Chapter 15
Human language processing in comparative computational psycholinguistics

1 Introduction

Computational psycholinguistics of human language processing has attracted considerable attention in both experimental psycholinguistics and natural language processing (NLP), thanks to recent advances in machine learning and large corpora. In computational psycholinguistics, computational models are constructed from symbolic generative models and artificial neural networks developed in NLP and, through the lens of information-theoretic complexity metrics (Hale 2016), evaluated against human behavioral and neural data collected through psycholinguistic experiments. However, the previous literature has focused almost exclusively on European languages with typologically similar characteristics (Bender 2011), so that the question whether the established conclusions in computational psycholinguistics can be generalized across languages remains to be empirically addressed. In this chapter, we advocate the comparative approach to computational psycholinguistics dubbed comparative computational psycholinguistics, which constructs and evaluates computational models of human language processing from comparative perspectives.

This chapter is organized as follows. Section 2 reviews the pipeline of computational psycholinguistics and, with some issues raised from comparative perspectives, proposes comparative computational psycholinguistics, building on the previous literature on comparative psycholinguistics and computational typology.

Acknowledgments: This chapter originates from the talk delivered at the International Symposium on Issues in Japanese Psycholinguistics from Comparative Perspectives (IJPCP) held online on September 11–12, 2021. I would like to thank excellent collaborators such as Ryo Yoshida, Hiroshi Noji, Tatsuki Kuribayashi, Masayuki Asahara, and especially Masatoshi Koizumi for inviting me to speak at the above-mentioned symposium and contribute to this volume. I also want to thank Kohei Kajikawa and two anonymous reviewers for proofreading the manuscript and providing insightful comments. This work was supported by JST PRESTO Grant Number JPMJPR21C2 and NINJAL Collaborative Research Project “Computational Psycholinguistics of Language Processing with Large Corpora”. All remaining errors are my own.

Open Access. © 2023 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. https://doi.org/10.1515/9783110778946-015
Section 3 presents the results of modeling *hierarchical syntactic structure* with Recurrent Neural Network Grammars (Dyer et al. 2016), demonstrating that hierarchical syntactic structure universally makes computational models more human-like, while optimal parsing strategies may diverge with respect to head directionality (Yoshida, Noji, and Oseki 2021). Section 4 then provides the results of modeling *cue-based memory retrieval* with Transformer architectures (Vaswani et al. 2017; Merkx and Frank 2021), suggesting that Transformer architectures are too powerful for those languages with few long-distance dependencies, which can be rendered more human-like through context limitations (Kuribayashi et al. 2021, 2022). Section 5 concludes this chapter and remarks some future directions.

2 Computational psycholinguistics from comparative perspectives

2.1 What is computational psycholinguistics?

In computational psycholinguistics, computational models are constructed from symbolic generative models and artificial neural networks developed in NLP and evaluated against human behavioral and neural data collected through psycholinguistic experiments (Crocker 1996; Lewis 2003; Hale 2017). In this sense, computational psycholinguistics is an interdisciplinary approach to human language processing at the intersection of experimental psycholinguistics and NLP. From experimental psycholinguistics, on one hand, computational psycholinguistics inherits both the scientific goal to elucidate human language processing and the human behavioral and neural data to be modeled computationally, while experimental manipulations are performed over computational models (e.g. model architecture, training data, etc.), not experimental stimuli themselves as in experimental psycholinguistics (e.g. syntactic complexity, semantic plausibility, etc.). From NLP, on the other hand, computational psycholinguistics borrows computational models such as symbolic generative models and artificial neural networks, but as computational models of human language processing with serious scientific commitments, not as pure engineering solutions as in NLP.\(^2\)

---

\(^2\) Computational psycholinguistics is usually referred to as *cognitive modeling* in the NLP community. For example, the submission track for computational psycholinguistics at the Association for Computational Linguistics ("ACL") conferences is called “Linguistic Theories, Cognitive Modeling, and Psycholinguistics”, and the designated workshop for computational psycholinguistics is named “Cognitive Modeling and Computational Linguistics (CMCL)".
Specifically, the pipeline of computational psycholinguistics generally consists of three components, as summarized in Figure 1 (Brennan and Hale 2019).

