supervisedly trained on morphological structures available from the CELEX database (e.g. [$_B$ [$_A$ [$_N$ *digit*] *al*] *ly*]). For testing, those trained PCFGs then computed morpheme probabilities of morphologically complex words as the ratio of prefix probabilities at position *k* to position *k*–1, where prefix probabilities are the sum of tree probabilities compatible with morphemes until position *k* (Earley, 1970; Stolcke, 1995). PCFGs are hierarchical models, which should accurately predict not only local dependencies of linear words (e.g. *digitally*), but also non-local dependencies of nested words (e.g. *unpredictable*).

2.5 Evaluation metrics

The information-theoretic complexity metric, *surprisal*, was employed as linking hypothesis that bridges the gap between representation and processing (Hale, 2001; Levy, 2008). Surprisal of morpheme *m*, $I(m)$, is defined as Equation (1):

$$I(m) = \log_2 \frac{1}{P(m)} = -\log_2 P(m) \quad (1)$$

where $P(m)$ is the probability of morpheme m computed by computational models via respective incremental algorithms. Surprisal was originally proposed to explain behavioral measures such as reading times in self-paced reading experiments and fixation durations in eye-tracking experiments (Boston et al., 2008; Demberg and Keller, 2008; Roark et al., 2009; Frank and Bod, 2011; Fossum and Levy, 2012). Recently, surprisal has also been extended to neural measures like N400 components in EEG experiments and BOLD signals in fMRI experiments (Frank et al., 2015; Brennan et al., 2016; Willems et al., 2016; Henderson et al., 2016; Nelson et al., 2017; Lopopolo et al., 2017).

Assuming further that morphological processing is incremental (cf. prefix stripping; Taft and Forster, 1975; Stockall et al., 2019), we compute surprisal of morphologically complex words as *cumulative surprisal*, the cumulative sum of surprisal of component morphemes. Cumulative surprisal of word *w*, $I(w)$, is defined as Equation (2):

$$I(w) = I(m_1, ..., m_n) = \sum_{i=1}^{n} I(m_i) \quad (2)$$

where $I(m)$ is the surprisal of morpheme m computed by computational models.

Two evaluation metrics are then derived from cumulative surprisal: neurological and error accuracies (cf. Frank et al., 2015; Sprouse et al., 2018). The neurological accuracy of model *M*, $NA(M)$, is defined as Equation (3):

$$NA(M) = D_B - D_M \quad (3)$$

where D_B and D_M are deviance defined as –2 times log-likelihoods of baseline and target models, respectively. Neurological accuracy quantifies decreases in deviance ($-\Delta D$) and evaluates how well computational models explain human neural activities beyond control predictors included in the baseline model (cf. Frank et al., 2015).

The error accuracy of model *M*, $EA(M)$, is defined as Equation (4):

$$EA(M) = \sum_{i=1}^{n} |\epsilon_B(w_i)| - |\epsilon_M(w_i)| \quad (4)$$

where $\epsilon_B(w)$ and $\epsilon_M(w)$ are residual errors of baseline and target models for word w, respectively. Error accuracy quantifies decreases in absolute residual errors ($-\Delta|\epsilon|$) and evaluates cost-benefit tradeoffs of computational models (cf. Sprouse et al., 2018). We compute error accuracies of computational models with respect to linear and nested morphological structures to address the question whether hierarchical models make better predictions for nested words than linear models.

2.6 Statistical analyses

We performed linear mixed-effects regression (Baayen et al., 2008) by averaging neural activities within the functionally defined region of interest (fROI) based on spatiotemporal cluster permutation regression (Maris and Oostenveld, 2007). In the previous literature (cf. Gwilliams et al., 2016), *lemma frequency* has been proposed as a significant predictor of the M170 and, thus, employed as the predictor of interest for spatiotemporal regression. Lemma frequency (cf. del Prado Martin et al., 2004) is defined as the sum of frequencies of words that share the same lemma. For example, the lemma frequency of *globalization* is the sum of frequencies of *globe*, *global*, *globalize*, and so on. Spatiotemporal regression in the left inferior temporal lobe and the 150-200 time window with log-transformed lemma frequency as target predictor and squared length as control predictor identified the significant cluster where the clear M170 peak can be observed, as shown in Figure 1. Finally, the neural activities were averaged over space and time within the fROI to compute by-trial dSPMs (Dale et al., 2000), which were then exported to R for mixed-effects regression.

