Dissociating Syntactic Operations via Composition Count

Kohei Kajikawa (kohei-kajikawa@g.ecc.u-tokyo.ac.jp)
Department of Language and Information Sciences, The University of Tokyo, Meguro-ku, Tokyo, Japan

Ryo Yoshida (yoshiryo0617@g.ecc.u-tokyo.ac.jp)
Department of Language and Information Sciences, The University of Tokyo, Meguro-ku, Tokyo, Japan

Yohei Oseki (oseki@g.ecc.u-tokyo.ac.jp)
Department of Language and Information Sciences, The University of Tokyo, Meguro-ku, Tokyo, Japan

Abstract

Computational psycholinguistics has traditionally employed a complexity metric called Node Count, which counts the number of *syntactic nodes* representing syntactic structures and predicts processing costs in human sentence processing. However, Node Count does not dissociate distinct syntactic operations deriving those syntactic structures, so that how much processing cost each syntactic operation induces remains to be investigated. In this paper, we introduce a novel complexity metric dubbed **Composition Count**, which counts the number of *syntactic operations* deriving syntactic structures, allowing us to understand the computational system of human sentence processing from the derivational, not representational, perspective. Specifically, employing Combinatory Categorial Grammar (CCG) which is equipped with multiple syntactic operations and thus suitable for the purpose here, we investigate (i) how much distinct syntactic operations of CCG contribute to predicting human reading times, and (ii) whether the same holds across languages. The results demonstrate that distinct syntactic operations of CCG have independent and cross-linguistic contributions to predicting human reading times, while Node Count turns out not to be robust cross-linguistically. In conclusion, these results strongly suggest the importance of Composition Count to dissociate distinct syntactic operations, not whole syntactic representations, and understand the computational system of human sentence processing.

Keywords: human sentence processing; reading time; Node Count; Composition Count; Combinatory Categorial Grammar

Introduction

Natural language has syntactic structures (Chomsky, 1957), which are essential for computing meanings (Heim & Kratzer, 1998). The previous literature has demonstrated that syntactic structures are built in human sentence processing, as evidenced by both behavioral and neural data (e.g., Roark, Bachrach, Cardenas, & Pallier, 2009; Fossum & Levy, 2012; J. R. Brennan, Stabler, Wagenen, Luh, & Hale, 2016; Nelson et al., 2017; Hale, Dyer, Kuncoro, & Brennan, 2018). Theoretically, it is assumed that the computational processes of constructing hierarchical structures are directly based on competence grammar (competence hypothesis, Chomsky, 1965; Bresnan & Kaplan, 1982; Berwick & Weinberg, 1983; Steedman, 2000; Marantz, 2005; Sag & Wasow, 2011; S. Lewis & Phillips, 2015). In light of this hypothesis, theories like *derivational theory of complexity* (DTC, Miller & Chomsky, 1963), which quantifies processing costs depending on the number of syntactic operations within competence grammar, have been proposed. Building upon this theoretical foundation, computational psycholinguistics has traditionally employed a complexity metric called Node Count, which counts the number of *syntactic nodes* representing syntactic structures and predicts processing costs in human sentence processing (J. R. Brennan et al., 2012, 2016; J. R. Brennan & Pylkkänen, 2017; Nelson et al., 2017; Bhattasali et al., 2018; Li & Hale, 2019; Stanojević et al., 2021; Li et al., 2022; Stanojević, Brennan, Dunagan, Steedman, & Hale, 2023).

However, Node Count does not dissociate distinct syntactic operations deriving those syntactic structures, so how much processing cost each syntactic operation induces remains to be investigated. In this respect, while the previous literature has almost exclusively employed Context-Free Grammar (CFG) in combination with Node Count, Combinatory Categorial Grammar (CCG, Steedman, 2000) is equipped with multiple syntactic operations with distinct syntactic and semantic properties and thus suitable for the purpose here. In fact, recent results have demonstrated that Node Count and the Reveal operation of CCG successfully predict processing costs in human sentence processing (Stanojević et al., 2021; Stanojević et al., 2023).

In this paper, we introduce a novel complexity metric dubbed **Composition Count**, which counts the number of *syntactic operations* deriving syntactic structures, allowing us to understand the computational system of human sentence processing from the derivational, not representational, perspective. Specifically, employing CCG which is equipped with multiple syntactic operations such as Function Application (FA), Function Composition (FC), and Type Raising (TR), we investigate (i) how much distinct syntactic operations of CCG contribute to predicting human reading times, and (ii) whether the same holds across languages, especially in both head-initial (English) and head-final (Japanese) languages. The results demonstrate that distinct syntactic operations of CCG have independent and cross-linguistic contributions to predicting human reading times: FA/TR and FC exhibit positive and negative effects, respectively, with the relative magnitude of the effects being FA $>$ TR $>$ FC. In contrast, Node Count turns out not to be robust cross-linguistically. In conclusion, these results strongly suggest the importance of Composition Count to dissociate distinct syntactic operations, not whole syntactic representations, and understand the computational system of human sentence processing.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Background