1. **Models**: symbolic generative models, artificial neural networks

   ![Diagram of models and humans](image)

   **Figure 1**: Pipeline of computational psycholinguistics (Brennan and Hale 2019).³

The first component is **models**: computational models are constructed from symbolic generative models and artificial neural networks developed in NLP. For example, symbolic generative models include *language models* (LMs; computational models to estimate the probabilities of the words within the sentences) such as *n*-gram models which sequentially process sentences given *n*–1 previous words and context-free grammars (CFGs) which hierarchically process sentences given their syntactic structures. In addition, artificial neural networks range from classic recurrent neural networks (RRNs; Elman 1990), through long short-term memory networks (LSTMs; Hochreiter and Schmidhuber 1997), to recent Transformer architectures (Vaswani et al. 2017).⁴

The second component is **humans**: human behavioral and neural data are collected through psycholinguistic experiments to be predicted with the computational models.⁵ For instance, human behavioral data can be divided into offline

---

³ Figure 1 exemplifies human neural data, especially electroencephalography (EEG), but notice importantly for the purpose here that this pipeline of computational psycholinguistics can be equally applied to human behavioral data.

⁴ See Goldberg (2017) and Jurafsky and Martin (2022: Ch. 5–11) for the details of various architectures of artificial neural networks.

⁵ In practice, computational psycholinguistics proper does not collect human behavioral and neural data through psycholinguistic experiments, but rather employ publicly available language resources.
measures like acceptability judgments and online measures like self-paced reading and eye-tracking. Similarly, human neural data can be classified into electrophysiological techniques like electroencephalography (EEG) and magnetoencephalography (MEG) and hemodynamic techniques like functional magnetic resonance imaging (fMRI), where the former and latter techniques exhibit higher temporal and spatial resolutions, respectively.

The third component is **models × humans**: computational models and human data are compared to evaluate which computational model most successfully predicts human behavioral and neural data. Importantly, in order to bridge the gap between probabilities estimated from computational models and processing complexities collected from human behavioral and neural data, information-theoretic complexity metrics are employed as **linking hypotheses** (Hale 2016). Specifically, there are two prominent information-theoretic complexity metrics proposed in computational psycholinguistics. The first information-theoretic complexity metric is **surprisal** (Hale 2001; Levy 2008), the negative logarithmic probability of words \( w \) in context \( c \) as defined in (1a) which quantifies how surprising words will be in context, hypothesizing that lower probability, hence higher surprisal, links to higher processing complexity. The second information-theoretic complexity metric is **entropy reduction** (Hale 2006), the non-negative reduction of entropy between two probability distributions \( W \) over words in context \( c \) as defined in (1b) which quantifies how uncertain words will be in context, hypothesizing that the higher divergence between two probability distributions, hence higher entropy reduction, links to higher processing complexity.\(^6\)

\[(1) \text{ Information-theoretic complexity metrics (Hale 2016):}^7\]

- **Surprisal**: \( I(w) = -\log_2 p(w) \)
- **Entropy**: \( H(W) = -\sum_{w \in W} p(w) \log_2 p(w) \)

Unlike traditional complexity metrics like node and action counts (i.e. the number of nodes/actions traversed between two words; Miller and Chomsky 1963) which can be applied only to computational models with hierarchical syntactic structures, information-theoretic complexity metrics are theory-neutral with respect to rep-
resentational assumptions of computational models, hence “causal bottleneck” (Levy 2008).8

Finally, the important logic behind this pipeline of computational psycholinguistics is that the computational model that most successfully predicts human behavioral and neural data is argued to be the most “human-like” computational model relative to baseline computational models. This logic is sometimes called the constructive approach or abductive inference in related fields such as cognitive robotics (Taniguchi et al. 2019).