Linear mixed-effects regression was implemented with `lme4` package (Bates et al., 2015) in R. The baseline regression model was first fitted with by-trial dSPMs as the dependent variable, control predictors as fixed effects, and by-subject and by-word random intercepts as random effects. For each computational model, the target regression model was then fitted with cumulative surprisal included as an additional fixed effect on top

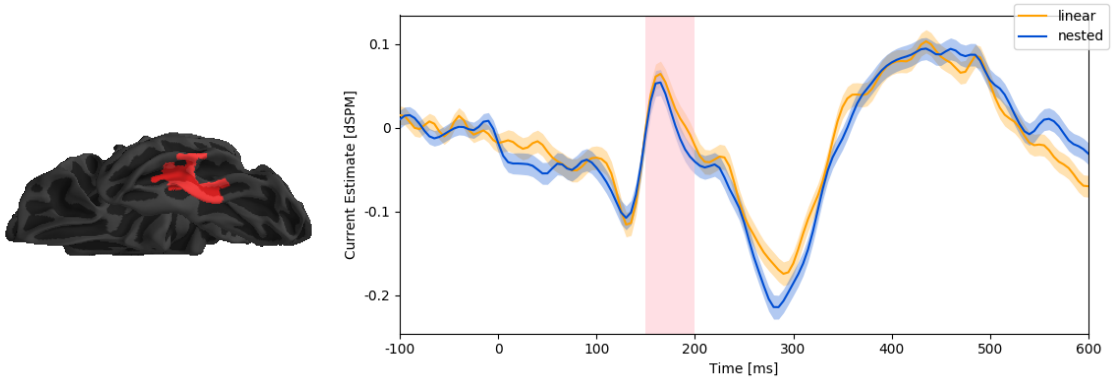


Figure 1: fROI for linear mixed-effects regression. Left: spatial extent defined as the significant cluster identified via spatiotemporal regression in the left inferior temporal lobe and the 150-200 time window with log-transformed lemma frequency as target predictor and squared length as control predictor; Right: temporal extent averaged over the significant cluster and categorized by linear and nested morphological structures. The *x*-axis is time in milliseconds, while the *y*-axis is neural activities in dSPM (Dale et al., 2000). Color indicates two morphological structures: yellow = linear, blue = nested. Pink vertical span marks the 150-200 ms time window.

of control predictors and random effects held constant. The control predictor was squared length (New et al., 2006) also included to functionally define the ROI. Mixed-effects models were fitted via Maximum Likelihood Estimation with `nlminb` optimizer in `optimx` package and the maximum number of iterations R permits. Given that the baseline and target models are minimally different only in cumulative surprisal, computational models can be evaluated with nested model comparisons via log-likelihood ratio tests based on χ^2-distribution with $df = 1$, where df is the difference in number of parameters between nested models.

3 Results

3.1 Neurological accuracy

Neurological accuracies of computational models are summarized in Figure 2, where the *x*-axis is computational models and the *y*-axis is neurological accuracies (i.e. decreases in deviance). The horizontal dashed line is $\chi^2 = 3.84$, the critical χ^2-statistic at $p = 0.05$ with $df = 1$.

