Syntactic dependency length shaped by strategic memory allocation

Weijie Xu
University of California, Irvine
weijie.xu@uci.edu

Richard Futrell
University of California, Irvine
rfutrell@uci.edu

Abstract

Human processing of nonlocal syntactic dependencies requires the engagement of limited working memory for encoding, maintenance, and retrieval. This process creates an evolutionary pressure for language to be structured in a way that keeps the subparts of a dependency closer to each other, an efficiency principle termed *dependency locality*. The current study proposes that such a dependency locality pressure can be modulated by the surprisal of the antecedent, defined as the first part of a dependency, due to strategic allocation of working memory. In particular, antecedents with novel and unpredictable information are prioritized for memory encoding, receiving more robust representation against memory interference and decay, and thus are more capable of handling longer dependency length. We examine this claim by analyzing dependency corpora of 11 languages, with word surprisal generated from GPT-3 language model. In support of our hypothesis, we find evidence for a positive correlation between dependency length and the antecedent surprisal in most of the languages in our analyses. A closer look into the dependencies with core arguments shows that this correlation consistently holds for subject relations but not for object relations.

1 Introduction

Language processing requires efficient use of bounded cognitive systems, creating evolutionary pressures that have been argued to shape the structure of human language (Gibson et al., 2019). In the domain of syntax, one source of evidence for this idea comes from the principle of dependency locality, which holds that linguistic units connected in a syntactic dependency tend to stay close in linear order, due to the limited resources of working memory (WM) to hold the subparts of a non-local dependency in working memory (Hawkins, 1994; Gibson, 1998, 2000; Ferrer-i-Cancho, 2004; Liu, 2008; Futrell et al., 2015, 2019; Temperley and Gildea, 2018; Futrell et al., 2020). With this basic finding, a natural next step is to see how far this efficiency-based account for dependency locality can go, with a more and more realistic characterization of the nature and constraints of WM.

We explore strategic memory allocation as one such constraint that may further shape the structure of syntactic dependencies. The idea is that limited WM resources are strategically allocated, subject to a trade-off between the economical investment of WM on each linguistic unit stored, and the minimization of potential cost in future processing tasks (Lieder and Griffiths, 2020; Lewis et al., 2014; Gershman et al., 2015). Specifically, we propose that linguistic units with novel and unpredictable information should receive prioritized WM resources for encoding and storage in memory, thus yielding more robust representation against memory interference and decay.

The result of this strategic memory allocation is that when an antecedent carries novel and unpredictable information, it can tolerate longer dependency length. In the current study, we test this hypothesis in 11 languages: Amharic, Danish, English, German, Italian, Japanese, Korean, Mandarin, Russian, Spanish, and Turkish. To preview the results, across all the dependency types, we find a general positive correlation between antecedent surprisal and dependency length for more than half of the languages in our analysis. A closer look into the dependencies with core arguments demonstrates that the effect consistently emerges for most of the languages in subject relations, but not in object relations.

2 Background

2.1 Dependency Locality

At the individual level, the processing of sentences with nonlocal dependencies requires active engage-

Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2024), pages 1–9

Language	Corpus	Genre	All Depends	Subject	Object
Amharic	ATT (Seyoum et al., 2018)	doc-by-doc	4,164	643	525
Danish	DDT (Johannsen et al., 2015)	sent-by-sent	45,976	4,203	3,963
English	GUM (Zeldes, 2017)	doc-by-doc	89,947	7,881	7,296
German	GSD (McDonald et al., 2013)	sent-by-sent	155,480	9,602	8,474
Italian	ISDT (Bosco et al., 2013)	doc-by-doc	208,939	10,323	11,735
Japanese	GSD (Tanaka et al., 2016)	sent-by-sent	113,771	5,005	4,018
Korean	Kaist (Chun et al., 2018)	doc-by-doc	154,609	9,855	24,690
Mandarin	GSDSimp (Nivre et al., 2020)	sent-by-sent	63,456	5,538	7,576
Russian	SynTagRus (Droganova et al., 2018)	doc-by-doc	329,745	32,822	25,065
Spanish	AnCora (Taulé et al., 2008)	doc-by-doc	333,728	21,472	31,143
Turkish	BOUN (Marşan et al., 2022)	sent-by-sent	45,914	3,861	4,680

