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

Tom S Juzek

Florida State University
Tallahassee, FL, USA
tjuzek@fsu.edu

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 English 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). Crucially, syntacticians focus on linguistic competence, which is a speaker's often implicit knowledge of a language (Chomsky, 1965).

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; Francis, 2021). These judgments are obtained when experts carefully examine and contrast linguistic sequences, determining whether 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 (see the Appendix for examples).

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.

16113

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.

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; Tenney et al., 2019), 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, 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, as well as various contributions in Schindler et al., 2020, and see Etxeberria et al., 2018 and Schoenmakers, 2023 for further nuances). Acceptability judgments can be influenced by extra-grammatical factors (Schütze, 2020; Łęska-Bayraktar and Żychliński, 2023), which contrasts to grammaticality judgments, where experts abstract away from extra-grammatical factors as much as possible (Schütze, 1996). When carefully controlled and analyzed, acceptability judgments can serve as a proxy for grammaticality (Schütze, 2020; Feldhausen and Buchczyk, 2021). 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.

Acceptability data are reliable (Langsford et al., 2018) and instructive for linguistic purposes, and an increasing number of studies make use of them, where recent examples include Fanselow et al. (2019), Hoot and Ebert (2021), Urtzi et al. (2022), and Lami and van de Weijer (2022). However, the collection of acceptability data 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. Others, for example Sprouse et al. (2013) for English, and Chen et al. (2020) for Mandarin, report on similar mid-sized datasets; these are, however, not publicly available. 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. Dataset Building

Our data are taken from two sources, representing two conditions. The first condition, referred to as the 'textbook condition', consists of 500 English sequences randomly sampled from CoLA, which itself is sourced from various syntax textbooks. The second condition, the 'journal condition', comprises 500 English sequences randomly sampled from the data from Juzek and Häussler (2020), who in turn sampled their items from the journal *Linguistic Inquiry*. In terms of their structure, items from both conditions are similar: a sequence relevant to syntax is presented alongside a grammaticality judgment. However, a *possible* difference lies in their grammatical status. 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. Examples of the data from which we sampled, that is examples from both CoLA and the data in Juzek and Häussler (2020), can be found in the Appendix.

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).

1/77

John was arrested.

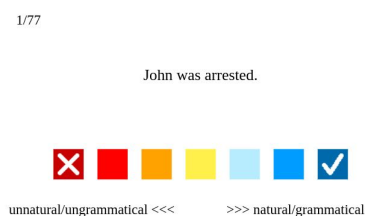unnatural/ungrammatical <<<   >>> natural/grammatical

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 remaining 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 dataset, 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. While under certain circumstances, fewer than ten items have been shown to give robust results (Mahowald et al., 2016), we adhered to a more conservative N, with an average of over 30 ratings per item. In total, the dataset consists of 34,490 acceptability judgments, making it the most extensive publicly available dataset of its kind.

3.2. The Dataset

The structure of the dataset is detailed in Table 1. The dataset 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 discussions in Wasow and Arnold (2005) and Gibson et al. (2013), and with previous results in Warstadt et al. (2019) and Juzek and Häussler (2020). For an analysis of paired items, with a higher convergence rate, 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.

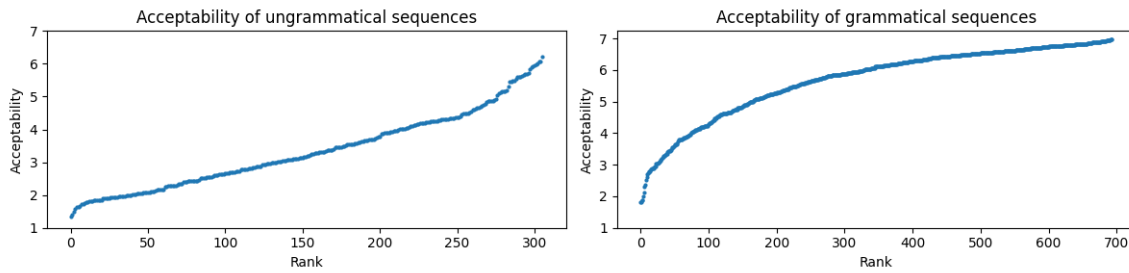(4)   *John perfectly rolled the ball.   (6.22)

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.

4.2. Gradience

As consistently observed in the literature, acceptability exhibits a gradient nature (e.g. Featherston, 2005; Wasow, 2007; Häussler and Juzek, 2020). 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).
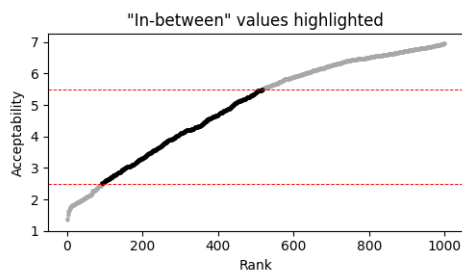

Figure 3: All items, with the mid-bin highlighted.

