Modeling Hierarchical Syntactic Structures in Morphological Processing

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

Charles Yang
Department of Linguistics & Psychology
University of Pennsylvania
charles.yang@ling.upenn.edu

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

Abstract

Sentences are represented as hierarchical syntactic structures, which have been successfully modeled in sentence processing. In contrast, despite the theoretical agreement on hierarchical syntactic structures within words, words have been argued to be computationally less complex than sentences and implemented by finite-state models as linear strings of morphemes, and even the psychological reality of morphemes has been denied. In this paper, extending the computational models employed in sentence processing to morphological processing, we performed a computational simulation experiment where, given incremental surprisal as a linking hypothesis, five computational models with different representational assumptions were evaluated against human reaction times in visual lexical decision experiments available from the English Lexicon Project (ELP), a "shared task" in the morphological processing literature. The simulation experiment demonstrated that (i) "amorphous" models without morpheme units underperformed relative to "morphous" models, (ii) a computational model with hierarchical syntactic structures, Probabilistic Context-Free Grammar (PCFG), most accurately explained human reaction times, and (iii) this performance was achieved on top of surface frequency effects. These results strongly suggest that morphological processing tracks morphemes incrementally from left to right and parses them into hierarchical syntactic structures, contrary to "amorphous" and finite-state models of morphological processing.

1 Introduction

Sentences are represented as hierarchical structures, not linear strings of words (Chomsky, 1957; Everaert et al., 2015). The hierarchical representations of sentences have been successfully modeled in sentence processing (Hale 2001; Levy 2008; Boston et al. 2008; Demberg and Keller 2008; Roark et al. 2009; Fossum and Levy 2012; cf. Frank and Bod 2011; Frank et al. 2012). In contrast, despite the theoretical agreement on hierarchical syntactic structures within words, especially derivational morphology, among various linguistic theories (Lieber, 1992; Anderson, 1992; Halle and Marantz, 1993; Aronoff, 1994), words have been argued to be computationally less complex than sentences (Langendoen 1981; Heinz and Idsardi 2011; cf. Carden 1983) and implemented by finite-state models as linear strings of morphemes (Beesley and Karttunen, 2003; Roark and Sproat, 2007; Virpioja et al., 2017), and even the psychological reality of morphemes has been denied by connectionist models (Baayen et al. 2011; Milin et al. 2017; cf. Anderson 1992). Consequently, the hierarchical representations of words have not been sufficiently considered in morphological processing, with a few exceptions (Libben, 2003, 2006; de Almeida and Libben, 2005; Pollatsek et al., 2010; Song et al., 2019).

In this paper, extending the computational models employed in sentence processing to morphological processing, we perform a computational simulation experiment where, given cumulative surprisal as a linking hypothesis (Hale, 2001; Levy, 2008), several computational models with different representational assumptions are evaluated against human reaction times (RTs) in visual lexical decision experiments available from the English Lexicon Project (ELP; Balota et al., 2007), a "shared task" in the morphological processing literature, with special focus on derivational morphology. The goal of this paper is to investigate whether morphological processing tracks morphemes and parses them into hierarchical syntactic structures.

Specifically, we employ five computational models with different representational assump-

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 43–52
Minneapolis, USA, June 7, 2019.

tions from sentence processing: two "amorphous" models, Letter Markov Model and Syllable Markov Model, with transition probabilities among letters and syllables, respectively, without reference to morpheme units and three "morphous" models, Markov Model, Hidden Markov Model, and Probabilistic Context-Free Grammars (PCFG), with conditional probabilities among morphemes, part-of-speech (POS) tags, and nonterminal nodes of hierarchical structures, respectively. Importantly, in the sentence processing literature, Markov Models and PCFGs have been exclusively compared (Frank and Bod, 2011; Fossum and Levy, 2012), but these computational models differ not only in the presence of hierarchical structures but also POS tags. Thus, we included HMM as an important "midpoint" model with POS tags but no hierarchical structures (cf. Lau et al., 2016). The prediction is that, if morphological processing tracks hierarchical syntactic structures, PCFG should outperform the alternative non-hierarchical models. Moreover, if morphological processing tracks morphemes, the "morphous" models should outperform the "amorphous" models.