Table 1: Dependency corpora used as datasets. 'Genre' refers to whether the texts in the corpus are organized as independent sentences (sent-by-sent), or as documents with larger coherent discourse size (doc-by-doc). 'All Depends' indicates the number of all the dependencies after data exclusion. 'Subject' is a subset of 'All Depends' and indicates the number of dependencies with subject relations. 'Object' indicates the number of dependencies with object relations.The original Russian corpus has over 1.2M tokens with over 600 documents; we randomly sampled 300 documents from the original corpus in our analysis in order to save the computational power.

ment of working memory. Consider the sentence with a nonlocal dependency as in (1b) compared to (1a). The language user needs to maintain the antecedent "nurse" active in WM for a longer period of time until it is retrieved later at the retrieval site "supervised." Under the Dependency Locality Theory (Gibson, 1998, 2000), the integration of the second part of the dependency should become increasingly difficult as the dependency length increases. This effect has been confirmed empirically in reading time studies (Bartek et al., 2011; Grodner and Gibson, 2005). This locality effect could be due to the memory decay of the antecedent's representation over time, or due to cumulative similarity-based interference introduced by the intervening materials between head and dependent (Lewis and Vasishth, 2005; Vasishth et al., 2019).

(1) a. The *nurse* supervised the administrator...
b. The *nurse* who was from the clinic in downtown LA supervised the administrator...

At the population level, this processing constraint functions as an evolutionary pressure that shapes language structure. It has been observed crosslinguistically that word order reflects the minimization of dependency length in general (Hawkins, 1994, 2004, 2014; Ferrer-i-Cancho, 2004; Liu, 2008; Futrell et al., 2015, 2020). For example, Futrell et al. (2020) point out the explanatory power of dependency locality principle for multiple typological phenomena, such as the contiguity of constituents, short-before-long and long-before-short constituent ordering preference, and the consistency in head direction.

2.2 Strategic Memory Allocation

Despite the general constraint of the limited memory capacity, WM is a highly flexible system that is dynamically optimized for the relevant cognitive tasks at hand or in the future (Sims et al., 2012; Sims, 2016; Van den Berg and Ma, 2018; Jakob and Gershman, 2023). One instantiation of this dynamic optimization of WM can be the strategic memory allocation: WM resources such as attention can be dynamically and strategically allocated in a way that prioritizes the information with novel and unexpected content given its context, resulting in higher memory precision and representation fidelity (Bruning and Lewis-Peacock, 2020).

Empirically, in the domain of language processing, deeper encoding for more informative referents have been shown to facilitate their retrieval later at the other side of the dependency (Hofmeister, 2011; Hofmeister and Vasishth, 2014; Karimi et al., 2019; Troyer et al., 2016). Theoretically, a more predictable unit is *a priori* more likely to be reconstructed successfully even if it is lost from memory, and thus if only a limited number of units can be stored, it would be less important to store the predictable ones, a dynamic observed in the sentence processing model of Hahn et al. (2022).

If WM resources can be dynamically and strategically allocated to prioritize novel and unexpected information, this can potentially shape the struc-

