A Tale Of Two Long Tails

Daniel D'souza [1 2] **Zach Nussbaum** [1] **Chirag Agarwal** [3] **Sara Hooker** [4]

Abstract

As machine learning models are increasingly employed to assist human decision-makers, it becomes critical to communicate the uncertainty associated with these model predictions. However, the majority of work on uncertainty has focused on traditional probabilistic or ranking approaches – where the model assigns low probabilities or scores to uncertain examples. While this captures what examples are challenging for the model, it does *not* capture the underlying source of the uncertainty. In this work, we seek to identify examples the model is uncertain about *and* characterize the source of said uncertainty. We explore the benefits of designing a targeted intervention – targeted data augmentation of high uncertainty examples over the course of training. We ask – does the rate of learning in the presence of additional information differ between atypical and noisy examples? Our results show that this is indeed the case, suggesting that well designed interventions over the course of training can be an effective way to characterize and distinguish between different sources of uncertainty.

1. Introduction

As machine learning models are increasingly implemented in real-world applications, it becomes important to estimate the uncertainty in the predictions of these models and ensure that model behavior is safe and trustworthy. Traditional approaches to uncertainty estimation use a probabilistic approach – where examples a model is uncertain about are assigned low probabilities or scores (Denker and LeCun, 1990; Hendrycks and Gimpel, 2016; Erfani et al., 2016; Ruff et al., 2018; Parzen, 1962; Rosenblatt, 1956; Hawkins, 1974; Vandeginste, 1988). While probabilities and other scores are a useful way to isolate examples where the model is uncertain, the estimate of uncertainty is fundamentally limited in that they capture what predictions are challenging for the model but *not* the underlying source of the uncertainty.

(a) Reducible Error (b) Irreducible Error

Figure 1. Examples of different predictive uncertainties. **Left:** An instance of the `horse` class representing error reducible using more data examples. **Right:** A `horse` image mislabelled as a `donkey`, representing irreducible error as the model cannot learn this class distribution even with more examples because of the corrupted label.

Most natural image and language datasets exhibit a long-tail distribution with an unequal frequency of attributes in training data (Zipf, 1999; Feldman, 2020). However, the nature of these low-frequency attributes differ considerably. Atypical examples are rare or unusual underrepresented attributes – data points sampled from sparsely populated regions of the input space. Poor model performance on atypical examples reflects *epistemic* uncertainty, where there is insufficient evidence for the model to learn the feature. In contrast, noisy examples are due to influences on the data-generating process, such as label corruption or input data perturbation, which impairs the learnability of the instance. These noisy examples are dominated by *aleatoric* uncertainty or *irreducible error* because the mapping between the input and output space is entirely stochastic. Recent work has suggested that labelling noise is widespread in widely used datasets, and can constitute a large fraction of the training set (Hooker et al., 2020; Northcutt et al., 2021; Beyer et al., 2020).

The need for a framework to estimate both the level *and* source of uncertainty is driven by the very different downstream remedies for different sources of uncertainty. For

[*]Equal contribution [1]ML Collective [2]ProQuest LLC [3]Harvard University [4]Google Research, Brain Team. Correspondence to: Daniel D'souza <ddsouza@umich.edu>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning., Copyright 2021 by the author(s).

sources of high *epistemic* uncertainty, such as low-frequency attributes or challenging fine-grained samples (atypical), a practitioner can improve model performance by either collecting more data that are similar or re-weighting examples to improve model learning of this instance (or other active learning techniques) (Budd et al., 2021; Zhang et al., 2019; Liu et al., 2020). In contrast, for causes of *aleatoric* uncertainty, such as noisy examples, solutions like down-weighting or elimination through data cleaning are advocated (Zhang et al., 2020; Li et al., 2019; Thulasidasan et al., 2019b; Liu and Guo, 2020; Schroder and Niekler, 2020).

Despite the importance of identifying the sources of predictive uncertainty, this subject has been relatively under-treated in ML literature. Probabilistic frameworks will accord high uncertainty to both underrepresented attributes and noisy examples, failing to distinguish between the two. While sub-fields have evolved separately around the treatment of low-frequency and noisy distributions (Wu et al., 2020; Yi and Wu, 2019; Thulasidasan et al., 2019a), only limited work to date has focused on the sources of uncertainty within a unified framework (Kendall and Gal, 2017; Depeweg et al., 2018).

