ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Four Quadrants of Difficulty: A Categorisation for Curriculum Learning in NLP

Vanessa Toborek[1,2], Sebastian Müller[1,2] and Christian Bauckhage[1,2,3]

[1]University of Bonn, Germany [2]Lamarr Institute, Germany
[3]Fraunhofer Institute IAIS, Germany

Abstract. Curriculum Learning (CL) aims to improve the outcome of model training by estimating the difficulty of training samples and scheduling them accordingly. In NLP, difficulty is commonly approximated using task-agnostic linguistic heuristics or human intuition, implicitly assuming that these signals correlate with what neural models find difficult to learn. We propose a four-quadrant categorisation of difficulty signals – human vs. model and task-agnostic vs. task-dependent – and systematically analyse their interactions on a natural language understanding dataset. We find that task-agnostic features behave largely independently and that only task-dependent features align. These findings challenge common CL intuitions and highlight the need for lightweight, task-dependent difficulty estimators that better reflect model learning behaviour.

1 Introduction

Curriculum Learning (CL) is based on the intuition that neural network training should be structured in ways that mimic human learning: starting from easier concepts before progressing to harder ones [1, 2]. Any CL strategy therefore consists of two components: a function that assigns a difficulty score to each training instance and a scheduler that determines when each instance becomes available during training. In NLP, estimating difficulty is particularly challenging because linguistic difficulty is multi-dimensional and hard to capture with a single measure [3]. As a result, CL research relies on a wide range of task-agnostic measures – such as sentence length, syntactic complexity, or readability scores – as well as approaches that draw directly on human intuition [4, 5]. At the same time, there exist conceptually distinct notions of difficulty grounded in psycholinguistics, annotation disagreement [6], or training-dynamics analysis [7]. All implicitly claim to approximate what is actually difficult for neural networks, yet their relationship to one another remains unclear. The field lacks a systematic analysis of how linguistic difficulty, human disagreement, and model learning difficulty relate. While existing work sometimes distinguishes task-specific from task-agnostic difficulty [8, 9], this distinction remains purely operational and does not aim to capture the broader conceptual space.

To address this gap, we introduce a principled, four-quadrant categorisation of difficulty that distinguishes human vs. model sources and task-agnostic vs. task-dependent information. This categorisation serves to make explicit which relationships are theoretically plausible and which are empirically demonstrable. We perform a systematic, cross-quadrant analysis across (1) task-agnostic human

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

	Task Dependency	Source	Selection of Difficulty Proxies
(1)	AGNOSTIC	HUMAN	Length, word rarity, FRE, age-of-acquisition, prevalence, concreteness, diversity, complexity, SLE
(2)	AGNOSTIC	MODEL	Perplexity
(3)	DEPENDENT	HUMAN	Inter-annotator disagreement
(4)	DEPENDENT	MODEL	Confidence, variability, correctness, loss

Table 1: Our proposed four-quadrant categorisation of difficulty proxies, distinguishing between human/ model sources and task-agnostic/dependent signals.

linguistic features, (2) task-agnostic model signals, (3) task-dependent human difficulty, and (4) task-dependent model difficulty on the SNLI dataset [10] using three different pre-trained language models. Our results show that task-agnostic difficulties behave orthogonally to task-dependent difficulties. Linguistic complexity fails to predict annotation disagreement or model learning difficulty; only task-dependent human and task-dependent model signals show meaningful alignment. These findings challenge a central assumption behind many CL strategies: that task-agnostic linguistic difficulty unilaterally captures the difficulty a model experiences during training.

The implications are twofold. First, the success of task-agnostic CL heuristics must stem from mechanisms other than *accurate* difficulty estimation, such as distributional reshaping during training by the CL scheduler. Second, annotation entropy, while highly informative, remains expensive. Our results therefore underscore the need for developing new, inexpensive ways to approximate task-dependent difficulty at pre-processing time, before model training begins.

2 A Categorisation of Difficulty Signals

The Four Quadrants of Difficulty We categorise difficulty signals into four quadrants defined by two dimensions: their source (human vs. model) and their task dependency (agnostic vs. dependent). Table 1 provides an overview.