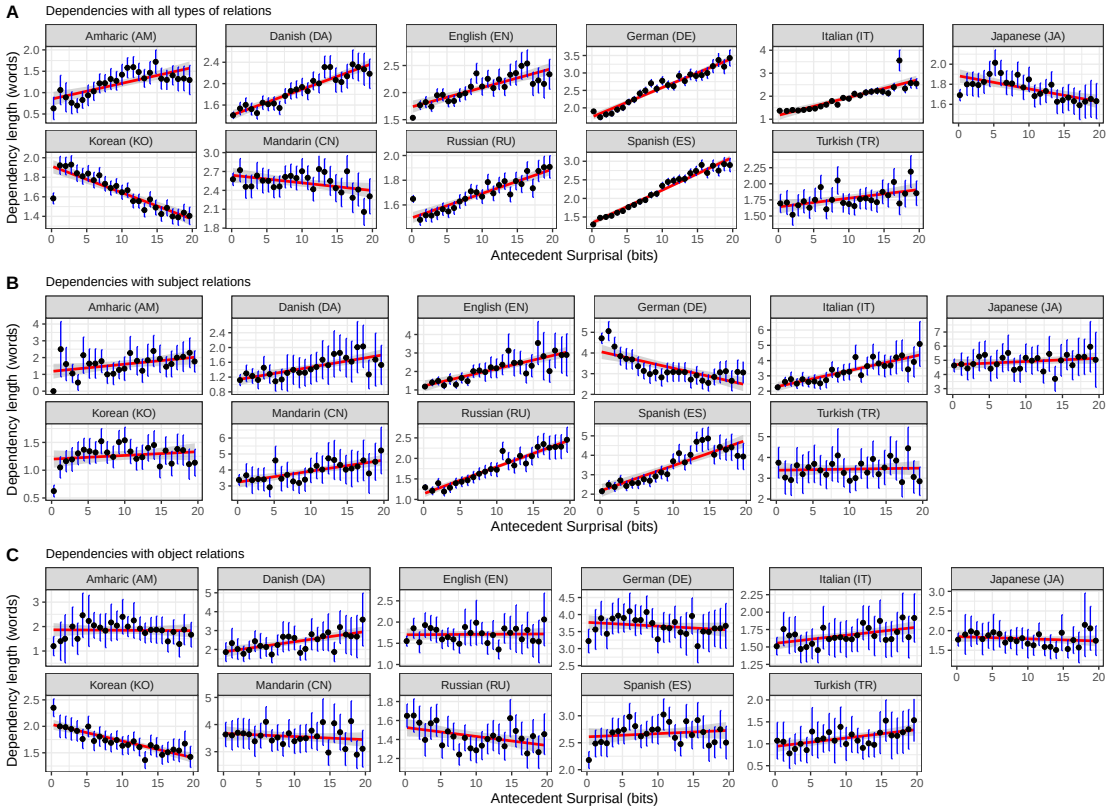

Figure 1: Average orthographic dependency length as a function of antecedent surprisal. Surprisal is binned into 25 categories, and the mean dependency length within each category is shown in black with a 95% confidence interval in blue. A linear fit to these points is shown in red.

ture of language by modulating the pressure of dependency locality. When the antecedent is less predictable but more informative, it should receive more WM resources and thus maintain a more robust memory representation, making it less likely to go through memory decay before it needs to be retrieved from memory at the other side of the dependency. Consequently, dependencies with less predictable antecedents are able to tolerate more intervening materials before the retrieval site, resulting in less pressure to put the subparts of a dependency local to each other, hence more likely to have longer dependency length.

3 Method

3.1 Data

We examine our hypothesis using the corpora taken from Universal Dependencies (UD) release 2.11 (Nivre et al., 2020), as described in Table 1. Some UD corpora consist of independent sentences, while others are organized document by document, thus providing longer and enriched discourse context for each token.

Token surprisal. For each token w in the dependency corpora, we obtain its surprisal $-\ln p(w|c)$ given the preceding context c using the GPT-3 base language model (`text-davinci-001`; Brown et al., 2020). We use the maximally allowed context window in the corresponding document or sentence. Due to the recent advancements in the performance of language models, they are increasingly applied to approximate human predictions in psycholinguistics literature (Levy, 2008). The surprisal generated from these models highly correlates with human processing difficulty indexed by behavioral measures such as reading times (Smith and Levy, 2013; Goodkind and Bicknell, 2018; Hao et al., 2020; Wilcox et al., 2020; Hoover et al., 2023), and this relationship has been shown to hold cross-linguistically (Wilcox et al., 2023; Xu et al., 2023).