2 Methods

2.1 Simulation Data

The simulation data was created by intersecting two corpora: CELEX (Baayen et al., 1995) and English Lexicon Project (ELP; Balota et al., 2007). These two corpora were selected because CELEX annotates morphological tree structures on which PCFG can be trained supervisedly, while ELP provides human reaction times (RTs) of visual lexical decision experiments against which computational models can be evaluated. First, every word except structurally ambiguous duplicates was extracted from the revised CELEX (O'Donnell, 2015) that only includes morphologically complex derived and monomorphemic words, hence 22,969 CELEX words.[1] Second, every word except those missing RTs or any control predictors to be included in the baseline model was extracted from the restricted ELP, hence 35,493 ELP words.[2] Finally, those sets of CELEX and ELP words were intersected, resulting in the simulation data of 13,244 morphologically complex derived and monomorphemic words.[3]

In order to make sure that model performance does not depend on the particular training/testing split, we adopted Monte Carlo cross-validation (MCCV), also known as repeated random subsampling, that repeatedly and randomly samples a subset of the full simulation data as the testing data and assigns the remaining data as the training data.[4] We only sampled bimorphemic words as the testing data, either suffixed (e.g. *teach+er*) or prefixed (e.g. *un+lock*), for the following two reasons. First, among morphologically complex words (n = 9,336), bimorphemic words account for more than 70% (n = 6,551), while trimorphemic, tetramorphemic, and supertetramorphemic words amount to only 24% (n = 2,277), 5% (n = 461), and 1% (n = 47), respectively. In other words, super-bimorphemic words can be nothing but outliers in the testing data. Second, given that computational models are multiplicative in nature (Yang, 2017), it is not fair to simultaneously test the words with different numbers of morphemes. That is, shorter words are exponentially more probable than longer ones, but shorter expressions are not necessarily more acceptable or easier to process (Lau et al., 2016; Sprouse et al., 2018). Given these two reasons, for each MCCV iteration, 10% of the bimorphemic words (n = 655) was randomly held out as the testing data and the remaining 90% (n = 13,244 – 655 = 12,589) was assigned as the train-

[1] The revised CELEX cleaned and expanded the original CELEX via hand annotation and heuristic parsing. See O'Donnell (2015, §7.2.2) for details.

[2] The restricted ELP only includes the words for which RT is available and computes paradigmatic lexical statistics like neighborhood density only among them. See Balota et al. (2007) for details.

[3] Another possibility would be that, like Virpioja et al. (2017), CELEX and ELP are independently used as training and testing data, respectively. While it is crucial in our computational simulation for morphemes to be consistent in training and testing data, however, morphological segmentations are not comparable across the two corpora, causing some morphemes to be unknown to computational models during testing, hence poor performance. Therefore, the intersection of the two corpora was necessary to ensure that morphemes are maximally identical in training and testing data.

[4] Another approach would be k-fold cross-validation (kFCV), that splits the full simulation data into k mutually exclusive and equally sized subsets and selects one subset for testing and k–1 subsets for training. kFCV is unbiased in that each word is guaranteed to get tested exactly once, but more variable because the number of iterations is restricted to k, the number of subsets. In contrast, MCCV is more robust than kFCV in that the number of iterations is not limited to the number of pre-split subsets (though biased because each word may be tested different times). That is, there is a general trade-off between variances and biases. Since the purpose here is just to ensure that model performance is robust among different training/testing splits, we adopted MCCV.

ing data. On the assumption that morphologically complex words are decomposed into component morphemes before morphological parsing, the testing words were represented as morpheme sequences (e.g. ['compute', 'ion', 'al']).[5] The number of iterations was set to 100 and the results presented below are all averaged across those 100 iterations, where the unparsed testing words were excluded (11 words per iteration on average).

2.2 Computational Models

The computational models were implemented with Natural Language Tool Kit (NLTK; Bird et al., 2009) in Python. The architectures of three types of computational models are summarized below: Markov Model, Hidden Markov Model, and Probabilistic Context-Free Grammar.

Markov Model: A Markov Model (also called *n*-gram model) was implemented with the `model` module. The Markov Model can be defined by an *n*-order Markov process that computes the transition probabilities of morphemes at position *i* given the *i*–*n* context, e.g. $P(m_i|m_{i-n}, m_{i-1})$. When *i* = 1, the 1st-order Markov Model (i.e. bigram model) computes the transition probabilities of morphemes at position *i* given the *i*–1 context, e.g. $P(m_i|m_{i-1})$. When *n* = 2, the 2nd-order Markov Model (i.e. trigram model) computes the transition probabilities of morphemes at position *i* given the *i*–2 context, e.g. $P(m_i|m_{i-1}, m_{i-2})$. Given the Markov assumption, the local probabilities of component morphemes in morphologically complex words are merely their transition probabilities.[6]

The transition probabilities are the model parameters empirically estimated from morpheme sequences in the training data via Maximum Likelihood Estimation with token weighting and Lidstone smoothing at $\alpha = 0.1$. The Markov Model is linear and string-oriented in that the transition probabilities merely track morphemes from left to right, which should effectively capture lexically specific dependencies among morphemes.