### 2.2 Issues with computational psycholinguistics

Despite the remarkable success thanks to recent advances in machine learning and large corpora, there exist several issues with computational psycholinguistics.9 One of the most urgent issues with the NLP community in general is that the previous literature has focused almost exclusively on European languages with typologically similar characteristics, especially Germanic languages like English. For example, Bender (2011) reported that, among the single-language studies published in the Association for Computational Linguistics (ACL 2008) and the European chapter of the Association for Computational Linguistics (EACL 2009), English accounted for 63% in ACL 2008 and 55% in EACL 2009, Germanic languages 71% in both ACL 2008 and EACL 2009, and surprisingly the European languages even 85% in ACL 2008 and 91% in EACL 2009.10

Unfortunately, computational psycholinguistics is not an exception: language resources naturalistically annotated with human behavioral and neural data such as Dundee Corpus (Kennedy and Pynte 2005) are mostly available in European languages, especially English. Language resources naturalistically annotated through human behavioral and neural data are summarized in Table 1 (Oseki and Asahara 2020). Accordingly, model evaluations of computational psycholinguistics have largely been limited to European languages, so that the question whether the established conclusions in computational psycholinguistics can be generalized

---

8 Those information-theoretic complexity metrics might be collectively called the Information(al) Theory of Complexity (ITC), in contrast with the traditional Derivational Theory of Complexity (DTC; Miller and Chomsky 1963; Fodor, Bever, and Garrett 1974).

9 Other issues with computational psycholinguistics include, but are not limited to: the plausibility of artificial neural networks as computational models of human language processing, the adequacy of information-theoretic complexity metrics as linking hypotheses between computational models and human data.

10 As Bender (2011: fn.18) herself correctly pointed out, given the recent trend in multilingual models and low-resource languages in the NLP community, the situation might have improved over the past years.
across languages remains to be empirically addressed. Therefore, comparative perspectives need to be brought into the field of computational psycholinguistics.

Table 1: Language resources naturalistically annotated through human behavioral and neural data (Oseki and Asahara 2020).

<table>
<thead>
<tr>
<th>Language Resource</th>
<th>Language</th>
<th>Self-Paced</th>
<th>Eye-Track</th>
<th>EEG</th>
<th>fMRI</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dundee Corpus</td>
<td>English/French</td>
<td>✓ (10)</td>
<td>✓ (10)</td>
<td></td>
<td></td>
<td>Kennedy and Pynte (2005)</td>
</tr>
<tr>
<td>Potsdam Sentence Corpus</td>
<td>German</td>
<td>✓ (144)</td>
<td></td>
<td></td>
<td></td>
<td>Kliegl et al. (2006)</td>
</tr>
<tr>
<td>Natural Stories Corpus</td>
<td>English</td>
<td>✓ (19)</td>
<td></td>
<td>✓ (78)</td>
<td></td>
<td>Futrell et al. (2018)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Shain et al. (2019)</td>
</tr>
<tr>
<td>Ghent Eye-Tracking Corpus (GECO)</td>
<td>English/Dutch</td>
<td>✓ (14)</td>
<td>✓ (19)</td>
<td></td>
<td></td>
<td>Cop et al. (2017)</td>
</tr>
<tr>
<td>UCL Corpus</td>
<td>English</td>
<td>✓ (117)</td>
<td>✓ (43)</td>
<td>✓ (24)</td>
<td></td>
<td>Frank et al. (2013)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Frank et al. (2015)</td>
</tr>
<tr>
<td>Alice Corpus</td>
<td>English</td>
<td>✓ (52)</td>
<td></td>
<td>✓ (29)</td>
<td></td>
<td>Brennan and Hale (2019)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Brennan et al. (2016)</td>
</tr>
<tr>
<td>Zurich Cognitive Language Processing Corpus (ZuCo)</td>
<td>English</td>
<td>✓ (12)</td>
<td>✓ (12)</td>
<td></td>
<td></td>
<td>Hollenstein et al. (2018)</td>
</tr>
<tr>
<td>BCCWJ-EEG</td>
<td>Japanese</td>
<td></td>
<td></td>
<td>✓ (40)</td>
<td></td>
<td>Oseki and Asahara (2020)</td>
</tr>
</tbody>
</table>

2.3 Comparative computational psycholinguistics

In order to address the question raised in the subsection above, we advocate the comparative approach to computational psycholinguistics dubbed *comparative computational psycholinguistics*. In comparative computational psycholinguistics, computational models are constructed from symbolic generative models and artificial neural networks developed in NLP and, through the lens of information-theoretic complexity metrics (Hale 2016), evaluated against human behavioral and neural data collected through psycholinguistic experiments just like computational psycholinguistics, but crucially from *comparative* perspectives.