Data transformation. Flat structures (e.g. foreign phrases, multiword proper names, fixed expressions, etc.) are merged such that the surprisal of the whole structure is the sum of all its components, and that the first word in the structure is treated as the head when calculating the dependency length with other sentence elements. Sen-

	Full Dataset		Subject Relations		Object Relations	
Language	L (words)	L (surprisal)	L (words)	L (surprisal)	L (words)	L (surprisal)
Amharic	$p = 0.175$	$+$	$+$	$p = 0.186$	$-$	$p = 0.876$
Danish	$+$	$+$	$+$	$+$	$p = 0.447$	$+$
English	$+$	$+$	$+$	$+$	$p = 0.743$	$+$
German	$+$	$+$	$-$	$-$	$-$	$-$
Italian	$+$	$+$	$+$	$+$	$p = 0.093$	$+$
Japanese	$p = 0.416$	$p = 0.775$	$p = 0.088$	$p = 0.985$	$p = 0.21$	$p = 0.94$
Korean	$-$	$-$	$p = 0.072$	$p = 0.156$	$-$	$-$
Mandarin	$p = 0.062$	$p = 0.331$	$+$	$+$	$-$	$p = 0.359$
Russian	$p = 0.395$	$p = 0.050$	$+$	$+$	$-$	$p = 0.454$
Spanish	$+$	$+$	$+$	$+$	$p = 0.058$	$+$
Turkish	$p = 0.161$	$p = 0.784$	$-$	$p = 0.59$	$p = 0.384$	$p = 0.083$

Table 2: Summary of statistical results for the effect of antecedent of surprisal on dependency length. "$+$" indicates a significant (at $p < 0.05$) positive correlation between antecedent surprisal and dependency length, while "$-$" indicates a significant negative correlation; p values are presented if the effect is not significant.

tences that are too short may have limited room for the dependency length to vary, so we exclude sentences containing less than five words. We exclude tokens that are punctuation. We also exclude tokens with a surprisal value greater than 20 bits. We then extract all the dependencies in which both the head and the dependent are spared from data exclusion. We also analyze two subsets of these dependencies: 1) a subset that only includes subject relations (marked as `nsubj` and `csubj`); and 2) a subset that only includes object relations (marked as `obj`, `iobj`, `ccomp`, and `xcomp`). These are considered core arguments in a sentence, whose head-dependent distance is less subject to grammatical constraints than other syntactic relations.

Dependency length. We analyze two variants of measures for dependency length L. The first variant takes L as the orthographic dependency length, measured as the number of intervening orthographic words between the head and the dependent. The second variant takes L as the sum of surprisal of all the intervening words. Instead of assuming that every word contributes to memory interference to the same extent, this information-theoretic dependency length is supposed to better handle low-informative words, such as function words, which induce less memory burden compared to high-informative content words (Gibson, 1998; Grodner and Gibson, 2005).

3.2 Statistical Analyses

For the full dataset, we fit linear mixed-effect models (Baayen et al., 2008) for each language to separately predict the two variants of dependency length L introduced above as a function of antecedent surprisal, with random slope and intercept by dependency type. We also include three control variables: sentence position in the text (if the corpus is doc-by-doc), sentence length (word counts of a sentence), and the antecedent position in the sentence. In the analyses of the two subsets of data with subject or object relations only, we fit a linear model with the same critical variable and control variables as above. All the continuous variables are z-scaled.

4 Results

Figure 1 shows the average orthographic dependency length as a function of antecedent surprisal, along with linear fits. The visualization for the information-theoretic dependency length yields similar patterns (see additional figure in Appendix). Table 2 summarizes the statistical results of all the six versions of analyses (two measures of dependency length L crossed with three datasets). The sign indicates the direction of the antecedent surprisal effect on dependency length, with a plus sign suggesting a significant positive correlation between antecedent surprisal and dependency length.[1]

For the analysis on the full dataset, Danish, English, German, Italian, and Spanish show significant positive effect of antecedent surprisal with both measures of dependency length. The effect is significant for Amharic only with the dependency

[1]Analysis code is available at https://github.com/weijiexu-charlie/Dependency-length-strategic-memory-allocation

length as surprisal. A significant negative effect of antecedent surprisal is found for Korean. The effect for other languages does not reach significance.