Hidden Markov Model (HMM): A HMM was implemented with the `hmm` module. A HMM generalizes the Markov Model by hypothesizing "hidden" structures behind visible strings. The HMM computes the transition probabilities of POS tags at position *i* given the *i*–1 context, e.g. $P(t_i|t_{i-1})$, and the emission probabilities of morphemes at position *i* given POS tags at the same position *i*, e.g. $P(m_i|t_i)$. Although the HMM, like the Markov Model, can be defined by an *n*-order Markov process over POS tags, only the Bigram HMM is investigated in this paper. The local probabilities of component morphemes in morphologically complex words are 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).[7]

While the local probabilities of component morphemes in structurally ambiguous words can be computed via a forward algorithm (`sum` of all paths) or a Viterbi algorithm (`max` of all paths), given that most probability mass was allocated to the best path and thus there were no substantial differences between forward and Viterbi algorithms, we adopted the forward algorithm. Both transition and emission probabilities are the model parameters empirically estimated from tagged morpheme sequences in the training data via Maximum Likelihood Estimation with token weighting and Lidstone smoothing at $\alpha = 0.1$. The HMM is structure-oriented in that hidden structures of POS tags are hypothesized behind visible strings, but still linear because the transition probabilities track POS tags from left to right.

Probabilistic Context-Free Grammar (PCFG): A PCFG was implemented with the `grammar` module. A PCFG is most representationally sophisticated among three types of computational models investigated in this paper and, crucially, can model hierarchical structures. The PCFG computes nonterminal production probabilities of right-hand sides given left-hand side nonterminals, e.g. $P(rhs|lhs)$, and terminal production probabilities of right-hand side terminals given left-hand side nonterminals, e.g. $P(m_i|t_i)$, corresponding to HMM emission prob-

[5] This is an empirical question whether morphological decomposition and morphological parsing are the same or different morphological computation(s). One possibility would be that top-down morphological parsing generates hierarchical structures while "emitting" morphemes as terminal nodes that provide cues to morphological decomposition.

[6] Bigram Markov Models append one word initial symbol <w> as the necessary context to estimate the probability of the first morpheme. Trigram Markov Models append two word initial symbols <w>, <w> to provide the context for the first morpheme, and so on.

[7] The term "prefix" as in prefix probabilities should not be confused with the term "prefix" in morphology (i.e. a type of affix linearly attached to the left of the base). The term "prefix" here means morpheme sequences that the incremental algorithm has encountered up to the current position.

abilities. The local probabilities of component morphemes in morphologically complex words are 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). Note that HMMs and PCFGs make different predictions even for bimorphemic words because derivational affixes are head-lexicalized in PCFGs (e.g. N → V *er*), while "emitted" from POS tags in HMMs.

Just like HMMs, while the local probabilities of component morphemes in structurally ambiguous words can be computed via an Earley algorithm (`sum` of all trees) or a Viterbi algorithm (`max` of all trees), we employed the Earley algorithm which may have interesting consequences for the incremental nature of morphological processing. Both nonterminal and terminal production probabilities are the model parameters empirically estimated from morphological tree structures in the training data via Maximum Likelihood Estimation with token weighting and Lidstone smoothing at $\alpha = 0.1$. The PCFG is hierarchical and structure-oriented in that the probabilities are defined over hierarchical structures permitted by the grammar.

2.3 Linking Hypothesis

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

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

The surprisal estimated by computational models has been demonstrated to explain self-paced reading times or eye-fixation durations in sentence processing (Boston et al., 2008; Demberg and Keller, 2008; Roark et al., 2009; Frank and Bod, 2011; Fossum and Levy, 2012) and remains to be extended to morphological processing (cf. Virpioja et al., 2017). Surprisal is a theory-neutral complexity metric in that computational models with different representational assumptions can be compared on the same probabilistic ground, unlike node counting (Miller and Chomsky, 1963) which only applies to the models with hierarchical structures. Thus, despite different representational assumptions, Markov Model, HMM, and PCFG can be equally evaluated through a lens of surprisal. Interestingly, Levy (2008) and Smith and Levy (2013) dubbed surprisal as a causal bottleneck: "surprisal serves as a causal bottleneck between the linguistic representations constructed during sentence comprehension and the processing difficulty incurred at a given word within a sentence" (Levy, 2008, p.1128). That is, various representational hypotheses assumed by different computational models can be evaluated via only one complexity metric ("the bottleneck"). See Hale (2016) for a review of information-theoretic complexity metrics.