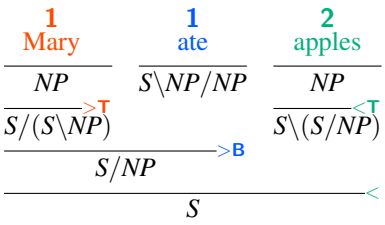
Node Count

Node Count assumes that, for a given syntactic structure of a sentence, under a specific parsing strategy, the total number of nodes traversed between each word corresponds to the processing costs of that word. For instance, let us examine the process of traversing a CCG syntactic structure of the sentence *Mary ate apples* as depicted in Figure 1a with a bottom-up parsing strategy. In this process, the total number of nodes traversed for each word is 1, 1, and 2, respectively. These numbers are just the processing costs postulated by Node Count. A more intuitive way to put it is that Node Count is a complexity metric that calculates the processing costs based on a *syntactic nodes* representing syntactic structures.

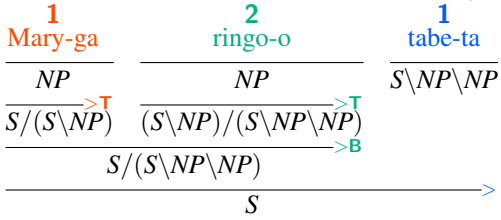
Historically, Node Count is a complexity metric that originates from the derivational theory of complexity (DTC), which was once denied by Fodor, Bever, and Garrett (1974). However, as Marantz (2005) correctly pointed out, it is premature to conclude the failure of DTC given the underdeveloped state of syntactic theories at the time. In addition, DTC is "just a name for standard methodology in cognitive science and cognitive neuroscience" (Marantz, 2005, p.439). In fact, DTC-inspired metrics like Node Count have been widely used in research on human sentence processing (e.g., Miller & Chomsky, 1963; Frazier, 1985; Hawkins, 1994; J. R. Brennan et al., 2012, 2016; J. R. Brennan & Pylkkänen, 2017; Nelson et al., 2017; Bhattasali et al., 2018; Hale et al., 2018; Li & Hale, 2019; Stanojević et al., 2021; Li et al., 2022; Stanojević et al., 2023).

Node Count, which directly calculates a processing cost from a syntactic structure, is different from the other two major types of complexity metrics used in psycholinguistics and computational psycholinguistics, memory-based metrics (e.g., Gibson, 2000; R. L. Lewis & Vasishth, 2005) and expectation-based metrics like surprisal (Hale, 2001; Levy, 2008). Indeed, Node Count partially overlaps with a memory-based metric, but they are conceptually different and it has been shown that Node Count can explain variances in neural data that the other metrics cannot (Bhattasali et al., 2018; Li & Hale, 2019; Stanojević et al., 2023).

However, Node Count does not dissociate distinct syntactic operations, assuming them as uniform processing loads. Consequently, it remains unclear whether distinct syntactic operations have effects on behavioral data differently and how much processing cost each operation incurs.



(a) Incremental derivation of *Mary ate apples* (English)



(b) Incremental derivation of *Mary-ga ringo-o tabe-ta* (Japanese)

Figure 1: Incremental CCG derivations of English and Japanese sentences. The operation applied to each word is shown in the same color as the word. The number above each word indicates the number of nodes that are constructed at each word (Node Count).

Combinatory Categorial Grammar

CCG is a formal linguistic theory that enables incremental construction of structures at a level that can be implemented on computers and its adequacy for modeling hemodynamic activity has already been shown (Stanojević et al., 2021; Stanojević et al., 2023). In addition, CCG is equipped with multiple syntactic operations with distinct syntactic and semantic properties. In CCG, all the syntactic and semantic representations are constructed in parallel, projected from each lexical item recursively via a small number of type-dependent syntactic operations. These syntactic representations are referred to as syntactic categories consisting of *atomic categories* like N, NP, or S, and *complex categories* like NP/NP or $S\backslash NP\backslash NP$. Complex categories are recursively built from basic categories with two types of slashes ("/", "\") representing the directions of arguments. X/Y is a function that takes Y as an argument to its right to yield X, while $X\backslash Y$ is also a function that takes Y as an argument to its left to yield X.

The syntactic operations mainly used in this study are shown below:[1]

Function Application (FA)

$$\begin{cases} X/Y \quad Y \implies_{>} X \\ Y \quad X\backslash Y \implies_{<} X \end{cases}$$

Type Raising (TR)

$$\begin{cases} X \implies_{>\mathbf{T}} T/(T\backslash X) \\ X \implies_{<\mathbf{T}} T\backslash(T/X) \end{cases}$$

Function Composition (FC)

$$\begin{cases} X/Y \quad Y/Z \implies_{>\mathbf{B}} X/Z \\ Y\backslash Z \quad X\backslash Y \implies_{<\mathbf{B}} X\backslash Z \end{cases}$$

Note that T is a meta-variable over categories. With FC and TR, CCG can construct flexible constituents, which contributes to deriving non-constituent coordination and long-distance dependencies. For example, an object relative clause has a constituent that lacks an object in *wh*-clause on its surface structure shown in Figure 2.