In order to clarify the essence of comparative computational psycholinguistics, here we will review several related approaches proposed in the previous lit-
erature, which share some (but not all) of the key features with comparative computational psycholinguistics: (i) computational psycholinguistics, (ii) comparative psycholinguistics, and (iii) computational typology. First of all, as already pointed out in the subsection above, computational psycholinguistics constructs and evaluates computational models of human language processing against human behavioral and neural data (Crocker 1996; Lewis 2003; Hale 2017), but lacks the comparative perspectives, hence the strong bias towards the European languages. Second, comparative psycholinguistics elucidates human language processing from comparative perspectives (Grillo and Costa 2014; Chacón et al. 2016), but through psycholinguistic experiments with experimental manipulations performed over experimental stimuli, not computational models. Finally, the computational approach to linguistic typology called computational typology has recently emerged in the NLP community which employs both computational models and massively comparative perspectives (Ackerman and Malouf 2013; Futrell, Levy, and Gibson 2020), but investigates linguistic universals, not human language processing. These related approaches taken together, comparative computational psycholinguistics can be regarded as the new interdisciplinary approach to human language processing from both computational and comparative perspectives. Related approaches with comparative computational psycholinguistics are summarized in Table 2.

Table 2: Related approaches with comparative computational psycholinguistics.

<table>
<thead>
<tr>
<th></th>
<th>Psycholinguistic</th>
<th>Computational</th>
<th>Comparative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computational psycholinguistics</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Comparative psycholinguistics</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Computational typology</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Comparative computational psycholinguistics</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

To recapitulate, this section reviewed the method and the problem with computational psycholinguistics and then proposed comparative computational psycholinguistics, in systematic comparisons with related approaches proposed in the previous literature. In the next two sections, we will present the results of modeling hierarchical syntactic structure (Section 3; Yoshida, Noji, and Oseki 2021) and cue-based memory retrieval (Section 4; Kuribayashi et al. 2021), where computational models are constructed and evaluated against human behavioral data from comparative perspectives, with particular emphasis on typologically different languages such as English and Japanese. Specifically, hierarchical syntactic structure and cue-based memory retrieval will be modeled with Recurrent Neural Network Grammars (Dyer et al. 2016) and Transformer architectures (Vaswani et al. 2017), respectively.
3 Modeling hierarchical syntactic structure

Linguistic theories assume that the grammar represents sentences as *hierarchical syntactic structure*, not just linear word sequence (Chomsky 1957; Everaert et al. 2015). Accordingly, given the Competence Hypothesis that the relationship between the grammar and the parser is maximally transparent (Chomsky 1965), psycholinguistic theories also hypothesize that the parser processes sentences hierarchically (i.e. building hierarchical structures), not just sequentially (i.e. tracking word sequences).¹¹

In sharp contrast with those linguistic and psycholinguistic theories, the NLP community has been dominated by recurrent neural networks (RNNs; Elman 1990) with the *recurrence* mechanism which propagates information through time and, despite the lack of explicit hierarchical structures, successfully processes sentences. One of the reasonable hypotheses for this success is that RNNs inductively learn hierarchical representations and *implicitly* represent sentences as hierarchical structures, as evidenced by acceptability judgment experiments where RNNs can capture long-distance dependencies like subject-verb agreement (Linzen, Dupoux, and Goldberg 2016; Warstadt, Singh, and Bowman 2019).

Nevertheless, the previous literature has also implemented the RNN architectures that *explicitly* represent sentences as hierarchical structures and, interestingly, demonstrated that those RNN architectures with syntactic supervision outperform RNNs in explaining long-distance dependencies (Kuncoro et al. 2018; Wilcox et al. 2019) and even human neural responses (Hale et al. 2018). The representative RNN architecture with syntactic supervision is *Recurrent Neural Network Grammars* (RNNGs; Dyer et al. 2016), a deep generative model which models not only sentences but also their hierarchical structures. Specifically, RNNGs adopt the stack LSTM (Dyer et al. 2015), an augmentation of Long Short-Term Memory networks (LSTMs; Hochreiter and Schmidhuber 1997) originally developed for dependency parsing, and estimate probability distributions over three parsing actions as defined in (2), where the composition function of REDUCE is bidirectional LSTMs. The architecture of RNNGs is summarized in Figure 2.