On the assumption that morphological processing proceeds incrementally from left to right, we propose that processing costs of morphologically complex words are proportional to *cumulative surprisal* of their component morphemes. The cumulative surprisal of word *w*, $CI(w)$, is defined as Equation (2):[8]

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

where $I(m)$ is the surprisal of morpheme *m* defined as Equation (1). In fact, the mathematical equivalence of the cumulative surprisal of word *w*, $CI(w)$, and the vanilla surprisal of word *w*, $I(w)$ can be proved simply via the combination of the chain rule and the Markov assumption.

2.4 Statistical Analyses

Ordinary linear regression models were fitted with the `lm` function in R.[9] The baseline regression model was first fitted with log-transformed by-item average RTs as the dependent variable and control predictors as independent variables. For each computational model, the target regression model was then fitted with cumulative surprisal as the independent variable of interest on top of control predictors in the baseline regression model.

[8] In sentence processing, the processing costs of words within sentences can be easily measured with self-paced reading or eye-tracking experiments, but the processing costs of morphemes within words cannot, so that cumulative surprisal should be computed to transform processing costs from morphemes to words.

[9] Another approach would be linear mixed-effects regression (Baayen et al., 2008) with by-iteration random effects without averaging across 100 MCCV iterations. However, because of methodological uncertainties and convergence failures, we followed the standard practice of cross-validation and averaged the results across 100 MCCV iterations.

That is, the target and baseline regression models minimally differ only in the presence of cumulative surprisal. Therefore, the cumulative surprisal estimated by computational models was evaluated with nested model comparisons via log-likelihood ratio tests based on the χ^2-distribution with *df* = 1, the difference in the number of parameters between two nested regression models. Furthermore, the control predictors were evaluated via one-sample *t*-tests on beta regression coefficients based on the *z*-distribution, given that *t*-statistics approximately follow the *z*-distribution with 500 $>$ observations.

Following Lignos and Gorman (2012), four control predictors were included in the baseline regression model relative to which cumulative surprisal was evaluated: squared length, number of syllables, orthographic neighborhood density, and surface frequency. All control predictors were obtained from the ELP.

Squared length: Length (i.e. number of letters) has inhibitory effects on visual word recognition: longer words are recognized more slowly. Since New et al. (2006) found that the quadratic term of length (i.e. number of letters squared) was closely correlated with RTs in the ELP (i.e. "U-shaped curve" of RTs as a function of length), we adopted squared length.

Number of syllables: New et al. (2006) also observed that number of syllables had "robust linear inhibitory effects" on visual word recognition independent of squared length and thus we adopted number of syllables.

Orthographic neighborhood density: Orthographic neighborhood density has been recognized to have inhibitory effects on visual word recognition: words in denser neighborhood are recognized more slowly. Yarkoni et al. (2008) proposed a new measure of orthographic neighborhood density called Orthographic Levenshtein Distance (OLD) which was shown to predict RTs in the ELP better than the classic measure known as Coltheart's N (Coltheart et al., 1977). Thus, we included a version of OLD computed based on 20 closest orthographic neighbors (OLD20).

Surface frequency: Frequency has facilitatory effects on visual word recognition and probably is the most important predictor in the psycholinguistics literature: more frequent words are recognized more quickly. In morphologically complex visual word recognition, theoretical interpretations of frequency crucially depend on the linguistic units over which frequency is computed. For example, surface frequency has been interpreted as an index of storage of morphologically complex words as unanalyzed wholes, whereas base frequency as a "litmus paper" of computation of morphologically complex words from component morphemes. Among various frequency norms such as the Brown Corpus (Kucera and Francis, 1967), the CELEX (Baayen et al., 1995), and the HAL (Burgess and Livesay, 1998), we used the SUBTLEX frequency norm (Brysbaert and New, 2009) which was demonstrated to predict RTs in the ELP better than the previous frequency norms. Specifically, we log-transformed a version of SUBTLEX frequency scaled per million, because frequency is known to follow the nonlinear Zipfian distribution (Zipf, 1949). Note that surface frequency is proportional to unigram probability estimated by "word unigram model", the model of storage discussed by Virpioja et al. (2017), simply because unigram probabilities are computed by dividing surface frequencies by the corpus size.