For subject relations, we find evidence for a positive effect of antecedent surprisal on dependency length for 7 out of 11 languages, namely Amharic, Danish, English, German, Italian, Mandarin, Russian, and Spanish. There is no significant effect for Japanese and Korean. For German and Turkish, however, the data support a negative antecedent surprisal effect. For object relations, there is little evidence for a positive antecedent surprisal effect on the orthographic dependency length, with the effect being significantly negative for Amharic, German, Korean, Mandarin, and Russian. For the dependency length as surprisal, there is a positive antecedent surprisal effect in Danish, English, Italian, and Spanish. But the effect is negative for German and Korean.

5 Discussion

In general, our results provide evidence for a positive correlation between antecedent surprisal and dependency length, indicating that dependencies whose antecedent is more surprising and informative are able to tolerate longer dependency length. This is especially true for subject relations, where 7 out of 11 languages in our analysis exhibit a positive antecedent surprisal effect on dependency length. For object relations, however, the data presents a mixed picture without clear support for an expected antecedent surprisal effect.

This asymmetry between the subject and the object can be due to the possibility that object relations may be under more grammatical pressure than subject relations to put head and dependent closer to each other. For example, according to the Accessibility Hierarchy proposed by Keenan and Comrie (1977) as a linguistic universal, the subject is more relativizable than the object crosslinguistically to form a relative clause.

It is worth noting that the correlation observed in the current study is compatible with some other theories as well. For example, the Uniform Information Density (UID) theory holds that language production should avoid abrupt fluctuation of information across linguistic units (Jaeger and Levy, 2006; Meister et al., 2021). Therefore, surprising antecedents may be followed by longer sequence of units for a smoother transition to the other side of the dependency, which is supposed to bear lower surprisal due to the high mutual information with its antecedent (Futrell et al., 2019). However, compared to UID, which is a computational-level theory (Marr, 1982), the strategic memory allocation proposed in the current study focuses on the processes more at the mechanistic level.

6 Conclusions

In a nutshell, we find empirical support for a positive correlation between the length of a dependency and the surprisal of its antecedent. A closer look into the dependencies with core arguments shows that this relationship consistently holds for subject relations, but not for the object, possibility due to the stronger grammatical constraint between the object and the verb. At the population level, this finding indicates that although working memory constraints exert a general pressure on language structure to organize in a way that minimizes dependency length, this pressure is further modulated by informativity. This crosslinguistic pattern is consistent with our hypothesis of strategic working memory allocation as an individual-level processing strategy, where less predictable but more informative linguistic units are prioritized in working memory to maintain a more robust representation against memory interference and decay, thus are more tolerant for longer dependency length.

Limitations

The results of the current study are contingent upon the reliability of the corpora and the language model we use. For the use of corpora, our analyses may be vulnerable to the potential inaccuracies in corpus annotations, especially for dependencies whose identity is ambiguous and controversial. In terms of the use of language model, as one of the state-of-the-art LLMs, GPT-3 provides high-quality estimation of token surprisal, especially in languages with substantial sample size, such as English. However, the accuracy of surprisal estimates may be compromised when the model's training data is limited, diminishing the extensibility of our analysis to understudied languages. This limitation is particularly relevant for languages of potential interest from a typological perspective.

References

R Harald Baayen, Douglas J Davidson, and Douglas M Bates. 2008. Mixed-effects modeling with crossed

random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Brian Bartek, Richard L Lewis, Shravan Vasishth, and Mason R Smith. 2011. In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(5):1178.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Allison L Bruning and Jarrod A Lewis-Peacock. 2020. Long-term memory guides resource allocation in working memory. *Scientific Reports*, 10(1):1–10.