Furthermore, owing to FC and TR, CCG can derive a left-branching structure in both head-initial (English) and head-final (Japanese) languages, which makes it possible to eagerly build syntactic structures with a bottom-up parsing strategy

[1] We used the syntactic operations employed in the CCG parsers developed by Stanojević and Steedman (2019) for English and Yoshikawa, Noji, and Matsumoto (2017) for Japanese.

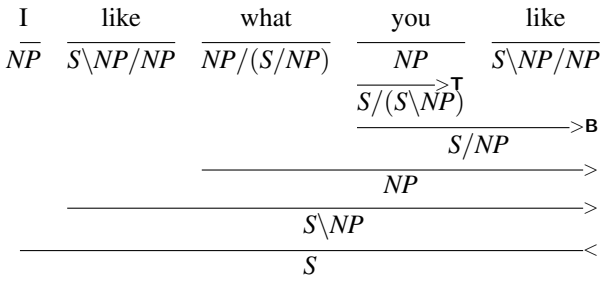


Figure 2: Derivation of an English objective relative clause. The syntactic operations used to derive each constituent are indicated on the right of the bar.

as shown in Figure 1.[2] Previous work has shown that the CCG left-branching structures tend to be more effective at explaining incremental sentence processing than CCG right-branching structures in English (Stanojević et al., 2021). This finding aligns with theoretical predictions concerning human memory capacity (Abney & Johnson, 1991; Resnik, 1992). In the case of a head-final language like Japanese, the construction of a left-branching CCG structure requires establishing the relationship between arguments before the verb, which is supported by the findings in psycholinguistics (e.g., Mazuka & Itoh, 1995; Kamide & Mitchell, 1999; Isono & Hirose, 2022).

By employing CCG as a theory of grammar, we can examine how much distinct syntactic operations with distinct syntactic and semantic properties contribute to predicting human reading times. There is no empirical reason to believe that these operations incur the same processing cost, as assumed by Node Count.

Methods

In this paper, we employed CCG to investigate how much the predictors based on the three main syntactic operations of CCG (Function Application (FA), Function Composition (FC), and Type Raising (TR)) contribute to predicting human reading times, and whether the same holds for both English and Japanese. In addition, we investigated whether Node Count itself can predict reading times, as there is no empirical evidence to suggest that Node Count is a robust predictor for predicting reading times, unlike its established efficacy with neural data. Specifically, we constructed linear mixed-effects regression models (Baayen, Davidson, & Bates, 2008) including predictors based on distinct syntactic operations and Node Count for modeling reading times, and examined the coefficients of the predictors.

[2]TR is considered a lexical rather than a syntactic operation in Steedman (2000), but following Stanojević et al. (2021); Stanojević et al. (2023), we treated it as a syntactic operation. Moreover, while in the previous studies on English, TR is not applied to noun phrases in object positions, we apply TR to such noun phrases as well, which was conducted to align conditions between English and Japanese. In Japanese, an SOV language, it is necessary to apply TR to noun phrases in object positions to construct a left-branching structure.

Reading time data

As behavioral data, we used log-transformed total reading times of eye-tracking from the Dundee corpus (Kennedy, Hill, & Pynte, 2003) for English and BCCWJ-EyeTrack (Asahara, Ono, & Miyamoto, 2016) for Japanese. The Dundee corpus includes the reading times of 10 English native speakers to 2,368 sentences, while BCCWJ-EyeTrack contains the reading times of 24 Japanese native speakers to 218 sentences. Note that while the reading times are annotated to space-separated words in English, they are annotated to each phrasal unit in Japanese. To align these differences, we put the word-by-word data points in the Dundee corpus together into phrasal units based on gold dependency annotation by Barrett, Agić, and Søgaard (2015); as shown in Figure 3, focusing on only dependency arrows extending from right to left, *words with the same dependent and the dependent itself* were combined into a single phrase. Furthermore, the words with an arrow extending from the root were treated as a single phrase. Through this grouping process, the annotations, such as word positions and dependency relations, were reconfigured.