Nested model comparisons via log-likelihood ratio tests revealed that while no "amorphous" models were statistically significant, all "morphous" models were statistically significant ($p < 0.01$). Among those "morphous" models, PCFG was most neurologically accurate: PCFG ($\chi^2 = 8.48$, $p < 0.01$) > Markov Model ($\chi^2 = 8.15$, $p < 0.01$) > HMM ($\chi^2 = 6.92$, $p < 0.01$) > Character ($\chi^2 = 0.19$, *ns*) > Syllable ($\chi^2 = 0.02$, *ns*).

3.2 Error accuracy

Error accuracies of computational models are summarized in Figure 3, where the *x*-axis is computational models and the *y*-axis is error accuracies (i.e. decreases in absolute residual errors), categorized into linear and nested morphological structures and averaged across individual derivational affixes. The horizontal dashed line indicates a "tie" borderline where computational models do not diverge from the baseline model. More positive and negative error accuracies mean better and worse predictions relative to the baseline model.

For linear words, all neurologically accurate "morphous" models made significant contributions, among which Markov Model made best predictions relative to the baseline model. For nested words, interestingly, PCFG was the only computational model which reduced residual errors, while linear models such as HMM and Markov Model made only slight or even worse predictions relative to the baseline model, respectively.

4 Discussion

In summary, our MEG experiment and computational modeling demonstrated that "morphous" models of morphological processing outperformed "amorphous" models and, importantly, PCFG was most neurologically accurate among those "morphous" models. We can conclude from these results that morphemes are neurologically represented in the human brain (pace Baayen