4.3. Challenges in Machine Learning

Transformers (Vaswani et al., 2017) demonstrate remarkable language abilities (Devlin et al., 2018). Hewitt and Manning (2019) provided documentation of explicit syntactic representations. Conversely, however, Chaves (2020), Chaves and Richter (2021), and others, have delineated the challenges that deep learning encounters with syntactic constructions. 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 excludes "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, which may suggest that syntactic (un)acceptability cannot be solely predicted by word order. 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, but also limited in scope, and thus, warrant further, more in-depth 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 dataset.

5. Next Steps

5.1. Scaling

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.

5.2. Further annotations

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 dataset. We are currently exploring possibilities to efficiently add such annotations. A scaled corpus with annotations for constructions could help with theory building and could inform research lines such as the work by Bader and Häussler (2010) and Bizzoni et al. (2020).

Prosodic information could also be of interest for syntactic analysis. Recent research is underlining the relevance of prosody-syntax interactions, for work on such effects see e.g. Wasow et al. (2015), Tang and Shaw (2021), González et al. (2022), and González and Reglero (2024). Information-theoretic measures like surprisal and perplexity (Shannon, 1948) and advanced measures like lossy surprisal (Futrell et al., 2020) could allow for advanced syntactic analyses. We also wish to add dependency parses (Tesnière, 2015; Nivre, 2008) to further facilitate dependency-based syntactic research along the lines of Gibson et al. (2000), Liu (2010).

5.3. Further analyses

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. Further, the setup allows for analyses such as the F_1-score (Chinchor, 1992; Van Rijsbergen, 1979) or the MCC (Chicco and Jurman, 2020). An alternative line of analysis could involve a detailed examination of the 'micro'-factors that underpin the observed 'macro'-trends. This could be done by carefully analyzing a selection of items from the dataset. Krielke (2024) offers an exemplary analysis in this regard, which could serve as a model for our further investigations.

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, by adding more items, additional annotations, more advanced analyses, and further validate our preliminary findings.

Data Availability

The dataset and all relevant scripts are on Github: github.com/tjuzek/sad.

Appendix

In the following, we give an illustration of textbook items and journal items, to demonstrate that in structure, the sequences are similar. The items are used to analyze syntactic structures and the sequences come with an evaluation of their grammaticality (indicated by the asterisk or the lack thereof).

Sequences 5 and 6 are ungrammatical items taken from textbooks (from Kim and Sells, 2008 and Baltin and Collins, 1991, respectively; we sampled them from CoLA), Sequences 7 and 8 come from the journal *Linguistic Inquiry* (both items come from Grosu and Horvath, 2006; we sampled them through Juzek and Häussler, 2020). Similarly, Sequences 9 and 10 are grammatical sequences from linguistics textbooks (from Adger, 2003 and Sportiche et al., 2013, respectively) and Sequences 11 and 12 are grammatical sequences from *Linguistic Inquiry* (both items come from Basilico, 2003).

(5) *You didn't leave, left you?

(6) *John seems will win.

(7) *John is too much to play with your kids old.

(8) *John is a more unusually than any of you is dressed student.

(9) I might be leaving soon.

(10) I saw John on Sunday.

(11) I really hate you right now.

(12) The guard made the prisoner unhappy.

Bibliographical References

David Adger. 2003. *Core syntax: A minimalist approach*. Oxford University Press.

Markus Bader and Jana Häussler. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics*, 46(2):273–330.

Mark Baltin and Chris Collins. 1991. *The handbook of contemporary syntactic theory*. John Wiley & Sons.

David Basilico. 2003. The topic of small clauses. *Linguistic Inquiry*, 34(1):1–35.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International conference on spoken language translation*, pages 280–290.

Rui P Chaves. 2020. What don't rnn language models learn about filler-gap dependencies? *Proceedings of the Society for Computation in Linguistics*, 3(1):20–30.

Rui P Chaves and Stephanie N Richter. 2021. Look at that! bert can be easily distracted from paying attention to morphosyntax. *Proceedings of the Society for Computation in Linguistics*, 4(1):28–38.

Zhong Chen, Yuhang Xu, and Zhiguo Xie. 2020. Assessing introspective linguistic judgments quantitatively: the case of the syntax of chinese. *Journal of East Asian Linguistics*, 29:311–336.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

Nancy Chinchor. 1992. Muc-4 evaluation metrics. In *Proceedings of the 4th conference on Message understanding-MUC4'92*. Association for Computational Linguistics.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge (MA).