---

¹¹ Since the failure of the Derivational Theory of Complexity (DTC) that processing complexity is a function of the number of derivational steps to generate the sentences in question (Miller and Chomsky 1963; Fodor, Bever, and Garrett 1974), psycholinguistic theories have ramified into those with and without the Competence Hypothesis, the latter of which claim that human language processing is insensitive to hierarchical structures (Frank and Bod 2011; Frank, Bod, and Christiansen 2012).
(2) Parsing actions of Recurrent Neural Network Grammars (Dyer et al. 2016):
   a. NT: introduce nonterminal symbols (e.g. NP, VP)
   b. GEN: generate terminal symbols (e.g. the, hungry, cat)
   c. REDUCE: compose symbols into phrases via composition function

Figure 2: Architecture of Recurrent Neural Network Grammars (Dyer et al. 2016).\textsuperscript{12}

However, RNNGs have been evaluated only in English, so that whether hierarchical syntactic structure universally makes computational models more human-like remains to be empirically verified. Moreover, the vanilla RNNG implemented by Dyer et al. (2016) adopts the top-down parsing strategy, but given the consensus in the parsing literature (Abney and Johnson 1991; Resnik 1992) that the top-down parsing strategy works effectively for right-branching structures instantiated by head-initial languages like English, the performance of RNNGs might have been overestimated by the accidental match between top-down parsing and head directionality. Therefore, in order to assess the robustness of RNNGs across languages, we should evaluate RNNGs with both top-down and left-corner parsing strategies against head-final languages like Japanese.

\textsuperscript{12} The vanilla RNNG implemented in DyNet by Dyer et al. (2016) estimates probability distributions over parsing actions collectively through stack ($S_t$), output buffer ($T_t$), and history of actions ($a_{<t}$), but the recent batched RNNG implemented in PyTorch by Noji and Oseki (2021) used in this chapter adopts the stack-only RNNG (Kuncoro et al. 2017) which estimates probability distributions over parsing actions based solely on stack ($S_t$).
3.1 Methods

The experiments follow the pipeline of computational psycholinguistics in Figure 1. First, three computational models were constructed from symbolic generative models and artificial neural networks developed in NLP and trained on the NINJAL Parsed Corpus of Modern Japanese (NPCMJ): Long Short-Term Memory (LSTM; Hochreiter and Schmidhuber 1997) and Recurrent Neural Network Grammars (RNNGs; Dyer et al. 2016) with top-down (Top-down RNNG) and left-corner (Left-corner RNNG) parsing strategies. Second, human behavioral data were collected through psycholinguistic experiments to be predicted with those computational models: BCCWJ-EyeTrack (Asahara, Ono, and Miyamoto 2016). Finally, three computational models and human behavioral data were compared to evaluate which computational model most successfully predicts human behavioral data, through information-theoretic complexity metrics like surprisal (Hale 2001; Levy 2008) and linear mixed-effects regression models (Baayen, Davidson, and Bates 2008). Following Goodkind and Bicknell (2018), perplexity (PPL; how successfully computational models predict next words) and psychometric predictive power (PPP; how successfully computational models predict human data relative to the baseline model with control variables like length and frequency) were adopted as the evaluation metrics.

3.2 Results

The results of modeling hierarchical syntactic structure are summarized in Figure 3 (Yoshida, Noji, and Oseki 2021). There are three important observations. First, RNNGs with both top-down and left-corner parsing strategies generally outperform LSTMs, demonstrating that hierarchical syntactic structure universally makes computational models more human-like (Kuncoro et al. 2018; Wilcox et al. 2019). Second, among those RNNGs, left-corner parsing strategies outperform top-down parsing strategies, indicating that optimal parsing strategies may diverge with respect to head directionality (Abney and Johnson 1991; Resnik 1992). Finally, there seems to be a linear correlation between perplexity and psychometric predictive power for RNNGs, whereas this linear correlation does not hold for LSTMs, contradicting the established conclusion (Goodkind and Bicknell 2018).
3.3 Summary and discussion

In summary, this section presented the results of modeling hierarchical syntactic structure with Recurrent Neural Network Grammars (Dyer et al. 2016), demonstrating that hierarchical syntactic structure universally makes computational models more human-like, while optimal parsing strategies may vary with respect to head directionality (Yoshida, Noji, and Oseki 2021). The main results are summarized below.