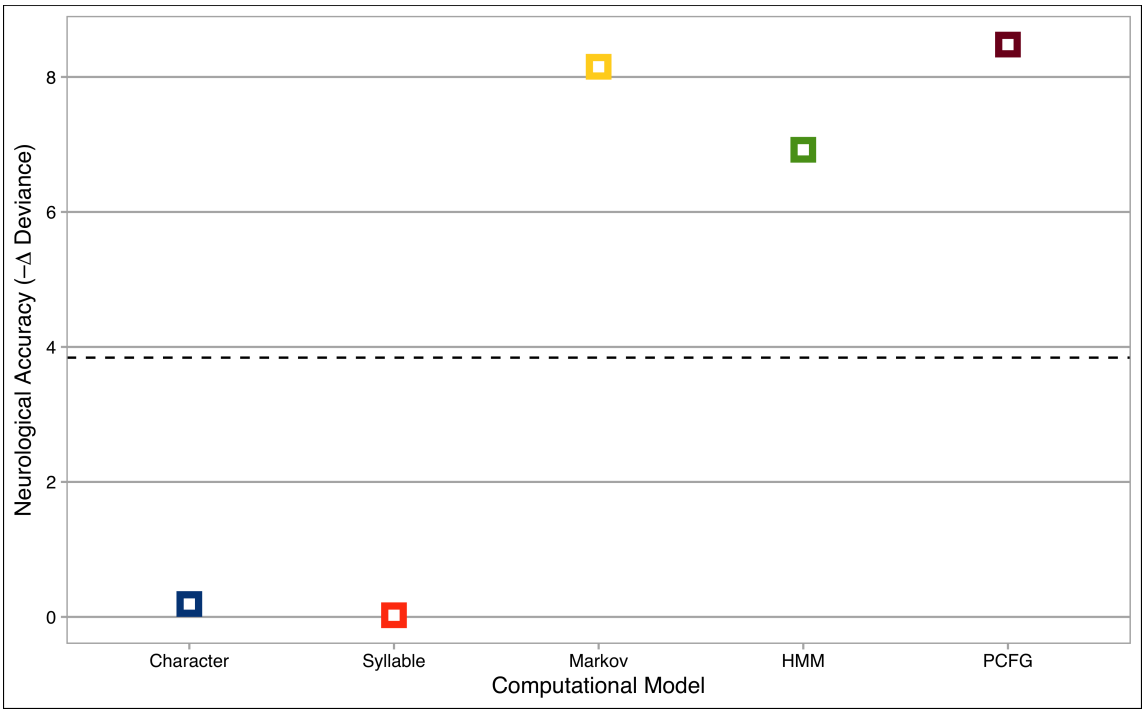


Figure 2: Neurological accuracies of computational models. The x-axis is computational models, while the y-axis is neurological accuracies (i.e. decreases in deviance). Points represent computational models: blue = Character Markov Model, orange = Syllable Markov Model, yellow = Morpheme Markov Model, green = Hidden Markov Model, purple = Probabilistic Context-Free Grammar. The horizontal dashed line is $\chi^2 = 3.84$, the critical χ^2-statistic at $p = 0.05$ with $df = 1$. All "morphous" models were statistically significant ($p < 0.01$).

et al., 2011; Milin et al., 2017) and parsed into hierarchical morphological structures (pace Virpioja et al., 2017; Hakala et al., 2018). In addition, this paper successfully generalized the computational modeling technique developed in the sentence processing literature (Frank et al., 2015; Brennan et al., 2016) to morphological processing.

Moreover, error accuracies of computational models indicated that PCFG better explained nested words with non-local dependencies between prefixes and suffixes than linear models such as Markov Model and HMM. This result follows straightforwardly from formal language theory, where linear and nested words are finite-state and context-free languages in the Chomsky hierarchy (Hopcroft and Ullman, 1979; Partee et al., 1990; Sipser, 1997), the former of which can be modeled by both linear and hierarchical models, but the latter of which can only be parsed by hierarchical models like PCFG. Furthermore, from the probabilistic perspective, linear models have trouble with transition probabilities from prefixes to roots in nested words (e.g. *unpredictable*) because prefixes (e.g. *un-*) and roots (e.g. *predict*) form no morphological constitutes (e.g. **unpredict*) and thus never appear in the training data.

Now the theoretical question arises why low-level visual evoked response components like M170 in the visual ventral stream "know" high-level linguistic representations like abstract hierarchical structures. One possibility is that, given the functional connectivity between the left fusiform gyrus and the left inferior frontal gyrus in visual word recognition (Pammer et al., 2004), M170 can be modulated in a top-down feedback manner by "Broca's area", the traditional "language" area proposed to process abstract hierarchical structures (Friederici, 2002, 2012). This possibility becomes even less surprising if visual cortex can be sensitive to abstract hierarchical structures (Dikker et al., 2009). Therefore, the functional connectivity between the left fusiform and inferior frontal gyri remains to be empirically investigated in the future research (Carreiras et al., 2014; Woodhead et al., 2014).

Nevertheless, there are several limitations with our computational modeling. One of the several important issues is that "amorphous" models in-