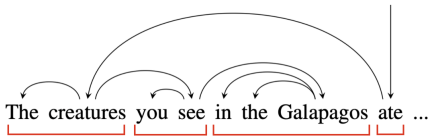


Figure 3: Composing phrasal unit based on dependency structure

Then, for the Dundee corpus, data points (i) not fixated and (ii) where the phrasal unit grouped crosses over separate lines were removed to align the unit with the Japanese dataset. For BCCWJ-EyeTrack, data points (i) not fixated and (ii) where the word units obtained by `sentencepiece` (Kudo & Richardson, 2018), the tokenizer used in GPT-2 (Radford et al., 2019), crossed over the phrasal units of BCCWJ-EyeTrack were removed. GPT-2 was utilized for calculating surprisal, one of the explanatory variables for reading times. Consequently, we used 182,736 data points from the Dundee corpus and 8,911 data points from BCCWJ-EyeTrack in the statistical analyses.

Syntactic structures

We assumed syntactic structures being constructed during sentence comprehension are incremental left-branching CCG derivations, as shown in Figure 1.

To obtain a left-branching CCG derivation for the Dundee corpus and BCCWJ-EyeTrack, we initially parsed the sentences in these corpora to get the best derivation with `ccgtools`,[3] an English CCG parser, and `depccg` (Yoshikawa et al., 2017),[4] a Japanese CCG parser, respectively. These

[3]`https://github.com/stanojevic/ccgtools`
[4]`https://github.com/masashi-y/depccg`

word	*Mary*	*ate*	*apples*	*Mary-ga*	*ringo-o*	*tabeta*
FA	0	0	1	0	0	1
FC	0	1	0	0	1	0
TR	1	0	1	1	1	0
NC	1	1	2	1	2	1

Table 1: Composition Counts and Node Count of each word of CCG derivations in Figure 1. *FA*, *FC*, and *TR* represent the Composition Counts of FA, FC, and TR, respectively. *NC* represents the Node Count.

parsers were trained to output right-branching CCG derivations. Subsequently, to rotate the right-branching derivations to left-branching, we performed TR on each *NP* node that was taken by verbs as an argument. [5] Finally, we applied the "tree-rotation" operation (Stanojević & Steedman, 2019) to the resulting type-raised right-branching trees. All of the type-raised right-branching trees were recursively rotated from bottom to top to achieve left-branching trees.

Complexity metrics

We used the total number of each syntactic operation per phrase as complexity metrics, assuming a bottom-up parsing strategy. Specifically, we focused on three representative syntactic operations: FA, FC, and TR, which have robust theoretical support in CCG and are implemented in a CCG parser. We refer to the count of these operations as *Composition Count*. Table 1 presents the Composition Counts for the English and Japanese derivation trees shown in Figure 1. Additionally, we also calculated the Node Count per phrase for comparison with the Composition Counts. The Node Count represents the sum of the Composition Count for the three syntactic operations and the count of other operations proper to the CCG parsers.

Statistical analyses

For the regression model to analyze reading time data, we used a linear mixed-effects model (Baayen et al., 2008). This was implemented with the `lmer` function in the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2023). The *p*-values were approximated with the `lmerTest` package (Kuznetsova, Brockhoff, & Christensen, 2017). The following regression model was used as a baseline regression model:

```
log(RT) ~ dependent + wlen + num_of_word
        + freq + prev_freq + surp
        + phraseN + lineN + screenN
        + prev_is_fixed
        + (1|article) + (1|subject)
```

In the baseline regression model, we used the predictors that have been suggested to have the effective power to model the human reading time data. `dependent` is the number of dependency relations to the phrasal unit from the previous phrases. This is intended to capture *anti-locality effect* (Konieczny, 2000). The dependency structure of each corpus was annotated manually by Barrett et al. (2015) and Asahara and Matsumoto (2016). `wlen` and `num_of_word` are the number of characters and the number of words composed of each phrase, respectively. The predictor `num_of_word` was not considered in previous studies, but it was included to take into consideration a discrepancy between the unit (word) handled by the CCG parsers and the unit (phrase) to which the reading time is assigned. `freq` is the frequencies of each phrasal unit that were estimated using Wikipedia word frequency generator[6] and National Institute for Japanese Language and Linguistics Web Japanese Corups (Asahara, Maekawa, Imada, Kato, & Konishi, 2014). `surp` is surprisal that was computed using GPT-2 models (Radford et al., 2019).[7] To determine the surprisal of each phrasal unit, following Wilcox, Gauthier, Hu, Qian, and Levy (2020), the cumulative sum of surprisal assigned to each sub-word was calculated. `phraseN`, `lineN`, and `screenN` are indexes of the positions of each phrase presented to the subjects. `prev_is_fixed` is whether the gaze to the previous phrasal unit is fixed or not. All numeric predictors were *z*-transformed.

First, the baseline model was fitted, and data points beyond three standard deviations were removed based on the distribution of residuals, which left 182,304 data points in English and 8,903 data points in Japanese for final statistical analysis.

We constructed four separate models by adding each of the three Composition Count-based and the one Node Count-based predictors to the baseline model individually to investigate how much each predictor contributes to predicting human reading times.