Task-agnostic Human Difficulty (TA-H) focuses on the linguistic difficulty of an input sample independent of any downstream task. Our analysis includes surface-level heuristics such as average sentence length and word rarity, alongside the Flesch Reading Ease score (FRE), a standard metric for sentence-level readability assessment [3]. We further include psycholinguistic features – age-of-acquisition (AOA), concreteness, and prevalence – which are informative predictors of lexical complexity [11]. For syntactic complexity, we consider part-of-speech tag diversity and the average depth of a sentence dependency parse tree (complexity). Finally, we include the learned, reference-less metric SLE, which correlates well with human perception of difficulty [12].

Task-agnostic Model Difficulty (TA-M) can be obtained in NLP as large pre-trained language models allow us to examine their behaviour prior to finetuning.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Perplexity reflects how well a pre-trained model predicts the input under its learned language distribution, and therefore captures aspects of fluency and lexical expectation derived from pre-training. We use the average perplexity over masked tokens in the input as a task-agnostic model signal.

Task-dependent Human Difficulty (TD-H) can be expressed as inter-annotator disagreement. Given multiple annotations per instance, disagreement reflects human uncertainty about the correct label. Such uncertainty may arise from heterogeneous sources, including lexical or syntactic complexity, inherent semantic ambiguity, underspecification in the input, or subjective annotator variation [6]. We quantify disagreement using annotation entropy, a label-distribution-based measure of uncertainty.

Task-dependent Model Difficulty (TD-M) is often derived from a model's training dynamics when fitted on a specific task [7]. We track several statistics: the model's average confidence in the correct label, its correctness across epochs, and its variability, defined as the standard deviation of confidence across training. We additionally monitor the mean and standard deviation of the loss. These metrics provide a post hoc view of how difficult each instance is for the model to learn.

Expected Interactions Between Quadrants The four-quadrant categorisation provides an analytical framework to reason about which interactions between difficulty signals are plausible, while leaving their strength and direction as empirical hypotheses:

H1: Internal coherence of TA-H. Linguistic difficulty is known to be multi-dimensional, with lexical, syntactic, and conceptual complexity capturing distinct aspects. Hence, we expect low internal correlation among TA-H features.

H2: TA-H $\leftrightarrow$ *TA-M.* Perplexity reflects a pre-trained language model's surprisal over an input sequence, driven by its learned lexical and syntactic expectations. Because several TA-H features partially relate to lexical predictability, we expect moderate correlations between perplexity and human linguistic features.

H3: TA-H $\leftrightarrow$ *TD-H.* Building on observations that linguistic complexity may contribute to human disagreement [6], we expect moderate correlations between linguistic difficulty and annotation entropy.

H4: TA-H $\leftrightarrow$ *TD-M.* Given that CL strategies using surface linguistic features have proven effective in some settings, we expect interactions between linguistic difficulty and model learning difficulty.

H5: TD-H $\leftrightarrow$ *TD-M.* Prior work on "dataset cartography" shows that instances with high label ambiguity tend to be harder for models to learn [7]. We expect a positive relationship between human disagreement and model training difficulty.

Although these hypotheses are grounded in prior work, the field lacks a systematic evaluation of whether these theoretically motivated interactions hold in practice. This categorisation of difficulty signals serves as the conceptual foundation for our empirical study. Importantly, none of these relationships follow directly from the quadrant definitions alone; rather, they constrain the space of meaningful interactions and motivate explicit hypotheses that require empirical validation.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