2.5 Evaluation Metrics

Two evaluation metrics are derived from surprisal: linguistic accuracy and psychological accuracy (Frank and Bod, 2011; Fossum and Levy, 2012).[10] The linguistic accuracy of model *M*, $LA(M)$, is defined as Equation (3):

$$LA(M) = -\frac{1}{n}\sum_{i=1}^{n} I(m_i) \tag{3}$$

where $I(m)$ is the surprisal of morpheme *m* defined as Equation (1). That is, the linguistic accuracy is the negative average surprisal over morphemes of morphologically complex words in the testing data. Note also that the linguistic accuracy is just the negative of the NLP evaluation metric *cross-entropy*. The linguistic accuracy may be cognitively interpreted as offline grammaticality judgment (Keller, 2000; Lau et al., 2016; Sprouse et al., 2018): the higher the linguistic accuracy is, the more grammatical the model "judges" the testing data never seen before. Note that the linguistic accuracy is completely independent of human be-

[10] Virpioja et al. (2017) call variants of linguistic and psychological accuracies as text prediction and cognitive prediction accuracies, respectively.

havior (i.e. human RTs), in contrast with the psychological accuracy introduced below.

The psychological accuracy of model M, $PA(M)$, is defined as Equation (4):

$$PA(M) = \Delta D_B - \Delta D_M \quad (4)$$

where ΔD is the delta deviance defined as –2 times log-likelihood and B is the baseline model without cumulative surprisal included. That is, the psychological accuracy is the decrease in delta deviance between the baseline model and the target model fitted to the testing data. The psychological accuracy may be cognitively interpreted as online morphological processing: the higher psychological accuracy is, the less costly the model "processes" the testing data never seen before. For example, suppose that the grammatical sentence *Colorless green ideas sleep furiously* (Chomsky, 1957) empirically turned out to be less costly. The most "human-like" model must assign the high probability, hence the less surprisal, to this sentence. Interestingly, Frank and Bod (2011) and Fossum and Levy (2012) inductively observed that linguistic and psychological accuracies are positively correlated (cf. Virpioja et al., 2017), suggesting that the relationship between representation and processing is transparent (Chomsky, 1965; Hale, 2001).

3 Results

3.1 Linguistic and Psychological Accuracies

Linguistic and psychological accuracies of computational models are summarized in Figure 1, where the x-axis is linguistic accuracy (negative average surprisal) and the y-axis is psychological accuracy (decrease in delta deviance). The accuracies are averaged across 100 MCCV iterations. Points represent computational models and vertical bars on the points are 95% confidence intervals of the psychological accuracy.[11] The horizontal dashed line is $\chi^2 = 3.84$, the critical χ^2-statistic at $p = 0.05$ with $df = 1$.

First, "morphous" models were psychologically more accurate than "amorphous" models. Nested model comparisons via log-likelihood ratio tests revealed that all "morphous" models were statistically significant ($p < 0.01$), but one of two "amorphous" models (i.e. Letter Markov Model) did not reached statistical significance. Second, the PCFG was psychologically most accurate among the five computational models: PCFG ($\chi^2 = 14.57$) > HMM ($\chi^2 = 13.83$) > Morpheme Markov Model ($\chi^2 = 13.65$) > Syllable Markov Model ($\chi^2 = 12.84$) < Letter Markov Model ($\chi^2 = 3.52$). Third, the PCFG was also linguistically most accurate, where the correlation between linguistic and psychological accuracies among five computational models was high ($r = 0.81$).

3.2 Control Predictors

Effects of control predictors are summarized in Figure 2, where the x-axis is t-statistic and the y-axis is control predictors. The t-statistics are averaged across 100 MCCV iterations. Points represent computational models and horizontal bars on the points are 95% confidence intervals of the t-statistic. Vertical dashed lines are $t = \pm 1.96$, the critical t-statistic at $p = 0.05$ with $df = \infty$.

All control predictors except visual predictors like squared length and number of syllables were statistically significant ($p < 0.05$). The surface frequency effects were most robustly observed among the four control predictors: Letter Markov Model ($t = -17.34$), Syllable Markov Model ($t = -16.71$), Morpheme Markov Model ($t = -16.19$), HMM ($t = -16.49$), and PCFG ($t = -16.58$). Note that surface frequency was most pronounced in combination with the PCFG among three "morphous" models, suggesting that cumulative surprisal estimated by the PCFG explains unique variances not covered by surface frequency.

