The Syntactic Acceptability Dataset (Preview): A Resource for Machine Learning and Linguistic Analysis

Anonymous submission

Abstract

We present a preview of the Syntactic Acceptability Dataset, a resource being designed for both syntax and computational linguistics research. In its current form, the dataset comprises 1,000 sequences from the syntactic discourse: Half from textbooks and half from the journal *Linguistic Inquiry*, the latter to ensure a representation of the contemporary discourse. Each entry is labeled with its grammatical status (“well-formedness” according to syntactic formalisms) extracted from the literature, as well as its acceptability status (“intuitive goodness” as determined by native speakers) obtained through crowdsourcing, with highest experimental standards. Even in its preliminary form, this dataset stands as the largest of its kind that is publicly accessible. We also offer preliminary analyses addressing three debates in linguistics and computational linguistics: We observe that grammaticality and acceptability judgments converge in about 83% of the cases and that “in-betweenness” occurs frequently. This corroborates existing research. We also find that while machine learning models struggle with predicting grammaticality, they perform considerably better in predicting acceptability. This is a novel finding. Future work will focus on expanding the dataset.

Keywords: computational linguistics, grammaticality, acceptability, gradience, data convergence

1. Introduction

One of the primary goals of syntactic theory is to identify the principles and processes that dictate the structure of sequences in a particular language and in human language in general. Syntax is primarily concerned with describing, explaining, and predicting the grammatical status of these sequences, particularly distinguishing between grammatical and ungrammatical sequences. Chomsky refers to these as members of the sets G and G’, respectively (Chomsky, 1975).

To build formalisms, syntacticians rely on various kinds of data, with a significant emphasis on their own expert judgments, often referred to as *grammaticality judgments* (Schütze, 1996). These judgments are obtained when experts carefully examine and contrast linguistic sequences, determining if a given sequence aligns with their grammatical formalisms. During this evaluation, linguists abstract away from extra-grammatical factors, such as memory limitations. This is illustrated in Sequence 1, taken from Chomsky et al. (1963).

(1) The rat the cat the dog chased killed ate the malt.

In recent years, sequences and their grammaticality evaluations have become increasingly accessible. The largest source of such data to date is the Corpus of Linguistic Acceptability (CoLA; Warstadt et al., 2019), which contains more than 10,000 sequences and their respective grammatical statuses (see Section 2 for why CoLA contains grammaticality judgments instead of acceptability judgments, as per standard usage in linguistics). The sequences in CoLA are sourced from syntax textbooks, and their grammaticality evaluations are provided by the authors of these textbooks.

1.1. Issues surrounding Grammaticality

Several issues arise when discussing grammaticality judgments. First, there is the matter of data adequacy and convergence. Sequence 2, taken from Landau (2007), is labeled as ungrammatical by the original author from whom the sequence was sourced. However, most laypeople consider the sequence to be acceptable (Francis, 2021, p. 207).

(2) *October 1st, he came back.

Furthermore, there is the question of *gradience*. Is grammaticality a binary concept, or do degrees of (un)grammaticality exist (Chomsky, 1975; Wasow, 2007; Francis, 2021)? Considering a sequence such as Sequence 3, taken from Francis (2021, p. 38), it becomes challenging to pinpoint which factors outside the traditional grammar influence the perception of the sequence as neither fully grammatical nor fully ungrammatical.

(3) Olson brings to the table a great deal of experience.

Thirdly, it has been observed that machine learning models struggle with the notion of grammaticality. Warstadt et al. (2019) trained various LSTM models on CoLA and observed accuracy results well below 80%. Given that contemporary models demonstrate high proficiency in various linguistic tasks (Devlin et al., 2018), the machine learning of grammaticality is of particular interest.

2. Acceptability

There are, however, many methods at the disposal of syntacticians to assist them in theory-building. These methods include self-paced reading tasks,

Gra	Acc	Nm-ac	Bi-ac	Src	Sequence
0	1.35	0.06	0	jl	John is too much to play with your kids old.
...	...	...	...	...	...
0	3.90	0.48	0	tb	I assumed to be innocent.
...	...	...	...	...	...
1	6.89	0.99	1	tb	I saw John on Sunday.

Table 1: The structure of the our dataset, including labels for grammaticality, acceptability, normalized acceptability, binary acceptability, source (textbook or journal), and the sequence.