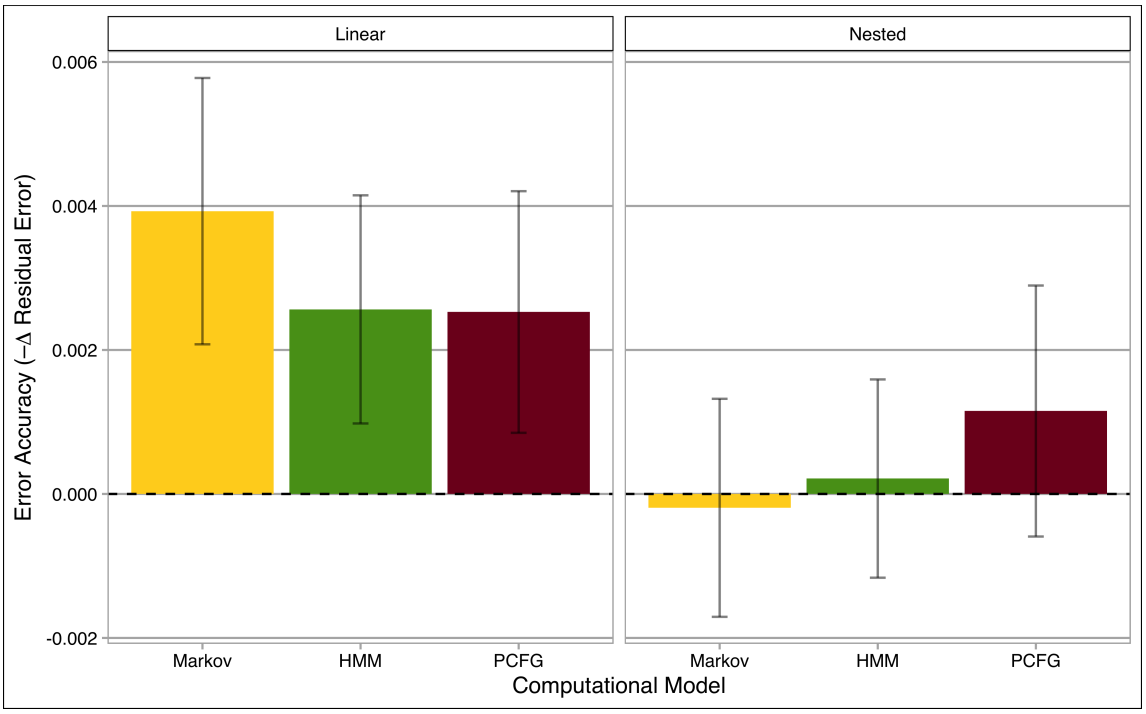


Figure 3: Error accuracies of computational models. The *x*-axis is computational models, while the *y*-axis is error accuracies (i.e. decreases in absolute residual errors), categorized into linear (Left) and nested (Right) morphological structures and averaged across individual derivational affixes. The horizontal dashed line indicates a "tie" borderline where computational models do not diverge from the baseline model, and more positive and negative error accuracies mean better and worse predictions relative to the baseline model.

vestigated in this paper are too simplistic as compared to computational models recently proposed in the morphological processing literature such as *Naive Discriminative Learning* (Baayen et al., 2011; Milin et al., 2017) or *Linear Discriminative Learning* (Baayen et al., 2018, 2019). Those state-of-the-art computational models of morphological processing remain to be constructed and evaluated against human neural activities and computational models investigated in this paper.

5 Conclusion

In this paper, we conducted a magnetoencephalography (MEG) experiment where participants performed visual lexical decision on morphologically complex words and, generalizing the computational modeling technique developed in the sentence processing literature (Frank et al., 2015; Brennan et al., 2016), computationally modeled morphological processing in the human brain, with special focus on the VWFA in the visual ventral stream. Five neuro-computational models of morphological processing were constructed and evaluated against human neural activities in order to investigate whether morphemes are neurologically represented in the human brain and parsed into hierarchical morphological structures: Character Markov Model and Syllable Markov Model as "amorphous" models without morpheme units, and Morpheme Markov Model, Hidden Markov Model (HMM), and Probabilistic Context-Free Grammar (PCFG) as "morphous" models with morpheme units structured linearly or hierarchically. Our MEG experiment and computational modeling demonstrated that "morphous" models of morphological processing outperformed "amorphous" models, PCFG was most neurologically accurate among those "morphous" models, and PCFG better explained nested words with non-local dependencies between prefixes and suffixes. These results strongly suggest that morphemes are neurologically represented in the human brain and parsed into hierarchical morphological structures. In conclusion, neuro-computational modeling of natural language must be a promising future direction in the cognitive computational neuroscience of language (Kriegeskorte and Douglas, 2018; Naselaris et al., 2018).

Acknowledgments

We would like to thank three anonymous reviewers of the *Society for Computation in Linguistics* and the members of the Neuroscience of Language Lab at New York University for valuable comments and suggestions. This work was supported by JSPS KAKENHI Grant Numbers JP18H05589 and JP19H04990 (YO) and the NYU Abu Dhabi Institute Grant Number G1001 (AM).

References

Harald Baayen, Yu-Ying Chuang, , and James Blevins. 2018. Inflectional morphology with linear mappings. *The Mental Lexicon*, 13:232–270.