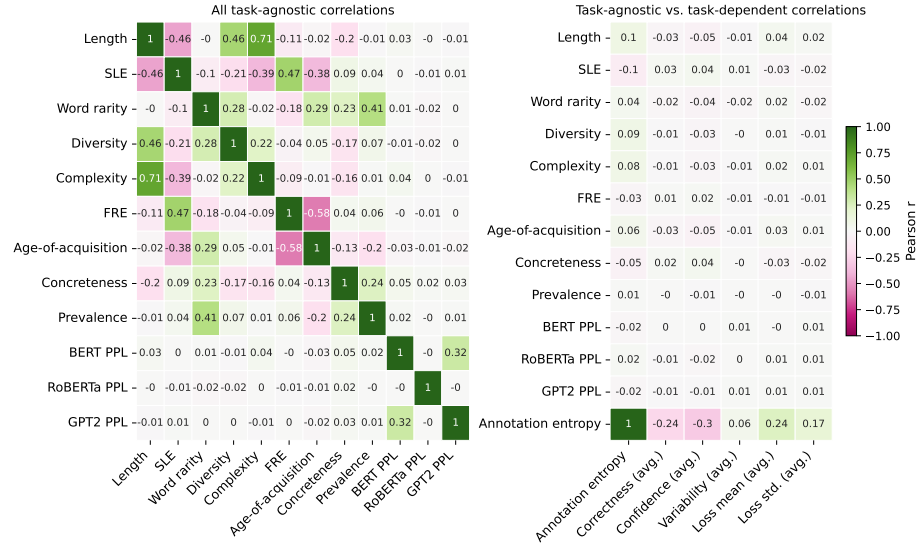

Fig. 1: Pearson correlation for the four difficulty quadrants. (Left) Correlation among task-agnostic human and task-agnostic model difficulty signals. (Right) Correlation for all task-agnostic difficulty signals with all task-dependent ones. All TD-M are averaged over all three models and ten random seeds each.

3 Testing the Quadrant Interactions

We empirically evaluate our hypotheses on the interactions between difficulty quadrants using the SNLI dataset [10]. It provides four independent annotator labels for each instance in the training split. This allows for computing annotation entropy as the TD-H difficulty proxy. TA-H features are calculated in a preprocessing step. For TA-M features we compute perplexities using pre-trained BERT-base [13], RoBERTa-base [14], and GPT-2-base [15] models. TD-M features are collected for 3 647 data points at twelve evenly spaced checkpoints during finetuning. BERT/RoBERTa are trained with batch size b=64 and learning rate lr=3×10^{-5}, GPT-2 with b=16 and lr=10^{-5}. All models use 5 epochs, AdamW (weight decay 0.01), and a linear lr schedule with 6% warm-up.

We evaluate the expected relationship between quadrants in four ways: (1) we compute the Pearson correlation (a) within the TA-H quadrant, (b) between TA-H and TD-H, and (c) between TA-H and TD-M. (2) To assess whether all TA features jointly predict either TD difficulty, we perform multivariate regression using either TD-H or TD-M features as targets. (3) Following the diagnosing approach of dataset cartography [7], we identify the top and bottom 25% of easy-to-learn and ambiguous data points. We then compare the distribution on task-agnostic features to test whether they meaningfully separate them. (4) As a sanity check, we replicate the established relationship between annotation entropy and model ambiguity [7] for the different model architectures.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

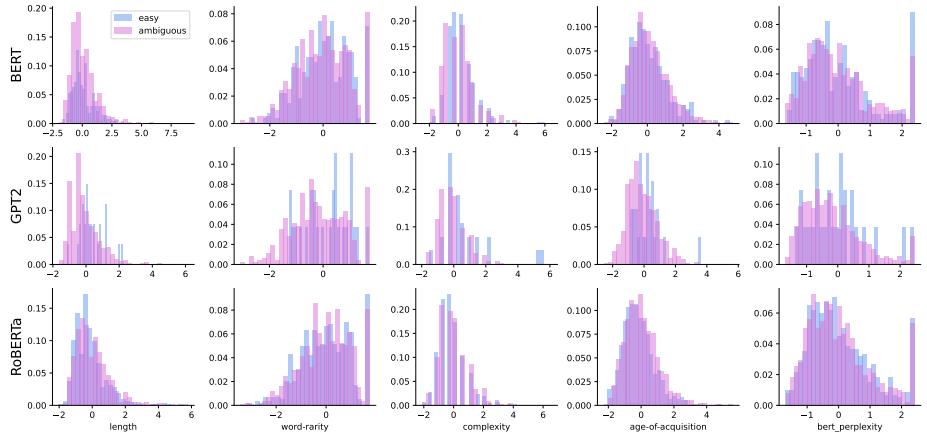

Fig. 2: Histograms of easy vs. ambiguous samples (based on the respective model confidence and variability) for selected task-agnostic proxies: length, word rarity, complexity, age of acquisition, and BERT perplexity.

Together, these analyses provide a comprehensive evaluation of whether difficulty signals across the four quadrants show the hypothesised interaction.