EEG measurements, eye-tracking, and eliciting acceptability judgments. As to the latter, the linguistic literature defines acceptability judgments as non-expert, 'naive' intuitions about the goodness of a sequence (see discussion in Häussler and Juzek, 2020, pp.235-236). Acceptability judgments can be influenced by extra-grammatical factors (Schütze, 2020), which contrasts to grammaticality judgments, where experts abstract away from extra-grammatical factors as much as possibl (Schütze, 1996). When carefully controlled and analyzed, acceptability judgments can serve as a proxy for grammaticality (Schütze, 2020). To illustrate the distinction between the two concepts, reconsider the examples from the previous section. Sequence 1 is grammatical according to most syntactic frameworks, yet many native speakers find it unacceptable. Sequence 2 is ungrammatical according to most frameworks, yet many native speakers find it acceptable.

While acceptability data is instructive, its collection is also resource-intensive. As a result, there are no large-scale datasets publicly available. To our knowledge, the largest datasets available are those by Lau et al. (2017) with 400 items, and Warstadt et al. (2019) with 200 items. The primary objective of this study is to produce and offer a publicly available dataset on a large(r) scale. We present initial acceptability judgments for 1,000 items, with an eventual goal of scaling this to approximately 15,000 items. The dataset encompasses sequences, grammaticality judgments, acceptability judgments (raw, normalized, and converted), and encodes its source (textbook vs journal). Even this preliminary dataset addresses the three issues mentioned earlier: data convergence, gradience, and challenges in machine learning. We will delve deeper into these three topics in Section 4.

3. Corpus Building

Our data is taken from two sources, representing two conditions. The first condition, referred to as the 'textbook condition', consists of 500 sequences randomly sampled from CoLA, which itself is sourced from various syntax textbooks. The second condition, the 'journal condition', comprises 500 sequences randomly sampled from the data from Juzek and Häussler (2020), who in turn sampled their items from the journal *Linguistic Inquiry*. It is anticipated that the sequences from textbooks are more foundational and well-established. Importantly, both sets of items are accompanied by grammaticality evaluations, as provided by their original sources. Examples of these sequences can be found in Table 1.

Of the sequences from textbooks, 71.4% were grammatical, compared to 67.4% from the journal. Consequently, our dataset exhibits an imbalance. We opted to sample singletons rather than minimal pairs, primarily because many items in the literature are presented without counterparts. For detailed discussions on this choice, refer to discussions in Warstadt et al. (2019) and Juzek and Häussler (2020).

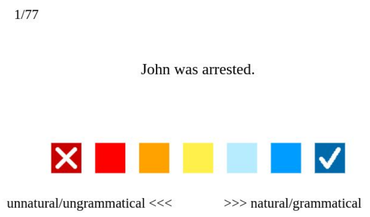

Figure 1: The interface of the judgment study.

3.1. Obtaining Acceptability Judgments

Acceptability judgments were obtained through a self-hosted rating platform using the interface shown in Figure 1. Participants were crowdsourced via Prolific.com. A prerequisite for participation was that participants on Prolific had set their first language as American English. On average, participants were paid $15/hr. After participants were provided with IRB information and instructions, they rated 77 items, 64 of which were critical items. We limited the number of items to avoid experimental fatigue. The first four items served as calibration items, taken from previous experiments to represent near-endpoints: two were unacceptable and two were acceptable. Participants then rated the re-

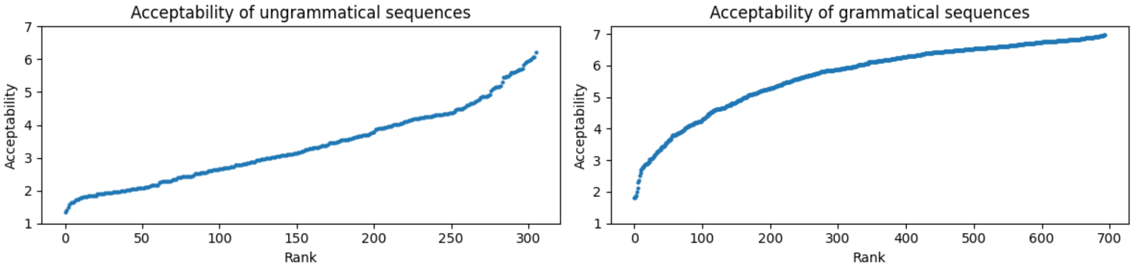

Figure 2: The items and their average acceptability ratings, sorted in ascending order, from unacceptable to acceptable. Left: Items evaluated as ungrammatical in the original source. Right: Items evaluated as grammatical.

maining items. We opted not to include filler items since no distractor items were needed for our study, and the critical set covers all parts of the scale.