4 Discussion

In summary, the results of the simulation experiment demonstrated that "morphous" models were more psychologically accurate than "amorphous" models, contrary to "amorphous" models of morphological processing (Baayen et al., 2011; Milin et al., 2017). Among three computational models with morpheme units, the PCFG was most accurate both linguistically and psychologically, suggesting that morphological processing tracks hierarchical syntactic structures, contrary to finite-state models of morphological processing (Beesley and Karttunen, 2003; Roark and Sproat, 2007; Virpioja et al., 2017). Interestingly, syntactic granularity was transparently mapped to psychological accuracy: PCFG with hierarchical

[11] Thanks to the central limit theorem, while the test statistic itself is χ^2-statistic, the samples of χ^2-statistic follow the Gaussian distribution, based on which 95% confidence intervals can be computed.

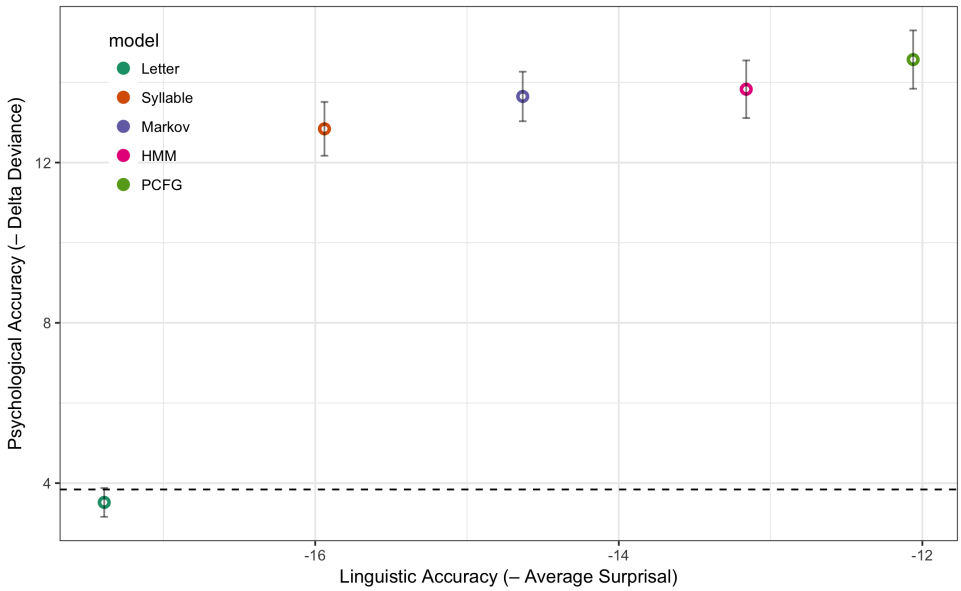

Figure 1: Linguistic and psychological accuracies of computational models, averaged across 100 MCCV iterations. The x-axis is linguistic accuracy (negative average surprisal), while the y-axis is psychological accuracy (decrease in delta deviance). Points represent computational models. Vertical bars on the points are 95% confidence intervals of the psychological accuracy. The horizontal dashed line is $\chi^2 = 3.84$, the critical χ^2-statistic at $p = 0.05$ with $df = 1$. All computational models except Letter Markov Model were statistically significant ($p < 0.01$).

structures was more accurate than HMM with POS tags but no hierarchical structures, which in turn was more accurate than Markov Model with neither hierarchical structures nor POS tags, meaning that hierarchical structures and POS tags made independent contributions for predicting human RTs in visual word recognition. In addition, given that the cumulative surprisal was computed by the PCFG via a probabilistic Earley parser (Earley, 1970; Stolcke, 1995), a top-down parser that incrementally computes probabilities morpheme by morpheme in morphologically complex words, this result may also indicate that morphological processing proceeds incrementally from left to right, despite the inherently non-incremental nature of visual word recognition.[12]

