On the Distribution of Deep Clausal Embeddings: A Large Cross-linguistic Study

Damián E. Blasi[1,2] **Ryan Cotterell**[3] **Lawrence Wolf-Sonkin**[4]
Sabine Stoll[1] **Balthasar Bickel**[1] **Marco Baroni**[5,6]
1: University of Zürich **2:** Max Planck Institute for the Science of Human History
3: Cambridge University **4:** Johns Hopkins University
5: Facebook AI Research **6:** Catalan Institution for Research and Advanced Studies

Abstract

Embedding a clause inside another ("the girl [who likes cars [that run fast]] has arrived") is a fundamental resource that has been argued to be a key driver of linguistic expressiveness. As such, it plays a central role in fundamental debates on what makes human language unique, and how they might have evolved. Empirical evidence on the prevalence and the limits of embeddings has however been based on either laboratory setups or corpus data of relatively limited size. We introduce here a collection of large, dependency-parsed written corpora in 17 languages, that allow us, for the first time, to capture clausal embedding through dependency graphs and assess their distribution. Our results indicate that there is no evidence for hard constraints on embedding depth: the tail of depth distributions is heavy. Moreover, although deeply embedded clauses tend to be shorter, suggesting processing load issues, complex sentences with many embeddings do not display a bias towards less deep embeddings. Taken together, the results suggest that deep embeddings are not disfavored in written language. More generally, our study illustrates how resources and methods from latest-generation big-data NLP can provide new perspectives on fundamental questions in theoretical linguistics.

1 Introduction

In a prominent intellectual tradition, the infinitude of human expressivity (Humboldt, 1836) is rooted in a machinery that allows syntactic embedding at arbitrary depth (Chomsky, 1957, 1995). Regardless of the controversy around this proposal in terms of computational theory (Pullum and Scholz, 2010; Watamull et al., 2014), it remains an open issue to what extent languages in fact deploy structures with arbitrarily deep embedding. Many languages contain specific constructions that cap embedding depth at phrasal levels to one (e.g. unlike English, Modern Greek compounds don't allow recursive embedding; Ralli, 2013), although more radical constraints (Mithun, 1984; Everett, 2005; Evans and Levinson, 2009) seem to be rare and are avoided when languages evolve over time (Widmer et al., 2017). In terms of sentence production, embedding depths seem to be capped at moderate levels, likely because deeper embeddings place increasing demands on the brain's processing system (Gildea and Temperley, 2010).

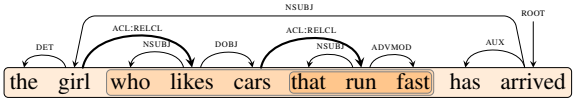

Figure 1: Example UD parse for sentence with maximum embedding $d = 2$.

Corpus studies of English, Pirahã, and a few other—mostly European—languages proposed constraints at one (Reich, 1969; Futrell et al., 2016), two (De Roeck et al., 1982), or three (Karlsson, 2010) levels of embedding. However, given that multiple embeddings might be vanishingly rare, a serious limitation of this work is the size of traditional corpora. If multiple embeddings are subject to constraints from processing load, these are likely to be probabilistic (rather than hard) in nature, and deeper embeddings are expected to be so rare that they are detectable only in very big data sets.

Here, we introduce a collection of large written corpora that we annotated using state-of-the-art parsers trained on Universal Dependencies (UD) treebanks (Nivre et al., 2018). We ask whether there are systematic patterns in the construction of complex nested clauses across languages. Instead of focusing on potential upper bounds of embedding depths we are interested in the distribution of syntactic dependencies in our corpora. We ask three questions: (i) How does embedding depth decline? (ii) Is the length of the clauses the same across levels of embedding? (iii) Can the rarity of deep embeddedness be explained by the rarity of

longer sentences, or is there a significant preference for simpler structures when sentence length is accounted for? The answer to these questions promises insights into the nature of constraints on the human parser, opening new research avenues on the computational complexity of human language.

2 Data

Corpora We focus on 17 languages, selected based on data and tool availability. We annotated 2 types of large, publicly available corpora: Wikipedia dumps from March 2017 and, where available, the WMT News Crawl corpora from 2007-2017 (Bei et al., 2018). Table 1 provides basic statistics of the annotated corpora.