For each participant, the 64 critical items were semi-randomly selected from the 1,000 items in our corpus, prioritizing items that had received the fewest ratings thus far. We interspersed four items testing language proficiency and five items testing general attention (of the sort "Please click on the leftmost button"). The platform also measured response times. If participants responded unrealistically fast, they received a warning. Those who were repeatedly non-cooperative were excluded immediately. After successfully completing the tasks, we collected basic demographics: gender, age, and first language. A total of 597 participants took part. We excluded users for the following reasons: less than 69 ratings were given (16 participants), unrealistically fast responses (2 participants), failing on language proficiency items (36 participants), failing on instructional items (2 participants), identifying as non-native speakers (2 participants). This commitment to quality is evident as the item with the lowest average rating had a score of 1.35 (an instructional item even averaged 1.03), while the highest-rated item had an average rating of 6.98. This indicates that participants utilized the entire rating scale. In total, the corpus consists of 34,490 acceptability judgments, making it the most extensive publicly available dataset of its kind.

3.2. The Corpus

The structure of the corpus is detailed in Table 1. The corpus includes average acceptability ratings given on a 7-point scale. These ratings were then normalized to values between 0 and 1. For the purpose of binary classification, ratings between 0 and 0.5 were converted to 0, while ratings between 0.5 and 1 were converted to 1. In cases where ratings were exactly 0.5, we used the respective grammaticality value to determine the binary label. Figures 2 and 3 illustrate the data distribution and structure, both of which will be discussed in the following section.

4. Preliminary Analyses

4.1. Data Convergence

83.3% of all items share their grammaticality label and (to binary form converted) acceptability label. This rate is slightly higher in the textbook condition, at 85.8%, and lower in the journal condition, at 80.8%. These figures align with previous results (for an analysis of paired items, see Sprouse et al., 2013). Moreover, the convergence rate for grammatical items (89.3%) is considerably higher than for ungrammatical ones (69.6%). This discrepancy is a novel finding and requires further investigation through a detailed item-by-item analysis: Apparently, there are numerous items that syntacticians label as ill-formed based on their formalisms, but which laypeople deem relatively acceptable. Sequence 4 serves as an example of this discrepancy. It was evaluated as ungrammatical in its original source, but received an average rating of 6.22 in our experiment. Moreover, the observed divergence rate of approximately 20% underscores the idea that grammaticality and acceptability are indeed two distinct concepts.

(4) *John perfectly rolled the ball. (6.22)

4.2. Gradience

As consistently observed in the literature, acceptability exhibits a gradient nature (e.g. Featherston, 2005; Wasow, 2007; Francis, 2021). The degree of this gradience is more pronounced than one might initially anticipate. In our results, when all items are ordered by rank in an ascending manner, as per Figure 3, and when the initial few items with the lowest ratings are disregarded, there is a near-linear increase in acceptability. Interestingly, towards the higher end of the rating scale, the curve begins to resemble a saturation curve. This is in contrast to the sharper S-curve that one might expect. When the rating scale is divided into thirds, 28.3% of all items are found in the middle bin. When the scale is divided into two bins: endpoint items (with ratings from 1 to 2.5 and 5.5 to 7) and in-between items (with ratings larger than 2.5 and smaller than 5.5), 42.6% of all items are are found in the middle bin. This distribution is illustrated in Figure 3. Our findings align with the discussion in Francis (2021).

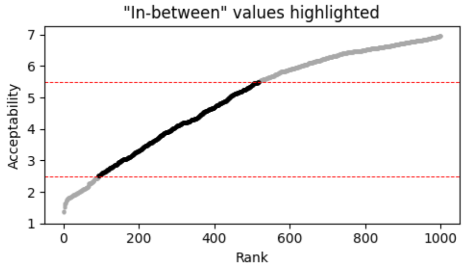

Figure 3: All items, with the mid-bin highlighted.