Moreover, the effects of surface frequency and cumulative surprisal were simultaneously observed, theoretically reflecting storage and computation, respectively. The simultaneous effects of surface frequency and cumulative surprisal were not surprising under either the single-route decomposition model of morphological processing (Taft, 1979, 2004; Taft and Forster, 1975), where storage and computation are indexed at functionally different stages of morphological processing (cf. Solomyak and Marantz, 2010; Fruchter and Marantz, 2015) or the dual-route model of morphological processing (Pinker and Prince, 1988; Pinker and Ullman, 2002), where storage and computation "routes" work in parallel. While Virpioja et al. (2017) interpreted the simultaneous effects of storage and computation as evidence in favor of the dual-route model of morphological processing, however, since RTs are an "end-point" measure of morphological processing, the two competing models cannot be conclusively dissociated. In fact, Virpioja et al. (2017, p.29) admits that "As the present study used simple RTs which provide an end-point measure of the entire recognition process, either or both of these alternatives about the word recognition process could be correct". Remember that surface frequency was most pronounced with the PCFG among three "morphous" models, indicating that the PCFG can explain unique uncorrelated variances not covered by surface frequency. Additionally, the recent conclusion reached by Virpioja et al. (2017) that derived words are primarily stored in the mental lexicon, not computed from their component morphemes, does not harmonize with the simultaneous effects of surface frequency and cumulative

[12] An anonymous reviewer insightfully pointed out that the Cohort Model (Marslen-Wilson, 1987) may harmonize with the present idea that a probabilistic parser applied to morphological processing incrementally contracts the mental lexicon from left to right, which remained to be investigated in future.

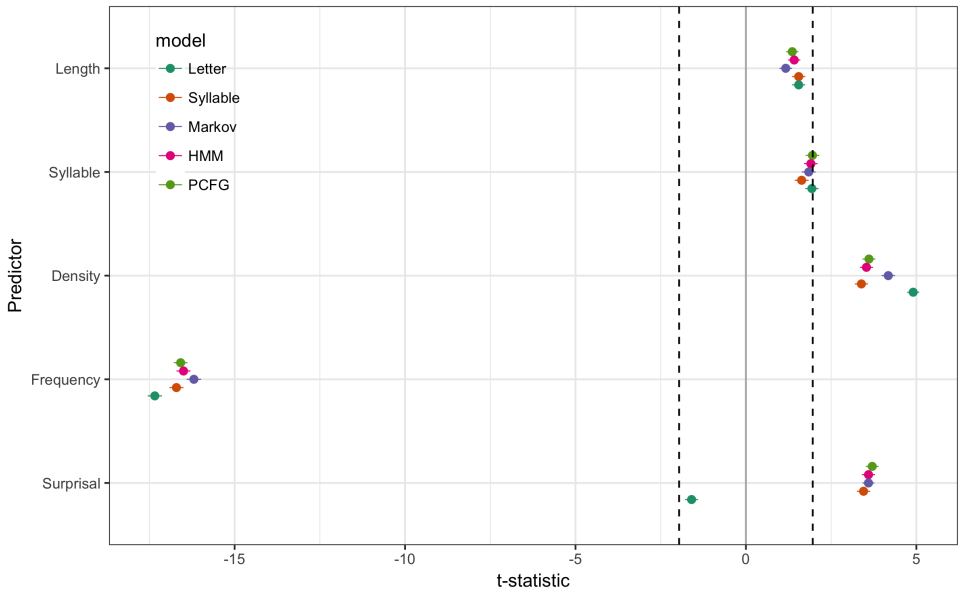

Figure 2: Effects of control predictors, averaged across 100 MCCV iterations. The x-axis is t-statistic, while the y-axis is control predictors. Points represent computational models. Horizontal bars on the points are 95% confidence intervals of the t-statistic. Vertical dashed lines are $t = \pm 1.96$, the critical t-statistic at $p = 0.05$ with $df = \infty$. All control predictors except visual predictors were statistically significant ($p < 0.05$).

surprisal, either.

Nevertheless, remember that we only sampled bimorphemic words as the testing data. However, as Libben (2003, 2006) pointed out, bimorphemic words are not sufficient to distinguish hierarchical structures and linear strings, and trimorphemic words are minimally required. In future, the computational models must be evaluated against trimorphemic words to make sure that the results will generalize beyond bimorphemic words.

5 Conclusion

In this paper, we performed a computational simulation experiment with human RTs in visual lexical decision experiments available from the ELP (Balota et al., 2007), a "shared task" in the morphological processing literature, and evaluated computational models with different representational assumptions via cumulative surprisal as a linking hypothesis (Hale, 2001; Levy, 2008), in order to investigate whether morphological processing tracks morphemes and parses them into hierarchical syntactic structures. Consequently, the results of the simulation experiment demonstrated that "morphous" models were psychologically more accurate than "amorphous" models and, importantly, a computational model with hierarchical syntactic structures, PCFG, was most psychologically accurate among five computational models, contrary to "amorphous" (Baayen et al., 2011) and finite-state (Beesley and Karttunen, 2003) models of morphological processing.