- Hierarchical syntactic structure universally makes computational models more human-like (Kuncoro et al. 2018; Wilcox et al. 2019).

---

13 See Yoshida, Noji and Oseki (2021) for further experimental manipulations on action beam size and parsing accuracy, which we omitted due to space limitations.
Optimal parsing strategies may vary with respect to head directionality (Abney and Johnson 1991; Resnik 1992).

- Perplexity and psychometric predictive power are linearly correlated for RNNGs, but not for LSTMs (Goodkind and Bicknell 2018).

In the next section, assuming that human language processing is not only expectation-based but also memory-based, we provide the results of modeling cue-based memory retrieval with Transformer architectures (Vaswani et al. 2017; Merkx and Frank 2021).

4 Modeling cue-based memory retrieval

In addition to hierarchical syntactic structure discussed in the previous section, the memory mechanism called cue-based memory retrieval has been proposed in the psycholinguistic literature (Lewis and Vasishth 2005; Lewis, Vasishth, and Van Dyke 2006). For example, in long-distance dependencies like subject-verb agreement, subjects should be stored in memory and selectively retrieved at verbs through such retrieval cues as number and person features of the subjects in order to correctly inflect the verbs.

In the same vein, the memory mechanism has also been implemented into artificial neural networks in the NLP literature. Specifically, RNNs involve the recurrence mechanism (Elman 1990) which propagates information through time but cannot “remember” information for long time, namely the vanishing gradient problem, while LSTMs employ the gate mechanism which not only “remembers” but also effectively “forgets” information through time, capturing long-distance dependencies like subject-verb agreement (Linzen, Dupoux, and Goldberg 2016). More recently, building on various insights from machine translation, Transformer architectures have dominated the NLP community (Vaswani et al. 2017) and achieved the state-of-the-art performance on various downstream tasks. The key innovation of Transformer architectures is the attention mechanism which dispenses with the recurrence and gate mechanisms and selectively attends previous information, and importantly has been cognitively interpreted as a computational model of human cue-based memory retrieval (Merkx and Frank 2021). The architecture of Transformers is summarized in Figure 4, in comparison with RNNs.

14 Cue-based memory retrieval has been computationally implemented within the framework of the cognitive architecture called Adaptive Control of Thought–Rational (ACT–R; Anderson 1983, 2007).
However, while Transformer architectures seem to be cognitively plausible for European languages like English with various long-distance dependencies (e.g. subject-verb agreement, wh-movement) and the so-called locality effect (i.e. local dependencies are less costly), whether this established conclusion can be transported to typologically different languages is not self-evident. That is, Transformer architectures might be too powerful for Asian languages like Japanese with few long-distance dependencies (e.g. no subject-verb agreement, no wh-movement) and the opposite anti-locality effects (i.e. non-local dependencies are less costly). Therefore, in order to assess the cognitive plausibility of Transformer architectures as a computational model of cue-based memory retrieval, we should evaluate Transformer architectures against those languages like Japanese.

4.1 Methods

The experiments here also follow the same pipeline of computational psycholinguistics in Figure 1. First, four computational models were constructed from symbolic generative models and artificial neural networks developed in NLP and trained on Wikipedia articles in both English and Japanese: n-gram models (N-gram, where \( n = \{3, 4, 5\} \)), Long Short-Term Memory (LSTM; Hochreiter and Schmidhuber 1997), and Transformer architectures (Vaswani et al. 2017) with large (Trans-lg) and small (Trans-sm) numbers of hyperparameters. Second, just like the previous section, human behavioral data were collected through psycholinguistic experiments to be predicted with those computational models: Dundee Corpus for English (Kennedy and Pynte 2005) and BCCWJ-EyeTrack for Japanese (Asahara, Ono, and Miyamoto 2016). Finally, four computational models and human behavioral data were compared to evaluate which computational model most successfully predicts human behavioral data, through information-theoretic complexity metrics like surprisal.
(Hale 2001; Levy 2008) and linear mixed-effects regression models (Baayen, Davidson, and Bates 2008). The evaluation metrics were the same as the previous section.