Noam 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*.

Urtzi Etxeberria, Susagna Tubau, Viviane Déprez, Joan Borràs-Comes, and M Teresa Espinal. 2018. Relating (un) acceptability to interpretation. experimental investigations on negation. *Frontiers in psychology*, 8:274247.

Gisbert Fanselow, Jana Häussler, and Thomas Weskott. 2019. Who cares what who prefers? a study in judgment differences between syntacticians and non-syntacticians. *Grammatical approaches to language processing: Essays in honor of Lyn Frazier*, pages 261–274.

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

Ingo Feldhausen and Sebastian Buchczyk. 2021. Revisiting subjunctive obviation in french: a formal acceptability judgment study. *Glossa: a journal of general linguistics. 2021; 6 (1): 59*.

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

Richard Futrell, Edward Gibson, and Roger P Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive science*, 44(3):e12814.

Edward Gibson, Steven T Piantadosi, and Evelina Fedorenko. 2013. Quantitative methods in syntax/semantics research: A response to sprouse and almeida (2013). *Language and Cognitive processes*, 28(3):229–240.

Edward Gibson et al. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000:95–126.

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.

Carolina González and Lara Reglero. 2024. Intonation correlates of canonical and non-canonical wh-in-situ questions in spanish. *Journal of Linguistics*, 60(1):29–50.

Carolina González, Lara Reglero, Barbara E Bullock, Cinzia Russi, and Almeida Jacqueline Toribio. 2022. Prosodic correlates of mirative and new information focus in spanish wh-in-situ questions. *A half century of Romance linguistics*, page 269.

Alexander Grosu and Julia Horvath. 2006. Reply to bhatt and pancheva's "late merger of degree clauses": The irrelevance of (non) conservativity. *Linguistic Inquiry*, 37(3):457–483.

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

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Bradley Hoot and Shane Ebert. 2021. The that-trace effect: Evidence from spanish–english code-switching. *Languages*, 6(4):189.

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.

Jong-Bok Kim and Peter Sells. 2008. *English syntax: An introduction*. CSLI publications Stanford, CA.

Marie-Pauline Krielke. 2024. Cross-linguistic dependency length minimization in scientific language: Syntactic complexity reduction in english and german in the late modern period. *Languages in Contrast*, 24(1):133–163.

Irene Lami and Joost van de Weijer. 2022. Compound-internal anaphora: evidence from acceptability judgements on italian argumental compounds. *Morphology*, 32(4):359–388.

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

Steven Langsford, Amy Perfors, Andrew T Hendrickson, Lauren A Kennedy, and Danielle J Navarro. 2018. Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: a journal of general linguistics*, 3(1):1–34.

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.

Paulina Łęska-Bayraktar and Sylwiusz Żychliński. 2023. A panoramic view of acceptability judgments in polish generative linguistics. *Poznan Studies in Contemporary Linguistics*, 59(4):705–740.

Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578.

Kyle Mahowald, Jeremy Hartman, Peter Graff, and Edward Gibson. 2016. Snap judgments: A small n acceptability paradigm (snap) for linguistic acceptability judgments. *Language*, pages 619–635.

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553.

Samuel Schindler, Anna Drożdżowicz, and Karen Brøcker. 2020. *Linguistic intuitions: Evidence and method*. Oxford University Press.

Gert-Jan Schoenmakers. 2023. Linguistic judgments in 3d: the aesthetic quality, linguistic acceptability, and surface probability of stigmatized and non-stigmatized variation. *Linguistics*, 61(3):779–824.

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.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Dominique Sportiche, Hilda Koopman, and Edward Stabler. 2013. *An introduction to syntactic analysis and theory*. John Wiley & Sons.

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.

Kevin Tang and Jason A Shaw. 2021. Prosody leaks into the memories of words. *Cognition*, 210:104601.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Lucien Tesnière. 2015. *Elements of structural syntax*. John Benjamins Publishing Company.

Etxeberria Urtzi, Tubau Susagna, Joan Borràs-Comes, and Espinal M Teresa. 2022. Polarity items in basque. *Natural Language & Linguistic Theory*, 40(2):485–504.

C.J. Van Rijsbergen. 1979. *Information Retrieval*. Butterworths.

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 Wasow and Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua*, 115(11):1481–1496.

Thomas Wasow, Roger Levy, Robin Melnick, Hanzhi Zhu, and Tom Juzek. 2015. Processing, prosody, and optional to. *Explicit and Implicit Prosody in Sentence Processing: Studies in Honor of Janet Dean Fodor*, pages 133–158.

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.