Harald Baayen, Douglas Davidson, and Douglas Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412.

Harald Baayen, Petar Milin, Dusica Filipovic Durdevic, Peter Hendrix, and Marco Marelli. 2011. An Amorphous Model for Morphological Processing in Visual Comprehension Based on Naive Discriminative Learning. *Psychological Review*, 118:438–481.

Harald Baayen, Yu-Ying Chuang andElnaz Shafaei-Bajestan, and James Blevins. 2019. The Discriminative Lexicon: A Unified Computational Model for the Lexicon and Lexical Processing in Comprehension and Production Grounded Not in (De)Composition but in Linear Discriminative Learning. *Complexity*, page Article 4895891.

R. H. Baayen, R. Piepenbrock, and L. Gulikers. 1995. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Douglas Bates, Martin Maechler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67:1–48.

Laurie Bauer. 2014. Grammaticality, acceptability, possible words and large corpora. *Morphology*, 24:83–103.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Marisa Boston, John Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2:1–12.

Jonathan Brennan, Edward Stabler, Sarah Van Wagenen, Wen-Ming Luh, and John Hale. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, 157-158:81–94.

Manuel Carreiras, Blair Armstrong, Manuel Perea, and Ram Frost. 2014. The what, when, where, and how of visual word recognition. *Trends in Cognitive Sciences*, 18:90–98.

Laurent Cohen, Stanislas Dehaene, Lionel Naccache, Stephane Lehericy, Ghislaine Dehaene-Lambertz, Marie-Anne Henaff, and Francois Michel. 2000. The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123:291–307.

Laurent Cohen, Stephane Lehericy, Florence Chochon, Cathy Lemer, Sohie Rivaud, and Stanislas Dehaene. 2002. Language-specific tuning of visual cortex? Functional properties of the Visual Word Form Area. *Brain*, 125:1054–1069.

Mathias Creutz and Crista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4:3.

Anders M. Dale, Arthur K. Liu, Bruce R. Fischl, Randy L. Buckner, John W. Belliveau, Jeffrey D. Lewine, and Eric Halgren. 2000. Dynamic Statistical Parametric Mapping: Combining fMRI and MEG for High-Resolution Imaging of Cortical Activity. *Neuron*, 26:55–67.

Stanislas Dehaene, Gurvan Le Clec'H, Jean-Baptiste Poline, Denis Le Bihan, and Laurent Cohen. 2002. The visual word form area: A prelexical representation of visual words in the fusiform gyrus. *NeuroReport*, 13:321–325.

Stanislas Dehaene, Laurent Cohen, Mariano Sigman, and Fabien Vinckier. 2005. The neural code for written words: a proposal. *Trends in Cognitive Sciences*, 9:335–341.

Stanislas Dehaene, Lionel Naccache, Laurent Cohen, Denis Le Bihan, Jean-Francois Mangin, Jean-Baptiste Poline, and Denis Riviere. 2001. Cerebral mechanisms of word masking and unconscious repetition priming. *Nature NeuroScience*, 4:752–758.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109:193–210.

Suzanne Dikker, Hugh Rabagliati, and Liina Pylkkänen. 2009. Sensitivity to syntax in visual cortex. *Cognition*, 110:293–321.

Jay Earley. 1970. An ecient context-free parsing algorithm. *Communications of the Association for Computing Machinery*, 13:94–102.

Victoria Fossum and Roger Levy. 2012. Sequential vs. hierarchical syntactic models of human incremental sentence processing. *Proceedings of the*

3rd Workshop on Cognitive Modeling and Computational Linguistics, pages 61–69.

Stefan Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22:829–834.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Angela Friederici. 2002. Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, 6:78–84.

Angela Friederici. 2012. The cortical language circuit: from auditory perception to sentence comprehension. *Trends in Cognitive Sciences*, 16:262–268.