Present Work. In this work, we seek to identify examples the model is uncertain about *and* characterize the source of said uncertainty. We leverage the key distinguishing difference between *Epistemic* and *Aleatoric* uncertainty – one is reducible given additional data and the other is not. We propose targeted data augmentation throughout training to amplify the difference in learning rate between atypical and noisy examples. Our results show well designed interventions over the course of training can be an effective way to cluster and distinguish between different sources of uncertainty.

2. Methodology

2.1. Sources of Uncertainty

We consider a supervised learning setting where we denote the training dataset $\mathcal{D}$ as:

$$\mathcal{D} \stackrel{\text{def}}{=} \big\{(x_1, y_1), \ldots, (x_N, y_N)\big\} \subset \mathcal{X} \times \mathcal{Y} \ , \qquad (1)$$

where $\mathcal{X}$ represents the data space and $\mathcal{Y}$ the set of outcomes associated with the respective instances. We consider a neural network as a function $f_w : \mathcal{X} \mapsto \mathcal{Y}$ with trainable weights w. Given the training dataset, f_w optimizes a set of weights w^* by minimizing an objective function L,

$$w^* = \arg\min_w L(w) \qquad (2)$$

Here, we aim to quantify the uncertainty associated with a model prediction, and to subsequently identify the source of the uncertainty by classifying examples as contributing disproportionately to *aleatoric* or *epistemic* uncertainty. This means we are interested in firstly obtaining a good measure of *predictive uncertainty*, the uncertainty related to the prediction $\hat{y}_i$ for an input instance $\mathbf{x}_i \in \mathcal{X}$. To this end, we leverage Variance of Gradients (VoG) (Agarwal and Hooker, 2020), a class-normalized variance gradient score for determining the relative ease of learning data samples within a given class. Important for our use case, VoG produces a per-example score that can be used to rank the entire dataset. Secondly, it measures change in gradient updates over the course of training – which we use to measure the impact of targeted interventions on model uncertainty. Given the predicted or true class label p, VoG first calculates the gradients of the pre-softmax activation layer l with respect to each pixel x_i and sums it across the color channels to arrive at $\mathbf{S} \in \mathbb{R}^{32\times32}$, *i.e.,*

$$\mathbf{S} = \frac{\partial A_p^l}{\partial x_i} \qquad (3)$$

The variance of gradients is then calculated across each pixel using the gradients from a set of K checkpoints, *i.e.,* $\{\mathbf{S}_1, \ldots, \mathbf{S}_K\}$.

$$\text{VoG}_p = \sqrt{\frac{1}{K}\sum_{t=1}^{K}(\mathbf{S}_t - \mu)^2} \qquad (4)$$

Finally, the score is averaged across all pixels N_p to compute a scalar VoG score.

$$\text{VoG} = \frac{1}{N_p}\sum_{p=1}^{N}(\text{VoG}_p) \qquad (5)$$

2.2. Characterizing the difference between atypical and noisy examples

An accurate uncertainty score $s(f_w, \mathbf{x}_i)$ for an instance $\mathbf{x}_i \in \mathcal{X}$ reflects the accumulation of uncertainty in the data curation process, the set of modeling choices and the training protocol itself. Thus, the possible outcome $\hat{y}_i$ depends upon the dataset $\mathcal{D}$ and the underlying model f_w. Intuitively, $s(f_w, \mathbf{x}_i)$ is a composition of both aleatoric ($s_a(f_w, \mathbf{x}_i)$) and epistemic knowledge ($s_e(f_w, \mathbf{x}_i)$).

One way to characterize uncertainty as aleatoric or epistemic is to ask whether it can be reduced through additional training data. In this work, we apply transformations θ to the training set $\mathcal{D}$, resulting in a new set $\mathcal{D}_A$. The stochasticity of the transformation parameters is responsible for generating new samples, *i.e., data augmentation*.