Results Figure 1 (left) shows the Pearson correlations among task-agnostic difficulty signals. Supporting *H1*, TA-H correlations are mostly moderate ($r < 0.5$) to low ($r < 0.2$), with only length-complexity ($r = 0.71$) and AOA-FRE ($r = -0.58$) standing out. This indicates that TA-H signals capture distinct facets of linguistic difficulty rather than a single underlying factor. Turning to *H2*, the bottom of Figure 1 (left) shows that perplexity does not correlate with TA-H difficulty signals, providing no evidence for the hypothesised weak alignment. Figure 1 (right) extends this picture: all TA features show virtually no correlation with TD-H (*H3*) or any TD-M signal (*H4*). This suggests that (a) SNLI label disagreement is likely driven by inherent ambiguity and less by linguistic complexity, and (b) surface-level linguistic properties are poor predictors of what models find difficult to learn. To test these relationships multivariately, we regress each task-dependent signal on all TA features using both linear and tree-based models. Predictive power remains negligible ($R^2 < 0.05$ for annotation entropy and $R^2 < 0.1$ for all TD-M metrics), reinforcing the weak link between TA-H and TD difficulty. Distributional analysis of "easy" vs. "ambiguous" instances, defined via dataset cartography, provides further evidence. As shown in Figure 2, the distributions of task-agnostic features overlap almost entirely for both groups across all models, indicating that linguistic difficulty does not separate easy-to-learn from ambiguous samples. Finally, the bottom row of Figure 1 (right) confirms the expected correlation structure among task-dependent signals (*H5*): annotation entropy aligns with correctness, confidence, and average loss, reflecting their shared dependence on label uncertainty.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

4 Lessons for Curriculum Learning

Our analysis on the SNLI dataset reveals that task-agnostic and task-dependent difficulty signals are largely independent, challenging the common implicit assumption that shallow linguistic proxies faithfully reflect model learning difficulty. Yet, both types of difficulty measures have been reported to yield successful curricula, suggesting that CL gains cannot be attributed to difficulty estimation alone. Instead, improvements may likely emerge from interactions between difficulty measures, task characteristics, and scheduling decisions, in particular from how difficulty is enforced and how difficulty exposure evolves throughout training. We further confirm that task-dependent model difficulty aligns with annotator entropy, supporting human disagreement as a meaningful signal of task-dependent difficulty, but one that is costly to obtain at scale. This motivates developing lightweight, pre-computable approximations of task-dependent difficulty that better reflect the model's learning behaviour and enable more principled curriculum design, as well as further work to identify the mechanisms by which a CL strategy improves training.

References

[1] J. L. Elman. Learning and Development in Neural Networks: The Importance of Starting Small. *Cognition*, 48(1), 1993.

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum Learning. In *ICML*, 2009.

[3] A. Battisti, D. Pfütze, A. Säuberli, M. Kostrzewa, and S. Ebling. A Corpus for Automatic Readability Assessment and Text Simplification of German. In *LREC*, 2020.

[4] M. Elgaar and H. Amiri. HuCurl: Human-induced Curriculum Discovery. In *ACL*, 2023.

[5] V. Toborek, S. Müller, T. Selbach, T. Horváth, and C. Bauckhage. Beyond Shallow Heuristics: Leveraging Human Intuition for Curriculum Learning. In *ICNLSP*, 2025.

[6] B. Plank. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *EMNLP*, 2022.

[7] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *EMNLP*, 2020.

[8] F. Christopoulou, G. Lampouras, and I. Iacobacci. Training Dynamics for Curriculum Learning: A Study on Monolingual and Cross-lingual NLU. In *EMNLP*, 2022.

[9] V. Toborek, F. Seiffarth, S. Müller, T. Horváth, and C. Bauckhage. Exploring Curriculum Learning for Languages: Lessons from Regular Language Tasks. In *Discovery Science*, 2025.

[10] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015.

[11] A. T. Desai, K. North, M. Zampieri, and C. Homan. LCP-RIT at SemEval-2021 Task 1: Exploring Linguistic Features for Lexical Complexity Prediction. In *SemEval-2021*, 2021.

[12] L. Cripwell, J. Legrand, and C. Gardent. Simplicity Level Estimate (SLE): A Learned Reference-Less Metric for Sentence Simplification. In *EMNLP*, 2023.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *ACL*, 2019.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, 2019.

[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 2019.