[5] The term *verb* refers to a complex category, where the resulted category is *S* and it includes some *NP*s as an argument like *S\NP* or *S\NP/NP* in English. In Japanese, it refers to a complex category, where the resulted category is *S*, it includes some *NP*s as an argument, and all slashes are backward such as *S\NP* or *S\NP\NP*.

[6] `https://github.com/IlyaSemenov/wikipedia-word-frequency`

[7] `https://huggingface.co/openai-community/gpt2` for English and `https://huggingface.co/rinna/japanese-gpt2-medium` for Japanese.

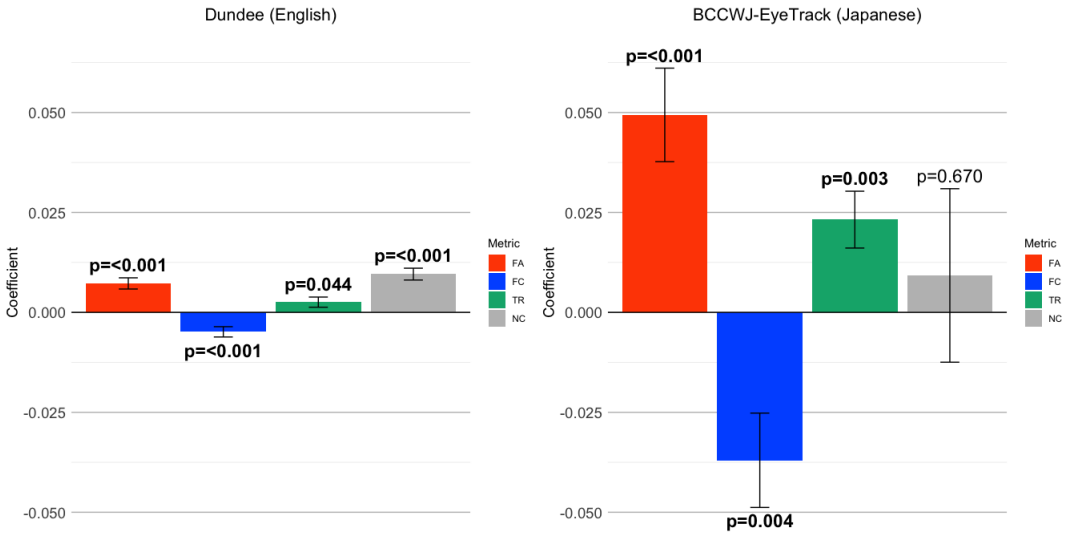


Figure 4: Results of Composition Counts and Node Count across the Dundee corpus (English) and BCCWJ-EyeTrack (Japanese). Bar graphs displaying the coefficients for the predictors based on Composition Counts or Node Counts within the Dundee corpus and BCCWJ-EyeTrack. The *x*-axis and *y*-axis represent complexity metrics and estimated coefficients, respectively. Colors correspond to complexity metrics: red = Function Application (FA), blue = Function Composition (FC), green = Type Raising (TR), and grey = Node Count (NC). Error bars indicate standard errors, on which *p*-values are presented. The log-likelihoods of each model were as follows: in the Dundee corpus, -123048 for FA, -123059 for FC, -123064 for TR, and -123045 for NC; in BCCWJ-EyeTrack, -8861 for FA, -8872 for FC, -8859 for TR, and -8876 for NC.

Results

The results of Composition Counts and Node Count across the Dundee corpus (English) and BCCWJ-EyeTrack (Japanese) are summarized in Figure 4. The *x*-axis and *y*-axis represent complexity metrics and estimated coefficients, respectively.[8] Colors correspond to each complexity metric. Error bars indicate standard errors, on which *p*-values are presented. In order to address the issue of multiple comparisons, the α-value was adjusted using the Holm-Bonferroni method.

There are three main results. First, all predictors based on Composition Count were statistically significant, demonstrating that distinct syntactic operations of CCG have independent and cross-linguistic contributions to predicting human reading times. Second, in contrast with Composition Count, the predictor based on Node Count reached statistical significance in English, but not in Japanese, indicating that Node Count turned out not to be robust cross-linguistically, which is rather surprising given that the effect of Node Count has been robustly observed for English in the previous literature. Finally, both FA and TR exhibited positive effects, while FC showed negative effects, with the relative magnitude of the effects being FA $>$ TR $>$ FC, despite the absolute magnitude being different across both languages.

[8]The magnitudes of the coefficients appear small, but this will be largely due to the log-transformation applied to the dependent variable, reading times. Compared to the coefficients of other explanatory variables, these values are not inherently small. Specifically, the scale of `surp` is 10^{-2}, and for `dependent`, it ranges form 10^{-3} to 10^{-2}.