Acknowledgments

We would like to thank Tal Linzen and CMCL anonymous reviewers for valuable suggestions. This work was supported by JSPS KAKENHI Grant Number JP18H05589.

References

Roberto de Almeida and Gary Libben. 2005. Changing morphological structures: The effect of sentence context on the interpretation of structurally ambiguous English trimorphemic words. *Language and Cognitive Processes*, 20:373–394.

Stephen Anderson. 1992. *A-morphous morphology*. Cambridge University Press, Cambridge.

Mark Aronoff. 1994. *Morphology by Itself*. MIT Press, Cambridge, MA.

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.

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

D.A. Balota, M.J. Yap, M.J. Cortese, K.A. Hutchison, B. Kessler, B. Loftis, J.H. Neely, D.L. Nelson, G.B. Simpson, and R. Treiman. 2007. The English Lexicon Project. *Behavior Research Methods*, 39:445–459.

Kenneth Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, University of Chicago Press.

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.

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41:977–990.

Curt Burgess and Kay Livesay. 1998. The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. *Behavior Research Methods*, 30:272–277.

Guy Carden. 1983. The Non-Finite = State-Ness of the Word Formation Component. *Linguistic Inquiry*, 14:537–541.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Max Coltheart, Eileen Davelaar, Jon Torfi Jonasson, and Derek Besner. 1977. Access to the internal lexicon. In Stanislav Dornic, editor, *Attention and Performance*, pages 535–555. Erlbaum, Hillsdale, NJ.

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

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

Martin Everaert, Marinus Huybregts, Noam Chomsky, Robert Berwick, and Johan Bolhuis. 2015. Structures, Not Strings: Linguistics as Part of the Cognitive Sciences. *Trends in Cognitive Sciences*, 19:729–743.

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 Frank, Rens Bod, and Morten Christiansen. 2012. How hierarchical is language use? *Proceedings of the Royal Society B*, 279:4522–4531.

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.

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

John Hale. 2016. Information-theoretical Complexity Metrics. *Language and Linguistics Compass*, 10:397–412.

Morris Halle and Alec Marantz. 1993. Distributed Morphology and the Pieces of Inflection. In Ken Hale and Samuel Keyser, editors, *The View from from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*, pages 111–176. MIT Press, Cambridge, MA.

Jeffrey Heinz and William Idsardi. 2011. Sentence and Word Complexity. *Science*, 333:295–297.

Frank Keller. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Ph.D. thesis, University of Edinburgh.

Henry Kucera and Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence.

Terence Langendoen. 1981. The Generative Capacity of Word-Formation Components. *Linguistic Inquiry*, 12:320–322.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2016. Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cognitive Science*, pages 1–40.

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

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.

Rochelle Lieber. 1992. *Deconstructing Morphology*. University of Chicago Press, Chicago.

Constantine Lignos and Kyle Gorman. 2012. Revisiting frequency and storage in morphological processing. *Proceedings of CLS*, 48:447–461.

William Marslen-Wilson. 1987. Functional parallelism in spoken word recognition. *Cognition*, 25:71–102.

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

George Miller and Noam Chomsky. 1963. Finitary models of language users. In Duncan Luce, Robert Bush, and Eugene Galanter, editors, *Handbook of Mathematical Psychology*, volume 2, pages 419–491. Wiley, New York.

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.

Timothy O'Donnell. 2015. *Productivity and Reuse in Language*. MIT Press, Cambridge, MA.

Steven Pinker and Alan Prince. 1988. On language and connectionism. *Cognition*, 28:73–193.

Steven Pinker and Michael Ullman. 2002. The past and future of the past tense. *Trends in Cognitive Sciences*, 6:456–462.

Alexander Pollatsek, Denis Drieghe, Linnaea Stockall, and Roberto de Almeida. 2010. The interpretation of ambiguous trimorphemic words in sentence context. *Psychonomic Bulletion and Review*, 17:88–94.

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.

Brian Roark and Richard Sproat. 2007. *Computational Approaches to Morphology and Syntax*. Oxford University Press, Oxford.

Nathaniel Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

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.

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.

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.

Charles Yang. 2017. Rage against the machine: Evaluation metrics in the 21st century. *Language Acquisition*, 24:100–125.

Tal Yarkoni, David Balota, and Melvin Yap. 2008. Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletion and Review*, 15:971–979.

George Zipf. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley, Cambridge, MA.