Joseph Fruchter and Alec Marantz. 2015. Decomposition, lookup, and recombination: MEG evidence for the Full Decomposition model of complex visual work recognition. *Brain and Language*, 143:81–96.

Joseph Fruchter, Linnaea Stockall, and Alec Marantz. 2013. MEG masked priming evidence for form-based decomposition of irregular verbs. *Frontiers in Human Neuroscience*, 7:798.

Laura Gwilliams, Gwyneth Lewis, and Alec Marantz. 2016. Functional characterisation of letter-specific responses in time, space and current polarity using magnetoencephalography. *Neuroimage*, 132:320–333.

Tero Hakala, Annika Hulten, Minna Lehtonen, Krista Lagus, and Riitta Salmelin. 2018. Information properties of morphologically complex words modulate brain activity during word reading. *Human Brain Mapping*, 39:2583–2595.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. *Proceedings of NAACL-2001*, pages 159–166.

Morris Halle. 1973. Prolegomena to a Theory of Word Formation. *Linguistic Inquiry*, 4:3–16.

Jennifer Hay. 2003. *Causes and Consequences of Word Structure*. Routledge, New York, NY.

P. Helenius, A. Tarkiainen, P. Cornelissen, P.C. Hansen, and R. Salmelin. 1999. Dissociation of Normal Feature Analysis and Deficient Processing of Letter-strings in Dyslexic Adults. *Cerebral Cortex*, 9:476–483.

John M. Henderson, Wonil Choi, Matthew W. Lowder, and Fernanda Ferreira. 2016. Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *NeuroImage*, 132:291–300.

Gregory Hickok and David Poeppel. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8:393–402.

John Hopcroft and Jeffrey Ullman. 1979. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading, MA.

Nikolaus Kriegeskorte and Pamela Douglas. 2018. Cognitive computational neuroscience. *Nature Neuroscience*, 21:1148–1160.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106:1126–1177.

Gwyneth Lewis, Olla Solomyak, and Alec Marantz. 2011. The Neural Basis of Obligatory Decomposition of Suffixed Words. *Brain and Language*, 118:118–127.

Gary Libben. 2003. Morphological parsing and morphological structure. In Egbert Assink and Dominiek Sandra, editors, *Reading Complex Words*, pages 221–239. Kluwer, New York.

Gary Libben. 2006. Getting at psychological reality: On- and off-line tasks in the investigation of hierarchical morphological structure. In G. Wiebe, G. Libben, T. Priestly, R. Smyth, and S. Wang, editors, *Phonology, Morphology, and the Empirical Imperative*, pages 349–369. Crane, Taipei.

Alessandro Lopopolo, Stefan L. Frank, Antal van den Bosch, and Roel M. Willems. 2017. Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PLoS ONE*, 12:e0177794.

Eric Maris and Robert Oostenveld. 2007. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164:177–190.

Petar Milin, Laurie Feldman, Michael Ramscar, Peter Hendrix, and Harald Baayen. 2017. Discrimination in lexical decision. *PLoS ONE*, 12.

Thomas Naselaris, Danielle Bassett, Alyson Fletcher, Konrad Kording, Nikolaus Kriegeskorte, Hendrikje Nienborg, Russell Poldrack, Daphna Shohamy, and Kendrick Kay. 2018. Cognitive Computational Neuroscience: A New Conference for an Emerging Discipline. *Trends in Cognitive Sciences*, 22:365–367.

Matthew J Nelson, Imen El Karoui, Kristof Giber, Xiaofang Yang, Laurent Cohen, Hilda Koopman, Sydney S Cash, Lionel Naccache, John T Hale, Christophe Pallier, et al. 2017. Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 114:3669–3678.

Boris New, Ludovic Ferrand, Christophe Pallier, and Marc Brysbaert. 2006. Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletion and Review*, 13:45–52.

Carolus Oldfield. 1971. The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, 9:97–113.