Discussion

All Composition Counts we proposed in this study to more precisely quantify the computational costs of individual syntactic operations significantly predict human reading times in both English and Japanese, suggesting that the operations theoretically licensed in linguistics are directly applicable to human sentence processing. In addition, we found that, while Node Count in English contributed to the prediction of reading times, thus supporting prior studies that have utilized this metric in modeling neural data, its predictive utility does not extend to Japanese reading time data. This discrepancy, not previously reported in the literature, suggests Node Count is not a robust predictor for capturing human sentence processing costs. As to the effects of the three distinct Composition Counts we examined, the relative magnitudes were found to be consistent across these languages.[9] Furthermore, owing to the Composition Counts, we have been able to detect the processing costs of each syntactic operation. In the following, we will delve into theoretical discussions concerning the implications of our findings for these operations.

[9]Here, we have yet to ascertain why the absolute magnitudes of the coefficients are different between English and Japanese. One possible explanation may lie in the theoretical analyses that the CCG parsers of the two languages assume. Specifically, case assignment and inflection are analyzed syntactically in the Japanese CCG parser, but not in the English CCG parser. Another explanation would be related to the characteristics of the datasets; for example, Shain, van Schijndel, Futrell, Gibson, and Schuler (2016) suggested that the limited number of participants or complex structures in the Dundee corpus may obscure the detection of processing costs associated with human sentence processing.

Function Application (FA) is the most fundamental operation of syntactic operations in CCG, commonly used for integrating arguments. In both English and Japanese, the processing costs of FA are notably large and positive. FA exhibits similarities to the F-L+ operation of the left-corner parser utilized by van Schijndel and Schuler (2013) both in terms of theoretical properties and the timing of its application. The F-L+ operation, as van Schijndel and Schuler noted, bears resemblance to the *integration* process in Dependent Locality Theory (DLT, Gibson, 2000), wherein a processing cost is incurred during the integration of a word into an incomplete parse. While there may be variations in the predicted magnitude of these costs, the anticipated timing for the occurrence of processing load is similar. In DLT, the integration cost is known to incur a positive cost in reading tasks (Gibson, 2000; Grodner & Gibson, 2005; Gibson & Wu, 2013; Levy & Keller, 2013, but see Konieczny, 2000; Vasishth & Lewis, 2006). In corpus-based studies, while Demberg and Keller (2008) observed a negative effect of DLT in the Dundee corpus, Shain et al. (2016) found the effect to be positive in the syntactically complex Natural Stories Corpus (Futrell et al., 2018) (but see Shain & Schuler, 2018). These align with our finding on the positive processing cost of FA. Regarding the F-L+ operation, van Schijndel and Schuler found a significant negative cost in the Dundee corpus. However, they suggested that the operation might capture an anti-locality effect (Konieczny, 2000). Our model, incorporating dependency relations as one baseline predictor, attributes anti-locality effects to these relations, thereby interpreting FA costs as positive.

Function Composition (FC) is an operation that merges complex categories, typically applied to combine a subject noun phrase (*NP*) with a verb in English and a subject *NP* with an object *NP* in Japanese to an incomplete parse during incremental processing. Semantically, FC demands multiple beta-reduction steps, making it theoretically more complex than FA, which involves just a single such step. However, contrary to the intuition, the effect of FC was minimal and even negative in both English and Japanese. This finding challenges the assumption that theoretical computational complexity necessarily translates into higher cognitive processing costs. Nonetheless, the critical insight is that the frequency and timing of structural-building operations, including FC, can sufficiently predict reading times, which suggests that, as Berwick and Weinberg (1983) pointed out, the distinctions of grammatical rules may be preserved as distinctions of parsing operations.

Type Raising (TR) is a unary operation that is applied to *NP*s directly combined with verbs. The effect was found to be positive. This aligns with prior research showing the positive processing cost of Type Shifting, which is theoretically similar to TR, through neural responses to *complement* and *aspectual coercion* (e.g., Pylkkänen & McElree, 2007; J. Brennan & Pylkkänen, 2008). In addition, given that Type Shifting is limited to the semantic domain while TR, within the syntax-semantics transparent framework of CCG, also entails syntactic considerations and thereby encompasses Type Shifting, our findings suggest the general existence of TR applied to noun phrases as an operation that takes positive processing cost.

Furthermore, our results intriguingly suggest that FA, a theoretically fundamental operation, incurs the highest processing load, challenging the view of Pylkkänen and McElree (2006). They, adopting Heim and Kratzer's (1998) semantics, argue that FA should be both preferred and the least costly in processing compared to other operations like Predicate Modification (PM), citing psycholinguistic research that argument phrases are processed more easily than adjunct phrases (e.g., Clifton, Speer, & Abney, 1991; Schütze & Gibson, 1999; Kennison, 2002). In Heim and Kratzer's framework, FA is used for argument phrases, whereas PM is for adjunct phrases, thereby implying a processing preference for FA over PM. Pylkkänen and McElree also hypothesize that the processing costs might influence their structure-building preferences, predicting FA to be the least costly. However, contrary to their prediction, our findings suggest that FA is, in fact, the most costly operation. One possible reason is that Heim and Kratzer's semantics do not directly predict processing mechanisms during real-time sentence comprehension. Therefore, the observed processing efficiency for arguments over adjuncts might not translate to a preference for FA over PM. In CCG, FA is used in processing both types of phrases. If we consider a left-branching structure incrementally constructed, adjuncts attach at a structurally lower position than arguments, requiring more FA at that time, which may explain the dispreference of adjuncts. Hence, our study suggests that fundamental operations may not always be less costly, and the processing costs of building structures could underlie their preferential use.