Parsing Each corpus was tokenized using UDPipe's (Straka et al., 2016) pre-trained UDv2.2 models (Straka, 2018) and then parsed as follows: We trained Dozat et al. (2017)'s parsing model, a state-of-the-art graph-based neural dependency parser, on the Universal Dependencies 2.2 dataset (Nivre et al., 2018). We used the hyperparameter configuration described in Dozat et al. (2017), and pre-trained FastText word embeddings for frequent words (Bojanowski et al., 2016). We are aiming to make the parsed corpora available as soon as possible.

Measuring embedding depth We approximate embedding relations through dependency graphs. Specifically, for our purposes we define **embedding** as any dependency such that (i) the dependent is the head of a clause and (ii) permuting head and dependent would lead to an ungrammatical sentence, unlike in "flat" syntactic structures such as coordinated clauses. Any given clause has a depth d, defined as the number of embedded dependencies that need to be traversed in order to reach the root of the sentence from the target clause. For example the sentence in Figure 1 has a maximum embedding depth of 2, since the clause "that runs fast" is 2 ACL:RELCL-arcs from the root, and there exists no other clause in the sentence with a greater such distance.

We do not presently differentiate between center embedding and tail embedding. The difference is eventually important from a cognitive and computational perspective, but our current interest is focused on the overall distribution of embeddings in large corpora. Knowing this distribution is a prerequisite for modelling the impact of more specific distinctions (Bickel, 2010), such as center vs. tail embedding, or the position of the head (verb-final vs. verb-initial), or different types of clausal embedding (e.g. complement clauses vs. chaining etc.)

Language	Sentences	Tokens
Arabic	2.9	108.0
Bulgarian	2.7	54.8
Catalan	6.4	185.7
Danish	2.5	52.7
Dutch	10.7	207.7
French	86.4	2,283.2
German	214.0	4,159.8
Greek	2.5	59.0
Hebrew	5.1	140.6
Italian	6.9	203.9
Latvian	6.3	136.8
Portuguese	9.0	253.1
Romanian	4.3	114.5
Russian	102.7	1,924.6
Slovenian	1.4	31.8
Spanish	53.8	1,688.1
Swedish	18.4	261.1

Table 1: Sentence and token counts in millions in the annotated Wikipedia and news crawl corpora.

3 Results

3.1 Maximum Embeddedness Depth

As a first step we explore the tail of the distribution of maximum embeddedness depth in our corpora. We focus on the 1-20 range, for which a majority of the languages in our sample have coverage. The probability distributions are reported in Figure 2.

The distributions decay in a continuous fashion rather than finding an abrupt cutoff. An important aspect of characterizing the tail of distributions is whether they can be approximated by an exponential function $(\Pr(x) \sim \exp(-ax); a > 0)$ or whether they have a so-called "long-tail" parametrized by a power law $(\Pr(x) \sim x^{-a}; a > 0)$. One of the essential differences between these types is that long-tailed distributions display a *large number of rare events* (Khmaladze, 1988) (i.e., in our case, very deep embeddings are attested), in contrast to the exponential regime where the overwhelming majority of observations are bound within a comparatively narrower range. Statistically distinguishing between these types is not always straightforward and several alternative distributional models might yield comparable empirical performance (Clauset et al., 2009). It is possible however at least to compare heuristically the observed data against reference distributions of each type. For this purpose, we included in Figure 2