Jayeol Chun, Na-Rae Han, Jena D Hwang, and Jinho D Choi. 2018. Building Universal Dependency Treebanks in Korean. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data conversion and consistency of monolingual corpora: Russian UD treebanks. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, volume 155, pages 53–66.

Ramon Ferrer-i-Cancho. 2004. Euclidean distance between syntactically linked words. *Physical Review E*, 70(5):056135.

Richard Futrell, Roger P Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–412.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Richard Futrell, Peng Qian, Edward Gibson, Evelina Fedorenko, and Idan Blank. 2019. Syntactic dependencies correspond to word pairs with high mutual information. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 3–13.

Samuel J Gershman, Eric J Horvitz, and Joshua B Tenenbaum. 2015. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, language, brain: Papers from the first mind articulation project symposium*, pages 94–126. The MIT Press.

Edward Gibson, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, 29(2):261–290.

Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. 2022. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–86, Online. Association for Computational Linguistics.

John A Hawkins. 1994. *A performance theory of order and constituency*. 73. Cambridge University Press.

John A Hawkins. 2004. *Efficiency and complexity in grammars*. Oxford University Press.

John A Hawkins. 2014. *Cross-linguistic variation and efficiency*. Oxford University Press.

Philip Hofmeister. 2011. Representational complexity and memory retrieval in language comprehension. *Language and Cognitive Processes*, 26(3):376–405.

Philip Hofmeister and Shravan Vasishth. 2014. Distinctiveness and encoding effects in online sentence comprehension. *Frontiers in Psychology*, 5:1237.

Jacob Louis Hoover, Morgan Sonderegger, Steven T Piantadosi, and Timothy J O'Donnell. 2023. The plausibility of sampling as an algorithmic theory of sentence processing. *Open Mind*, 7:350–391.

T Florian Jaeger and Roger Levy. 2006. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19.

Anthony MV Jakob and Samuel J Gershman. 2023. Rate-distortion theory of neural coding and its implications for working memory. *Elife*, 12:e79450.

Anders Johannsen, Héctor Martínez Alonso, and Barbara Plank. 2015. Universal dependencies for Danish. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, page 157.

Hossein Karimi, Michele Diaz, and Fernanda Ferreira. 2019. "A cruel king" is not the same as "a king who is cruel": Modifier position affects how words are encoded and retrieved from memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(11):2010.

Edward L. Keenan and Bernard Comrie. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, 8(1):63–99.

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

Richard L Lewis, Andrew Howes, and Satinder Singh. 2014. Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, 6(2):279–311.

Richard L Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29:375–419.

Falk Lieder and Thomas L Griffiths. 2020. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA.

Büşra Marşan, Salih Furkan Akkurt, Muhammet Şen, Merve Gürbüz, Onur Güngör, Şaziye Betül Özateş, Suzan Üsküdarlı, Arzucan Özgür, Tunga Güngör, and Balkız Öztürk. 2022. Enhancements to the BOUN Treebank Reflecting the Agglutinative Nature of Turkish. *arXiv preprint arXiv:2207.11782*.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Binyam Ephrem Seyoum, Yusuke Miyao, and Baye Yimam Mekonnen. 2018. Universal Dependencies for Amharic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Chris R Sims. 2016. Rate–distortion theory and human perception. *Cognition*, 152:181–198.

Chris R Sims, Robert A Jacobs, and David C Knill. 2012. An ideal observer analysis of visual working memory. *Psychological Review*, 119(4):807.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto. 2016. Universal Dependencies for Japanese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1651–1658, Portorož, Slovenia. European Language Resources Association (ELRA).

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of 6th International Conference on Language Resources and Evaluation*, volume 2008, pages 96–101.

David Temperley and Daniel Gildea. 2018. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4:67–80.

Melissa Troyer, Philip Hofmeister, and Marta Kutas. 2016. Elaboration over a discourse facilitates retrieval in sentence processing. *Frontiers in Psychology*, 7:374.

Ronald Van den Berg and Wei Ji Ma. 2018. A resource-rational theory of set size effects in human visual working memory. *ELife*, 7:e34963.