We evaluate the impact of providing additional information for all training examples $\forall \mathbf{x}_i \in \mathcal{X}$ relative to not providing additional information. This amounts to comparing standard data augmentation to no data augmentation. In addition, we also explore the benefits of designing a targeted intervention

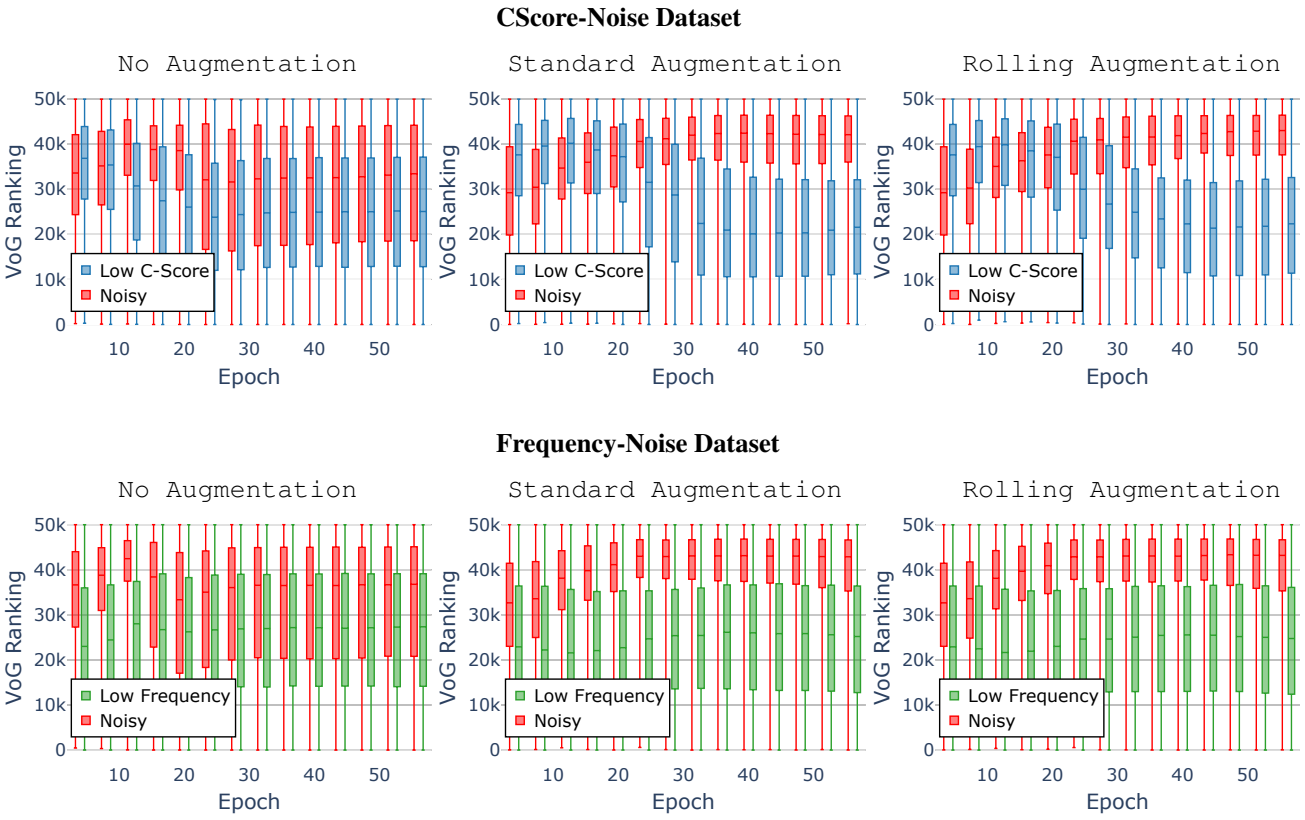

Figure 2. VoG ranking for atypical and noisy subsets across training for different augmentation variants

– where rolling augmentation only selectively augments samples where the model is uncertain. For the first 3 epochs, rolling augments 100% i.e all of the samples. From the 4^{th} epoch, we utilize the VoG score to rank and begin selectively augmenting the top X% VoG percentile, with X reducing with every following epoch. X begins at 100% and is then reduced by 2% for every following epoch till we reach 20%. At this point, we continue to augment the images in the top 20% of the VOG ranking till the end of the training.

Given our hypothesis that additional information will help distinguish between reducible and irreducible error, we expect differences in the distributions of atypical and noisy to be amplified for the two augmentation schemes *relative* to no augmentation.