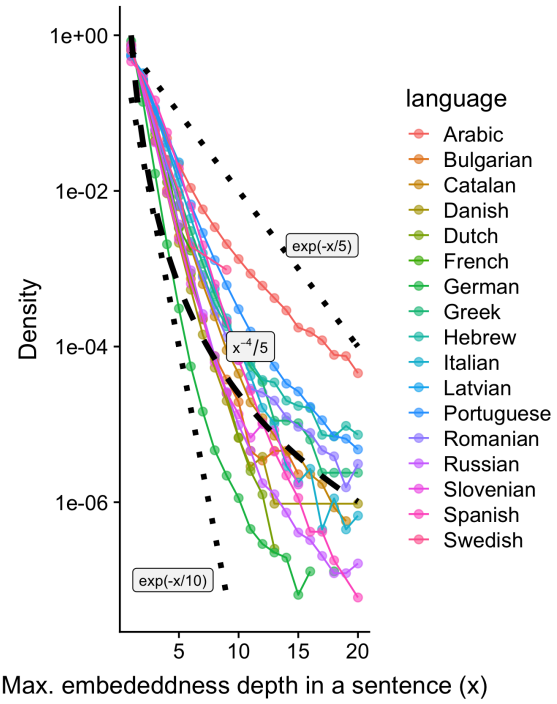

Figure 2: Distribution of maximum embeddedness depth in our corpora across languages. Dotted lines correspond to exponential distributions and the dashed line to a power-law distribution.

two exponential distributions flanking the empirical ones ($\exp(-x/5)$ and $\exp(-x/10)$), and a power-law distribution ($x^{-4/5}$). It can be observed that, while there is a relatively sharp and exponential decrease for the lowest values of embedding depth, the tail of the distributions become progressively less rapidly decaying, sometimes paralleling the behavior of the reference power-law distribution.[1]

3.2 Clause Depth and Length

As mentioned in the introduction, it is generally accepted that deeper levels of embeddedness imply a larger burden to the human parser. Given the ample evidence that linguistic behavior involves cost-avoiding strategies (Zipf, 1949), we expect that, all else being equal, clauses of larger d will be shorter, to minimize time spent in states with heavy processing demands. We model clause length (measured in number of orthographic words) in a Poisson regression model, with clause depth as independent variable. The results in Table 2 confirm across-the-board mean clause length reduction in function of depth.

[1]Visual inspection suggests that a small proportion of the deepest embeddings are found in degenerate text, e.g., misprocessed tables. Future work should estimate how such noise affects our statistics.

Language	Slope	SE (10^{-4})	Intercept
Arabic	−0.07	0.9	2.79
Bulgarian	−0.23	7	1.94
Catalan	−0.26	2	2.31
Danish	−0.23	6	1.93
Dutch	−0.32	3	2.03
French	−0.42	100	2.45
German	−0.26	1	1.90
Greek	−0.25	4	2.10
Hebrew	−0.25	2	2.31
Italian	−0.30	1	2.31
Latvian (n.s.)	−0.15	800	1.48
Portuguese	−0.16	1	2.02
Romanian	−0.28	3	2.32
Russian	−0.22	1	2.02
Slovenian	−0.27	9	1.98
Spanish	−0.31	0.7	2.38
Swedish	−0.27	100	1.85

Table 2: Estimates and standard errors (SE) of slopes for embedding depth and estimates of intercepts for Poisson regression model with clause length as dependent variable. Only the slope coefficient for Latvian is not significant ($\alpha = 0.01$).

3.3 Large Complex Sentences

Deep embeddings might be rare simply because complex, multi-clause sentences are rare in general. To assess this possibility, we test whether we can detect a bias against deep embedding *when taking sentence complexity (in counts of clauses) into consideration.*

For this, we introduce a minimal model for evaluating the presence of a bias against deep embeddings. We focus on complex matrix clauses with a large number of embedded clauses, as those are the ones where such a bias is most likely to be detectable if it exists. In practical terms, we evaluate main clauses with 8 or more total embedding dependencies only and at least one clause hosting two or more embedding dependencies. We consider the 14 corpora that contained at least 500 sentences satisfying these conditions.

We characterize these matrix clauses with their dependents as directed trees τ (with direction from parent to daughter nodes/clauses). Thus, a clause will be represented by a node with out-degree equal to the number of embedded dependencies hosted by the clause. The in-degree will be either 0 for main clauses and 1 for subordinate clauses. The matrix clause is then the root of such a tree, and the leaves are clauses which do not host any embedded clauses themselves.

To evaluate the observed distribution, we generate a baseline set of trees with no bias against deep embeddings. The baseline trees have (i) the

same number of clauses (n), and (ii) the same distribution of embedded dependencies (P(k)), i.e. the same out-degree distribution, as the observed trees.

Under the null hypothesis that there is no bias against deep embeddings, the distribution of observed tree depths should be compatible with the distribution of depths arising from the baseline set of trees.