Conclusion

In this paper, we introduced a novel complexity metric dubbed **Composition Count** to investigate (i) how much distinct syntactic operations of CCG contribute to predicting human reading times, and (ii) whether the same holds across languages. The results demonstrated that distinct syntactic operations of CCG have independent and cross-linguistic contributions to predicting human reading times: FA/TR and FC exhibited positive and negative effects, respectively, with the relative magnitude of the effects being FA $>$ TR $>$ FC. In contrast, Node Count turned out not to be robust cross-linguistically. These results strongly suggest the importance of Composition Count to dissociate distinct syntactic operations, not whole syntactic representations, and understand the computational system of human sentence processing.[10]

[10]Code for reproducing our experiments is available at `https://github.com/osekilab/CompositionCount`.

Acknowledgements

We thank Shinnosuke Isono for his helpful feedback. We also thank the three anonymous reviewers for their insightful comments. This work is supported by JSPS KAKENHI Grant Number 24H00087 and JST PRESTO Grant Number JPMJPR21C2 and JST SPRING Grant Number JPMJSP2108.

References

Abney, S. P., & Johnson, M. (1991). Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*, *20*(3), 233–250.

Asahara, M., Maekawa, K., Imada, M., Kato, S., & Konishi, H. (2014). Archiving and analysing techniques of the ultra-large-scale web-based corpus project of NINJAL, Japan. *Alexandria*, *25*(1–2), 129–148.

Asahara, M., & Matsumoto, Y. (2016). BCCWJ-DepPara: A syntactic annotation treebank on the 'Balanced Corpus of Contemporary Written Japanese'. In *Proceedings of the 12th workshop on Asian language resources (ALR12)* (pp. 49–58). Osaka, Japan: The COLING 2016 Organizing Committee.

Asahara, M., Ono, H., & Miyamoto, E. T. (2016). Reading-time annotations for "Balanced Corpus of Contemporary Written Japanese". In (pp. 684–694). Osaka, Japan: The COLING 2016 Organizing Committee.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Barrett, M., Agić, Ž., & Søgaard, A. (2015). The dundee treebank. In *Proceedings of the 14th international workshop on treebanks and linguistic theories (tlt14)* (pp. 242–248).

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Berwick, R. C., & Weinberg, A. S. (1983). The role of grammars in models of language use. *Cognition*, *13*(1), 1–61.

Bhattasali, S., Fabre, M., Luh, W.-M., Saied, H. A., Constant, M., Pallier, C., ... Hale, J. (2018). Localising memory retrieval and syntactic composition: an fMRI study of naturalistic language comprehension. *Language, Cognition and Neuroscience*, *34*(4), 491–510.

Brennan, J., & Pylkkänen, L. (2008). Processing events: behavioral and neuromagnetic correlates of Aspectual Coercion. *Brain and language*, *106*(2), 132–143.

Brennan, J. R., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and language*, *120*(2), 163–173.

Brennan, J. R., & Pylkkänen, L. (2017). MEG evidence for incremental sentence composition in the anterior temporal lobe. *Cognitive science*, *41*, 1515–1531.

Brennan, J. R., Stabler, E. P., Wagenen, S. E. V., Luh, W.-M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157–158*, 81–94.

Bresnan, J., & Kaplan, R. (1982). Introduction: Grammars as mental representations of language. In J. Bresnan (Ed.), *The mental representation of grammatical relations* (pp. i–lii). Cambridge, MA: MIT Press.

Chomsky, N. (1957). *Syntactic structures*. Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Clifton, C., Speer, S., & Abney, S. P. (1991). Parsing arguments: Phrase structure and argument structure as determinants of initial parsing decisions. *Journal of Memory and Language*, *30*(2), 251–271.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Fodor, J., Bever, T., & Garrett, M. (1974). *The psychology of language.* New York: McGraw Hill.

Fossum, V., & Levy, R. P. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics (CMCL 2012)* (pp. 61–69). Montréal, Canada.

Frazier, L. (1985). Syntactic complexity. In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 129–189). Cambridge University Press.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S., & Fedorenko, E. (2018). The natural stories corpus. In N. Calzolari et al. (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA).

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 95–126). The MIT Press.

Gibson, E., & Wu, H.-H. I. (2013). Processing chinese relative clauses in context. *Language and Cognitive Processes*, *28*(1-2), 125–155.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, *29*(2), 261–290.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics.*

Hale, J., Dyer, C., Kuncoro, A., & Brennan, J. (2018). Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2727–2736). Association for Computational Linguistics.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge University Press.