2.3. Experimental Framework

Dataset Construction. To understand how atypical and noisy examples are learned, we stratify atypical and noisy subsets where ground truth is known for these examples. We briefly describe the details below:

Atypical sub-sets. We construct two different atypical subsets based upon two different notions of typicality: 1) *Frequency* and 2) *Consistency*:

1. **Frequency.** We artificially create a known frequency disparity between examples. We sample a fraction of the dataset at random p. Of the remaining dataset, we sample a fraction t and create two copies of each example. t is selected such that the overall dataset size is the same as the original unmodified dataset.

2. **Consistency.** For the consistency setting, we utilize the C-Score (Jiang et al., 2020) as a pseudo-measure of how *typical* an example is *wrt* the other samples in the dataset. We directly use the pre-computed C-scores [1] available for Cifar-10 as a continuous measure of typicality.

For modeling a noisy subset, we follow (Zhang et al., 2016) and assign uniformly shuffled labels to a percentage of the training data. More specifically, this decision models noisy data as mislabelled instances. For both variants, we maintain a ratio of 20% noisy, 20% atypical, and 60% typical examples and ensure that it is the same size as the original dataset.

Frequency-Noise Dataset. We uniformly sample 20% (A) from Cifar-10 training set as atypical candidates. We sample another random 20% (N) from the remaining examples, uni-

[1] Available from https://pluskid.github.io/structural-regularity/

Table 1. Testing accuracies for different augmentation variants of LongTail datasets.

DATASET	VARIANT	TEST ACC
FREQUENCY	NO AUG.	60.6%
	STANDARD AUG.	67.7%
	ROLLING AUG.	72.0%
C-SCORE	NO AUG.	72.1%
	STANDARD AUG.	76.9%
	ROLLING AUG.	78.7%

formly shuffle the labels and consider them as Noisy candidates. Finally, we sample 30% from the remaining Cifar-10 dataset, create two copies and add that as Typical candidates.

CScore-Noise Dataset. For the `CScore-Noise` dataset, we consider the bottom 20%(A) C-Score ranked images as atypical. Similar to the `Frequency-Noise` dataset, we then sample another random 20%(N) from the remaining examples, uniformly shuffle the labels and add these as Noisy candidates. The remaining 60% Cifar-10 dataset is considered as Typical.

We note that `CScore-Noise` preserves all original datapoints in Cifar-10 training set, whereas `Frequency-Noise` downsamples the number of original examples in order to maintain the same training set size. While `Frequency-Noise` and `CScore-Noise` differ in the construction of the atypical subset, both have the same fraction of noisy examples.

Training details. For all our Cifar-10 experiments, we use a WideResNet (Zagoruyko and Komodakis, 2016) architecture. We train for 60 epochs using stochastic gradient descent (SGD). For training variants where augmentation is present, we use standard data augmentation by applying random horizontal flips and crops with padding. We use a base learning rate schedule of 0.1 and adaptively dampen it by a factor of 0.2 at the 10^{th}, 20^{th} and 30^{th} training epochs. For our baseline variant on a clean dataset (no artificial stratification of noisy and atypical) we report a top-1 test-set accuracy of 93.11% for Cifar-10.

3. Results

We address the following key questions: Q1) Can the presence of additional information amplify differences in the rate of learning of atypical and noisy examples? and Q2) Do different atypical subsets exhibit different separability?

Q1) Characterizing differences between atypical and noisy subsets across training. In Fig. 2, we plot the distribution of ranks based on class-normalized VoG scores for both noisy and atypical samples across training. We now describe the effect of different augmentation variants on separating the noisy and atypical subsets.

No Augmentation. In Fig. 2, we observe a large overlap in the distribution of atypical and noisy examples in a training setting without augmentation. While there is a perceptible drop in Atypical VoG ranks relative to Noisy VoG ranks after ten epochs, the level of overlap between atypical and noisy remains high.

Standard Augmentation. We observe better separation between the noisy and atypical distributions using standard augmentation during training (Fig. 2). The addition of information, even if done uniformly, provides the model with additional examples of the atypical instances. Atypical VoG ranks fall more markedly, while Noisy VoG ranks remain at the top of the distribution.

Rolling Augmentation. In Fig. 2, it is clear that rolling augmentation provides a notable improvement over no augmentation at helping distinguish between atypical and noisy examples. We note that rolling Augmentation leads to a slightly more pronounced separation of noisy and atypical examples than standard augmentation.