As an ilustration consider the sentence "[The girl [who likes cars [that run fast]] arrived [as I cooked the pasta [that you gave me]]] ". In the Universal Dependencies convention, denoting each clause by its own head, this can be represented by the tree in Figure 3(a). This sentence has an embedding depth of 2, it has five clauses (nodes) in total, and three clauses have non-zero out-degree. One possible alternative tree with the same number of nodes and the same out-degree distribution is given in Figure 3(b), which has an embedding depth of 3. Hence, the same number of clauses and embedded dependencies distribution yields a tree that is *deeper*. In the case of large complex sentences with many clauses, there exist multiple such trees that could be leveraged to determine whether the depth of the empirically observed sentence is unusually low or high given what is expected under the baseline.

It should be stressed that this scheme of comparison considers each observed sentence independently: the statistics of other sentences in the same language play no role. In order to evaluate the overall bias in a language, we compare the difference between the observed depth of each sentence against the mean value of 100 permuted baseline trees derived from them, and we aggregate the results of all sentences within a language. If the resulting distribution of depth differences has its probability mass systematically above or below zero, this would speak against the null hypothesis of no bias.

Surprisingly, we find no outstanding systematic pattern in the comparison. While the median and mean values of the differences vary across languages, the distributions hover around zero with a modest variation (so that in general we observe that the empirical values are an average of 1 SD from the reference sample mean). Figure 4 displays the cumulative distribution of the difference between empirical and mean reference embeddedness depth across languages.

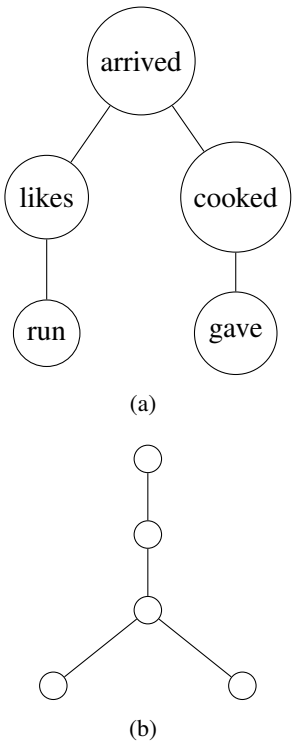

Figure 3: (a) UD-style clause dependencies for the sentence "[The girl [who likes cars [that run fast]] arrived [as I cooked the pasta [that you gave me]]]"; (b) Example alternative tree with same number of nodes and out-degree distribution.

4 Conclusions

We empirically addressed one central issue in theoretical linguistics, namely the nature and distribution of nested clausal embeddings in natural languages. Large corpora and automated annotation tools are crucial to address this question, as deep embeddings are expectedly rare. Our results confirm that there is no sharp boundary on maximum embedding depth. More deeply embedded sentences appear to be shorter (in number of words), and this is in accordance with the hypothesis that they impose a heavier processing load than shallower clauses. However, surprisingly, when sentence complexity (in number of clauses) is accounted for, there is no clear bias against deeper embeddings.

This is a first large quantitative exploration of the matter. In future work, we will extend our set of languages, aiming at more typological variety (Indo-European languages are greatly over-represented in our current data). Moreover, our results rely on automated annotation, and we have no estimate of the extent to which they are affected by annotation error. Finally, we have glossed over potential dif-

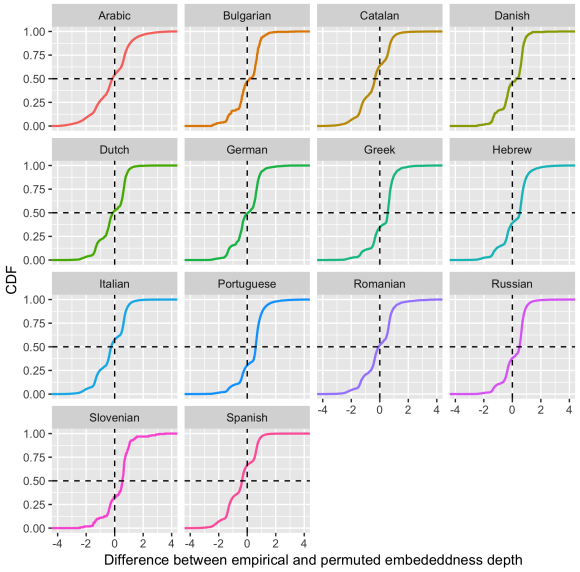

Figure 4: Cumulative distribution function of the difference in embeddedness depth for 14 languages with at least 500 sentences with more than 8 clauses.

ferences in embedding preferences stemming from differences in types of dependencies (e.g. center vs. tail embedding) and their linearizations (e.g. head-initial vs. head-final), although these differences are likely to play an important role.