Shravan Vasishth, Bruno Nicenboim, Felix Engelmann, and Frank Burchert. 2019. Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*, 23(11):968–982.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 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*, page 1707–1713.

Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721, Singapore. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

A Appendix

Additional figure:

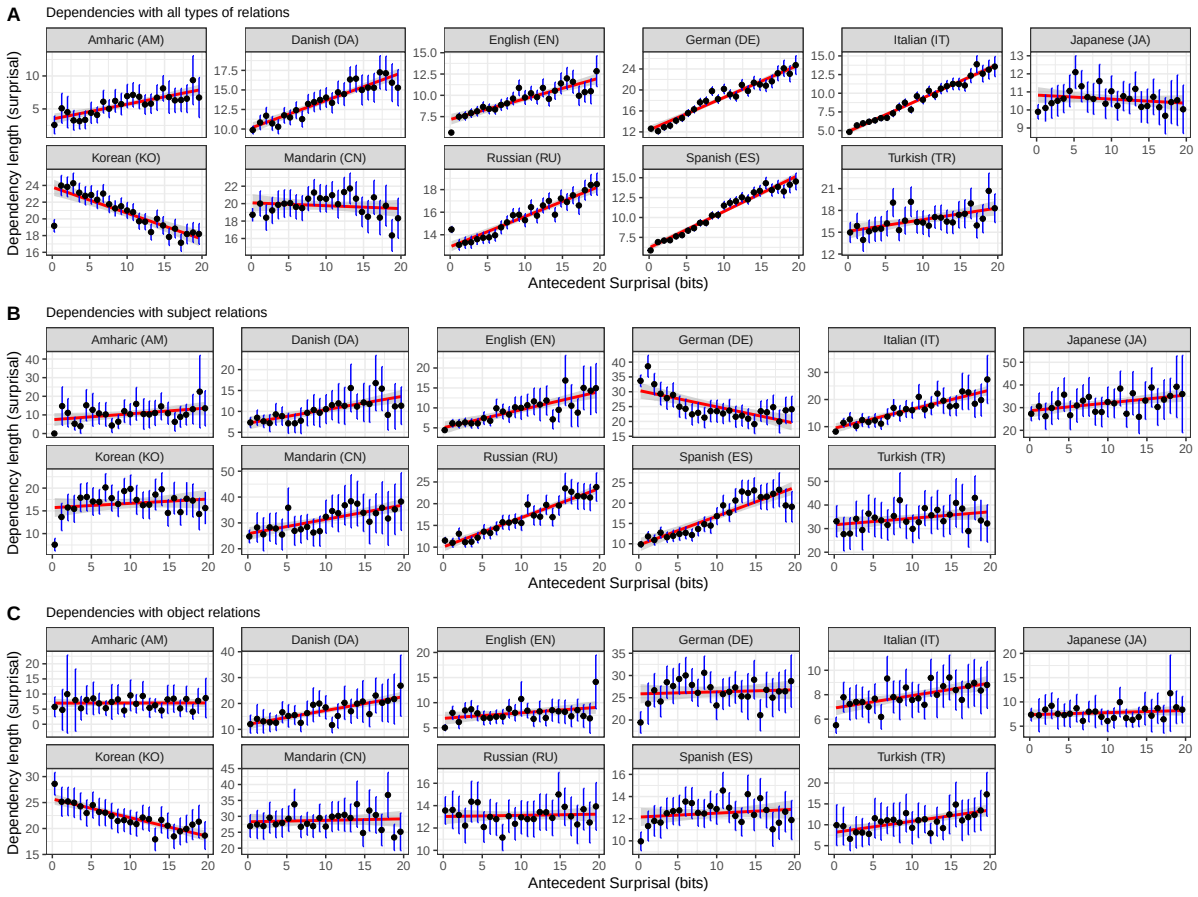

Figure 2: Average information-theoretic dependency length as a function of antecedent surprisal with subject relations. Surprisal is binned into 25 categories, and the mean dependency length within each category is shown in black with a 95% confidence interval in blue. A linear fit to these points is shown in red.