Q2) Differences between datasets. We note that this separability is more visible on the `CScore-Noise` dataset than the `Frequency-Noise` dataset. We believe this is because our frequency constraints limited the training dataset to 70% of the typical 50,000 examples (in order to make 2 copies and maintain the same training set size, only 30% of the dataset was sampled and expanded).

4. Conclusion

We leverage targeted augmentation interventions to characterize examples as dominated by *aleatoric* and *epistemic* uncertainty. We empirically show how augmentation protocols (both targeted and standard) amplify the differences in distribution between noisy and atypical examples. The slight improvement of targeted rolling augmentation over standard opens up interesting questions around 1) gaining a better understanding of what types of augmentations aid distinction between atypical and noisy, 2) experimenting with different augmentation protocols.

The motivation for this work is to eventually treat the atypical and noisy subsets in a dataset appropriately during training. Common remedies for noisy examples would include data cleaning or isolating a subset for re-labelling, whereas atypical examples may benefit from data augmentation, reweighting or additional data collection. An interesting area of future work is leveraging these targeted augmentations to inform these downstream remedies.

REFERENCES

Chirag Agarwal and Sara Hooker. Estimating example difficulty using variance of gradients, 2020.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv e-prints*, art. arXiv:2006.07159, June 2020.

Samuel Budd, Emma C. Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062, Jul 2021. ISSN 1361-8415. doi: 10.1016/j.media.2021.102062. URL http://dx.doi.org/10.1016/j.media.2021.102062.

John S. Denker and Yann LeCun. Transforming neural-net output levels to probability distributions. In *Proceedings of the 3rd International Conference on Neural Information Processing Systems*, NIPS'90, page 853–859, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc. ISBN 1558601848.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1184–1193. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/depeweg18a.html.

Sarah M Erfani, Sutharshan Rajasegarar, Shanika Karunasekera, and Christopher Leckie. High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recognition*, 58:121–134, 2016.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 954–959, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369794. doi: 10.1145/3357713.3384290. URL https://doi.org/10.1145/3357713.3384290.

Douglas M Hawkins. The detection of errors in multivariate data using principal components. *Journal of the American Statistical Association*, 69(346):340–344, 1974.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *CoRR*, abs/1610.02136, 2016. URL http://arxiv.org/abs/1610.02136.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget?, 2020.

Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C. Mozer. Characterizing Structural Regularities of Labeled Data in Overparameterized Models. *arXiv e-prints*, art. arXiv:2002.03206, February 2020.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf.

P. Langley. Crafting papers on machine learning. In Pat Langley, editor, *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pages 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Learning to learn from noisy labeled data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5046–5054, 2019. doi: 10.1109/CVPR.2019.00519.

Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Y. Liu and Hong-Yi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. *ArXiv*, abs/1910.03231, 2020.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks, 2021.

Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

M Rosenblatt. Remarks on some nonparametric estimates of a density function. annals of mathematical statistics. 1956.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

Christopher Schroder and Andreas Niekler. A survey of active learning for text classification using deep neural networks, 2020.

Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. Combating label noise in deep learning using abstention. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6234–6243. PMLR, 09–15 Jun 2019a. URL http://proceedings.mlr.press/v97/thulasidasan19a.html.

Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. Combating label noise in deep learning using abstention, 2019b.

Bernard GM Vandeginste. Robust regression and outlier detection. pj rousseeuw and am leroy, john wiley & sons, new york, 1987. no. of pages: 329. price:£ 31.95. isbn: 0 471 85233 3, 1988.

Pengxiang Wu, Songzhu Zheng, Mayank Goswami, Dimitris Metaxas, and Chao Chen. A topological filter for learning with label noise, 2020.

Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016. URL http://arxiv.org/abs/1605.07146.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

J. Zhang, Lingqiao Liu, P. Wang, and Chunhua Shen. To balance or not to balance: An embarrassingly simple approach for learning with long-tailed distributions. *ArXiv*, abs/1912.04486, 2019.

Zizhao Zhang, Han Zhang, Sercan O. Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

G.K. Zipf. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Cognitive psychology]. Routledge, 1999. ISBN 9780415209762. URL https://books.google.com/books?id=w1Z4Aq-5sWMC.