4.3. Challenges in Machine Learning

Transformers (Vaswani et al., 2017) demonstrate remarkable language abilities (Devlin et al., 2018). Here, we ask how machine learning of grammaticality compares to machine learning of acceptability. Our dataset is relatively small for machine learning and can thus be viewed as a scarce-data learning scenario (Wang et al., 2020). We fine-tuned Transformers on our data, with pre-trained models from Wolf et al. (2020) ("bert-base-uncased"), for four conditions: 1) predicting grammaticality, 2) predicting acceptability, 3) predicting end-point acceptability (which exclude "in-between" items as per Figure 3). Additionally, we include 4) a baseline condition where we sampled sentences from the Leipzig Corpora Collection for English (Goldhahn et al., 2012) (labelled 'good') and scrambled the word order of 500 sentences (labelled 'bad'), then fine-tuned a Transformer to make predictions on these. Linear confusion matrices for these conditions are presented in Table 2.

As expected, the baseline model performs well. We observe that the models struggle with grammaticality, but they perform better on acceptability items. Furthermore, their performance on end-point acceptability is considerably better. These findings regarding acceptability are novel and warrant further research. Further, these findings also motivate the distinction between grammaticality and acceptability.

Condition	tn	fp	fn	tp	Accu
Grammatic.	0	45	0	105	70%
Acceptab.	23	21	5	101	83%
End-p. acc.	9	3	2	73	94%
Baseline	39	0	4	107	97%

Table 2: Linear confusion matrices for transformers, fine-tuned on our data in the different conditions, as per Section 4.3. Test data is 15% of the corpus.

5. Next Steps

While 1,000 items are a good start, the ideal situation for syntactic theory building would be that syntacticians can look up the acceptability of all relevant items. For this, we wish to expand our dataset to all items in Warstadt et al. (2019) and Juzek and Häussler (2020), resulting in a dataset of about 15,000 items. This would also help solidify our insights regarding the machine learning of acceptability. Further, ideally, syntacticians would not only be able to look up items but also search for syntactic constructions. This would require expert annotations for the items in the corpus. We are currently exploring possibilities to efficiently add such annotations. Thirdly, while we used response times for exclusions, we still need to do further analyses on the collected response times. For example, it could be interesting to see if there is a correlation between unacceptability and increased response times.

6. Concluding Remarks

We have introduced a preview of the Syntactic Acceptability Dataset, which comprises 1,000 sentences sourced from syntactic textbooks and the journal *Linguistic Inquiry*. Each item in the dataset is accompanied by grammaticality evaluations and high-quality acceptability ratings. Even in its current form, this dataset is considerably larger than any other acceptability dataset currently available, and it has already provided insights into several debates. The dataset aids in understanding issues related to data convergence (with grammaticality and acceptability converging in about 83% of cases, and a higher rate for textbook sequences), gradience (items with intermediate ratings are common), and machine learning challenges (grammaticality proves more difficult to predict than acceptability). In the next phase, we aim to expand the dataset and further validate our preliminary findings.

7. Data Availability

The corpus and all relevant scripts are on Github (anonymously): github.com/arizus/sad.

8. Bibliographical References

N. Chomsky. 1975. *The Logical Structure of Linguistic Theory*. Springer US.

Noam Chomsky, George Armitage Miller, R Luce, R Bush, and E Galanter. 1963. Introduction to the formal analysis of natural languages. *1963*, pages 269–321.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sam Featherston. 2005. The decathlon model of empirical syntax in: Reis, marga, kepser, stephan stephan (eds.), linguistic evidence. empirical, theoretical and computational perspectives.

Elaine Francis. 2021. *Gradient acceptability and linguistic theory*, volume 11. Oxford University Press.

Dirk Goldhahn, Thomas Eckart, Uwe Quasthoff, et al. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43.

Jana Häussler and Tom Juzek. 2020. Linguistic intuitions and the puzzle of gradience. *Linguistic intuitions: Evidence and method*, pages 233–254.

Tom S Juzek and Jana Häussler. 2020. Data convergence in syntactic theory and the role of sentence pairs. *Zeitschrift für Sprachwissenschaft*, 39(2):109–147.

Idan Landau. 2007. Epp extensions. *Linguistic Inquiry*, 38(3):485–523.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

Carson T. Schütze. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press, Chicago, IL.

Carson T. Schütze. 2020. Acceptability ratings cannot be taken at face value. In Samuel Schindler, Anna Drożdżowicz, and Karen Brøcker, editors, *Linguistic Intuitions: Evidence and Method*, chapter 11, pages 189–214. Oxford University Press, Oxford.

Jon Sprouse, Carson T Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, 134:219–248.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Thomas Wasow. 2007. Gradient data and gradient grammars. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, volume 43, pages 255–271. Chicago Linguistic Society.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.