Heim, I., & Kratzer, A. (1998). *Semantics in generative*

grammar. Oxford: Blackwell.

Isono, S., & Hirose, Y. (2022). Locality effect before the verb as evidence of pre-verb reactivation. *The Japanese Society for Language Sciences 23rd Annual International Conference*.

Kamide, Y., & Mitchell, D. C. (1999). Incremental pre-head attachment in Japanese parsing. *Language and Cognitive Processes*, *14*(5-6), 631–662.

Kennedy, A., Hill, R., & Pynte, J. (2003). The dundee corpus. In *Proceedings of the 12th european conference on eye movement.*

Kennison, S. M. (2002). Comprehending noun phrase arguments and adjuncts. *Journal of Psycholinguistic Research*, *31*(1), 65–81.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29*, 627–645.

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In E. Blanco & W. Lu (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations* (pp. 66–71). Brussels, Belgium: Association for Computational Linguistics.

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Levy, R. P. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, R. P., & Keller, F. (2013). Expectation and locality effects in german verb-final structures. *Journal of Memory and Language*, *68*(2), 199–222.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Lewis, S., & Phillips, C. (2015). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research*, *44*(1), 27–46.

Li, J., Bhattasali, S., Zhang, S., Franzluebbers, B., Luh, W.-M., Spreng, R. N., … Hale, J. (2022). Le petit prince multilingual naturalistic fmri corpus. *Scientific Data*, *9*(530).

Li, J., & Hale, J. (2019). Grammatical predictors for fMRI time-courses. In R. C. Berwick & E. P. Stabler (Eds.), *Minimalist parsing* (pp. 159–173). Oxford University Press.

Marantz, A. (2005). Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review*, *22*(2-4), 429–445.

Mazuka, R., & Itoh, K. (1995). Can Japanese speakers be led down the garden path. *Japanese sentence processing*, 295–329.

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2, pp. 419–491). Wiley.

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., … Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, *114*(18), E3669–E3678.

Pylkkänen, L., & McElree, B. (2006). The syntax-semantics interface: On-line composition of sentence meaning. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics (second edition)* (Second Edition ed., pp. 539–579). London: Academic Press.

Pylkkänen, L., & McElree, B. (2007). An MEG Study of Silent Meaning. *Journal of Cognitive Neuroscience*, *19*(11), 1905–1921.

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners..

Resnik, P. (1992). Left-corner parsing and psychological plausibility. In *COLING 1992 volume 1: The 14th International Conference on Computational Linguistics.*

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 324–333).

Sag, I. A., & Wasow, T. (2011). Performance-compatible competence grammar. In R. D. Borsley & K. Börjars (Eds.), *Non-transformational syntax: Formal and explicit models of grammar* (pp. 359–377). Wiley-Blackwell.

Schütze, C. T., & Gibson, E. (1999). Argumenthood and english prepositional phrase attachment. *Journal of Memory and Language*, *40*(3), 409–431.

Shain, C., & Schuler, W. (2018). Deconvolutional time series regression: A technique for modeling temporally diffuse effects. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2679–2689). Brussels, Belgium: Association for Computational Linguistics.

Shain, C., van Schijndel, M., Futrell, R., Gibson, E., & Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In D. Brunato, F. Dell'Orletta, G. Venturi, T. François, & P. Blache (Eds.), *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)* (pp. 49–58). Osaka, Japan: The COLING 2016 Organizing Committee.

Stanojević, M., Bhattasali, S., Dunagan, D., Campanelli, L., Steedman, M., Brennan, J., & Hale, J. (2021). Modeling Incremental Language Comprehension in the Brain with Combinatory Categorial Grammar. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 23–38). Online: Association for Computational Linguistics.

Stanojević, M., Brennan, J. R., Dunagan, D., Steedman, M.,

& Hale, J. T. (2023). Modeling structure-building in the brain with CCG parsing and Large Language Models. *Cognitive Science*, *47*(7), e13312.

Stanojević, M., & Steedman, M. (2019). CCG Parsing Algorithm with Incremental Tree Rotation. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 228–239). Minneapolis, Minnesota: Association for Computational Linguistics.

Steedman, M. (2000). *The syntactic process*. MIT press.

van Schijndel, M., & Schuler, W. (2013). An analysis of frequency- and memory-based processing costs. In L. Vanderwende, H. Daumé III, & K. Kirchhoff (Eds.), *Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 95–105). Atlanta, Georgia: Association for Computational Linguistics.

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*(4), 767–794.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd annual meeting of the cognitive science society* (pp. 1707–1713).

Yoshikawa, M., Noji, H., & Matsumoto, Y. (2017). A* CCG parsing with a supertag and dependency factored model. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 277–287). Vancouver, Canada: Association for Computational Linguistics.