### 4.2 Results

The results of modeling cue-based memory retrieval are summarized in Figure 5 (Kuribayashi et al. 2021). For the Dundee Corpus in English, the correlation between perplexity and psychometric predictive power was negative, corroborating the established conclusion that Transformer architectures are cognitively plausible for European languages like English with various long-distance dependencies. In contrast, for the BCCWJ-EyeTrack in Japanese, while the correlation between perplexity and psychometric predictive power was negative with perplexity > 400, the correlation became positive with perplexity < 400, suggesting that Transformer architectures might be too powerful for Asian languages like Japanese with few long-distance dependencies.

![Figure 5: Results of modeling cue-based memory retrieval (Kuribayashi et al. 2021).](image)

If the attention mechanism of Transformer architectures is too powerful for those languages with few long-distance dependencies which do not require such “skilled” cue-based memory retrieval, the prediction is that context limitations make Transformer architectures only accessible to local information and thus more human-like in those languages with few long-distance dependencies. The results of context limitations are summarized in Figure 6 (Kuribayashi et al. 2022), where surprisal

---

15 See Kuribayashi et al. (2021) for further experimental manipulations on number of updates and data size, which we omitted due to space limitations.
of Transformer architectures is computed given \( n-1 \) previous words as in \( n \)-gram models. Interestingly, when the same architecture is compared (e.g., GPT2-xs-Wiki), while psychometric predictive power did not change significantly through context limitations in English, context limitations can render Transformer architectures more human-like in Japanese.

![Figure 6: Results of context limitations (Kuribayashi et al. 2022).](image)

### 4.3 Summary and discussion

In summary, this section provided the results of modeling cue-based memory retrieval with Transformer architectures (Vaswani et al. 2017; Merkx and Frank 2021), suggesting that Transformer architectures are too powerful for those languages with few long-distance dependencies, which can be rendered more human-like through context limitations (Kuribayashi et al. 2021, 2022). The main results are summarized below:

- Transformer architectures are cognitively plausible for European languages like English with various long-distance dependencies (Merkx and Frank 2021).
- In contrast, Transformer architectures are too powerful for Asian languages like Japanese with few long-distance dependencies (Kuribayashi et al. 2021).
- Context limitations can render Transformer architectures more human-like in Japanese (Kuribayashi et al. 2022).

Now several theoretical implications will be discussed in light of the results above. First, as pointed out above, while human language processing has traditionally been assumed to be both expectation-based (“look-ahead” prediction of next words)

---

**Note:**

16 Note that the y-axes in Figures 3, 5, 6 all represent psychometric predictive power of the computational models, but their scales are not directly comparable due to various differences in training data, test data, computational models, among others.
and memory-based ("look-behind" retrieval of previous words), hence the radical
dichotomy between expectation-based and memory-based theories in psycholin-
guistics, those two theories should not be mutually exclusive, merely reflecting dif-
ferent aspects of human language processing (Demberg and Keller 2008; Futrell,
Gibson, and Levy 2020). Second, the Transformer architectures with context limi-
tations can be regarded as a hybrid computational model of expectation-based and
memory-based theories, suggesting the possibility that cue-based memory retrieval
itself is universal, while what counts as “cue” is parametrized across languages.

5 Conclusion

To summarize, this chapter advocated the comparative approach to computational
psycholinguistics dubbed comparative computational psycholinguistics, which con-
structs and evaluates computational models of human language processing from
comparative perspectives. Specifically, we presented the results of modeling hier-
archical syntactic structure with Recurrent Neural Network Grammars (Dyer et al.
2016), demonstrating that hierarchical syntactic structure universally makes comput-
tional models more human-like, though optimal parsing strategies may vary with
respect to head directionality (Yoshida, Noji, and Oseki 2021). Then, we provided
the results of modeling cue-based memory retrieval with Transformer architectures
(Vaswani et al. 2017; Merkx and Frank 2021), suggesting that Transformer architec-
tures are too powerful for those languages with few long-distance dependencies,
which can be rendered more human-like through context limitations (Kuribayashi
et al. 2021, 2022). For future directions, this comparative approach to computational
psycholinguistics should be extended to (i) typologically more diverse languages like
Kaqchikel Maya and Tongan (Koizumi et al. 2014) and (ii) human neural data like
EEG/MEG and fMRI (Hale et al. 2022).

In conclusion, we believe that comparative computational psycholinguistics
will be a promising approach to human language processing from both computa-
tional and comparative perspectives, towards machines that process natural lan-
guages like humans.
References