References

Chao Bei, Hao Zong, Yiming Wang, Baoyong Fan, Shiqi Li, and Conghu Yuan. 2018. An empirical study of machine translation for the shared task of WMT18. In *Proceedings of the Third Conference on Machine Translation*, pages 344–348, Belgium, Brussels. Association for Computational Linguistics.

Balthasar Bickel. 2010. Capturing particulars and universals in clause linkage: a multivariate analysis. In Isabelle Bril, editor, *Clause-hierarchy and clause-linking: the syntax and pragmatics interface*, pages 51–101. Benjamins, Amsterdam.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton, Berlin, Germany.

Noam Chomsky. 1995. *The Minimalist Program*. MIT Press, Cambridge, MA.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Anne De Roeck, Roderick Johnson, Margaret King, Michael Rosner, Geoffrey Sampson, and Nino Varile. 1982. A myth about centre-embedding. *Lingua*, 58(3-4):327–340.

Timothy Dozat, Peng Qi, and Christopher D Manning. 2017. Stanford's graph-based neural dependency parser at the conll 2017 shared task. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30.

Nicholas Evans and Stephen C. Levinson. 2009. The myth of language universals: language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32:429–448.

Daniel Everett. 2005. Cultural constraints on grammar and cognition in pirahã: Another look at the design features of human language. *Current anthropology*, 46(4):621–646.

Richard Futrell, Laura Stearns, Daniel L. Everett, Steven T. Piantadosi, and Edward Gibson. 2016. A corpus investigation of syntactic embedding in Pirahã. *PLoS ONE*, 11:e0145289.

Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.

Wilhelm von Humboldt. 1836. *Über die Verschiedenheit des menschlichen Sprachbaus und ihren Einfluss auf die geistige Entwickelung des Menschengeschlechtes*. Dümmler, Berlin.

Fred Karlsson. 2010. Syntactic recursion and iteration. *Recursion and Human Language*, pages 43–67.

Estate V Khmaladze. 1988. The statistical analysis of a large number of rare events. *Department of Mathematical Statistics*, (R 8804).

Marianne Mithun. 1984. How to avoid subordination. *Proceedings of the 10th Annual Meeting of the Berkeley Linguistics Society*, pages 493–509.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Sandra Bellato, Kepa Bengoetxea, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Rogier Blokland, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat,

Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roca, Olga Rudina, Shoval Sadde, Shadi Saleh, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Yuta Takahashi, Takaaki Tanaka, Isabelle Tellier, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jonathan North Washington, Seyi Williams, Mats Wirén, Tsegay Woldemariam, Tak-sum Wong, Chunxiao Yan, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 2018. Universal dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Geoffrey K. Pullum and Barbara C. Scholz. 2010. Recursion and the infinitude claim. In Harry van der Hulst, editor, *Recursion and human language*, pages 113–137. De Gruyter Mouton, Berlin.

Angela Ralli. 2013. *Compounding in modern Greek*. Springer, Berlin.

Peter A. Reich. 1969. The finiteness of natural language. *Language*, pages 831–843.

Milan Straka. 2018. CoNLL 2018 shared task - UDPipe baseline models and supplementary materials. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Milan Straka, Jan Hajic, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *LREC*.

Jeffrey Watamull, Marc D. Hauser, Ian G. Roberts, and Norbert Hornstein. 2014. On recursion. *Frontiers in Psychology*, 4:doi:10.3389/fpsyg.2013.01017.

Manuel Widmer, Sandra Auderset, Paul Widmer, Johanna Nichols, and Balthasar Bickel. 2017. NP recursion over time: evidence from Indo-European. *Language*, 93:1–36.

GK Zipf. 1949. Human behaviour and the principle of least-effort.