Yohei Oseki, Charles Yang, and Alec Marantz. 2019. Modeling Hierarchical Syntactic Structures in Morphological Processing. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 43–52.

Kristen Pammer, Peter Hansen, Morten Kringelbach, Ian Holliday, Gareth Barnes, Arjan Hillebrand, Krish Singh, and Piers Cornelissen. 2004. Visual word recognition: the first half second. *NeuroImage*, 22:1819–1825.

Barbara Partee, Alice ter Meulen, and Robert Wall. 1990. *Mathematical Methods in Linguistics*. Springer, Dordrecht.

Jonathan Peirce. 2007. PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, 162:8–13.

Jonathan Peirce. 2009. Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, 2:10.

Ingo Plag and Harald Baayen. 2009. Suffix Ordering and Morphological Processing. *Language*, 85:109–152.

Fermin Moscoso del Prado Martin, Aleksandar Kostic, and Harald Baayen. 2004. Putting the bits together: an information theoretical perspective on morphological processing. *Cognition*, 94:1–18.

Liina Pylkkänen and Alec Marantz. 2003. Tracking the time course of word recognition with MEG. *Trends in Cognitive Sciences*, 7:187–189.

Lawrence Rabinar. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77:257–286.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 324–333.

Michael Sipser. 1997. *Introduction to the Theory of Computation*. PWS Publishing, Boston, MA.

Olla Solomyak and Alec Marantz. 2010. Evidence for Early Morphological Decomposition in Visual Word Recognition: A Single-Trial Correlational MEG Study. *Journal of Cognitive Neuroscience*, 22:2042–2057.

Yoonsang Song, Youngah Do, Jongbong Lee, Arthur Thompson, and Eileen Waegemaekers. 2019. The reality of hierarchical morphological structure in multimorphemic words. *Cognition*, 183:269–276.

Jon Sprouse, Sagar Indurkhya, Beracah Yankama, Sandiway Fong, and Robert C. Berwick. 2018. Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *Linguistic Review*, 35:575–599.

Linnaea Stockall, Christina Manouilidou, Laura Gwilliams, Kyriaki Neophytou, and Alec Marantz. 2019. Prefix Stripping Re-Re-Revisited: MEG Investigations of Morphological Decomposition and Recomposition. *Frontiers in Psychology*, 10:1964.

Andreas Stolcke. 1995. An ecient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21:165–201.

M. Taft. 1979. Recognition of affixed words and the word frequency effect. *Memory and Cognition*, 7:263–272.

M. Taft. 2004. Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology*, 57:745–765.

M. Taft and K. I. Forster. 1975. Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, 14:638–647.

A. Tarkiainen, P. Helenius, P. C. Hansen, P. L. Cornelissen, and R Salmelin. 1999. Dynamics of letter string perception in the human occipitotemporal cortex. *Brain*, 122:2119–2132.

Sami Virpioja, Minna Lehtonen, Annika Hulten, Henna Kivikari, Riitta Salmelin, and Krista Lagus. 2017. Using Statistical Models of Morphology in the Search for Optimal Units of Representation in the Human Mental Lexicon. *Cognitive Science*, pages 1–35.

Roel M. Willems, Stefan L. Frank, Annabel D. Nijhof, Peter Hagoort, and Antal van den Bosch. 2016. Prediction During Natural Language Comprehension. *Cerebral Cortex*, 26:2506–2516.

Z.V.J. Woodhead, G.R. Barnes, W. Penny, R. Moran, S. Teki, C.J. Price, and A.P. Leff. 2014. Reading Front to Back: MEG Evidence for Early Feedback Effects During Word Recognition. *Cerebral Cortex*, 24:817–825.

Linda Zirkel. 2010. Prefix combinations in English: structural and processing factors. *Morphology*, 20:239–266.

Etyan Zweig and Liina Pylkkänen. 2009. A Visual M170 Effect of Morphological Complexity. *Language and Cognitive Processes*, 24:412–439.