Changing the World by Changing the Data

Anna Rogers
Center for Social Data Science
University of Copenhagen
arogers@sodas.ku.dk

Abstract

NLP community is currently investing a lot more research and resources into development of deep learning models than training data. While we have made a lot of progress, it is now clear that our models learn all kinds of spurious patterns, social biases, and annotation artifacts. Algorithmic solutions have so far had limited success. An alternative that is being actively discussed is more careful design of datasets so as to deliver specific signals. This position paper maps out the arguments for and against data curation, and argues that fundamentally the point is moot: curation already is and will be happening, and it is changing the world. The question is only how much thought we want to invest into that process.

1 Introduction

The key ingredient behind the recent successes in NLP is Transformer-based language models. The paradigm of pre-training followed by fine-tuning on downstream tasks was popularized by BERT (Devlin et al., 2019), and is actively developed (Rogers et al., 2020b). In December 2020 the human performance baselines on SuperGLUE (Wang et al., 2019a) were surpassed twice, making the community wonder if it is possible to formulate benchmarks not solvable in this paradigm.

However, the successes are not the full story. It is becoming increasingly clear that much of the remarkable performance is down to benchmarks that do not actually require sophisticated verbal reasoning skills due to annotation artifacts and spurious patterns correlating with the target labels (Gururangan et al., 2018; McCoy et al., 2019; Paullada et al., 2020).The social biases in NLP models are also attracting more attention (Sheng et al., 2019; Davidson et al., 2019; Hutchinson et al., 2020).

The "garbage in, garbage out" principle suggests that the situation will not change without a dramatic reappraisal of how NLP data is collected, both for pre-training and task-specific resources. But that seemingly uncontroversial conclusion is at the core of the interdisciplinary tension between NLP understood as a deep learning (DL) application area, and the more qualitative approaches of computational linguistics and AI ethics. How deep that tension goes is illustrated by the recent heated (and sometimes less than professional[1]) debate around "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜" by Bender, Gebru et al (2021).

This position paper brings together the arguments for and against curating data[2] from linguistic and ethical perspectives (§2). It makes the case that curation is unavoidable and already happening, and that any data choices that we make, explicitly or implicitly, *will* affect the real world (§3). Thus the debate is only about how much thought we should put into this process. If we are to at least try to steer it, we have to overcome the interdisciplinary tension and reconsider what counts as "NLP work" (§4). §5 outlines some policies that could help.

2 To Curate or Not to Curate?

2.1 Why Change the Data?

The core argument for active curation/design of the data that goes into NLP models is that the models are representations of the data they were trained on, and thus data work is necessary to make sure that the models can learn what we need them to learn. The supporting evidence for this position

[1] https://www.theverge.com/22309962/timnit-gebru-google-harassment-campaign-jeff-dean

[2] In this paper "data curation" is interpreted broadly as making choices about what should be included in a NLP resource (either for pre-training or task-specific data). The phenomena to be included/excluded could be defined in terms of *what* is said (e.g. soccer commentary), *how* it is expressed (e.g. with or without expletives), and/or *who* is speaking or being addressed (e.g. teenage soccer fans).

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 2182–2194
August 1–6, 2021.

comes independently from several directions: the studies finding that the models fail to learn a certain phenomenon and/or learn something undesirable.

2.1.1 Social biases

Our world is far from perfect, and written texts contain plenty of evidence of all kinds of social biases based on gender, race, social status, ability, age, etc. Models may learn these biases (from pre-training and/or task data) and even amplify them, putting the minority groups at a disadvantage by direct psychological harm and propagation of stereotypes (Blodgett et al., 2020; Bender et al., 2021). In this context, "data curation" means selecting data based on its sociocultural characteristics (Jo and Gebru, 2020). Fundamentally, this is about fair representation for different social groups.

Some dismiss Bender et al. (2021) as "political", or even "advocacy rather than research" (Lissack, 2021). However, "papers advocate for specific research agendas all the time" (Venkatasubramanian, 2021). NLP in particular has a growing subfield of bias mitigation (see e.g. the survey on such work for gender bias by Sun et al. (2019)) that pursues exactly the same social justice agenda, but does not receive the same pushback.

2.1.2 Privacy concerns.

Models may memorize specific facts in training data, and if those facts happen to be personally identifiable information, this is a security concern. For instance, Carlini et al. (2020) showed that GPT-2[3] was able to memorize personal contact information, even if it only appeared on a few web pages. A big problem is that this is not a bug, but a feature: we do want our language models to represent some facts about presidents – just not about private citizens. Deciding what should not be remembered is clearly a data curation issue.

2.1.3 (Lack of) progress towards NLU

DL models are data-hungry, and so far we have heavily relied on the sources that are easy to scale: web texts for pre-training, and crowdsourcing for annotation or generating shorter texts. Combined with most funding and effort allocated to model development, this meant a less clear view of what was in the data. Consequently, the recent years witnessed a lot of findings along the following lines.

[3]Google legal department reportedly requested edits to the article by Carlini et al. (2020), in particular to avoid mentions of Google technology (Dave, 2021).

DL models learn spurious patterns present in the data. These patterns can be the results of the heuristics used by crowd workers (Gururangan et al., 2018), small samples of workers creating large parts of data with traces (Geva et al., 2019), or simply random patterns in the task or pre-training data. For example, words like *football* may frequently occur in abusive tweets, but this should not give the model the idea that all sports fans are violent (Wiegand et al., 2019). The result is that many current datasets can (and do) get "solved" with shallow cues such as lexical co-occurrence (Jia and Liang, 2017; McCoy et al., 2019). The larger the resource, the more difficult it is to avoid them (Gardner et al., 2021).

DL models are surprisingly vulnerable to basic perturbations. ACL 2020 best paper award went to Ribeiro et al. (2020)'s demonstration that even the successful, commercially deployed NLP systems cannot handle many core linguistic phenomena like negation. Pre-trained language models by themselves do not necessarily cope with them either (Ettinger, 2020). This suggests that the current resources do not provide the signal to learn the necessary linguistic paradigms.

DL models struggle to learn rare phenomena. Linguistic phenomena generally follow Zipf distribution (Zipf, 1945), which means that most of them are rare in naturally occurring data, and thus harder for the models to learn. This applies even to the large pre-training datasets. For example, Zhang et al. (2020) compared the learning rates for different linguistic phenomena as RoBERTa was pre-trained on more and more data. English irregular verb forms (highly frequent) were acquired in under 10M of training tokens, but the model struggled with island effects even after 30B tokens. Such results suggest that if something needs to be learned, the model needs to be provided with a sufficiently strong signal (and it may still fail even then (Geiger et al., 2019)).

The bottom line is that the distributions of linguistic phenomena in the current NLP resources do not seem to provide the signal with which the current models could learn to perform human-level language "understanding". We do not even know the full spectrum of abilities that would qualify for that. Choosing which aspects of a given "task" (or language, in case of pre-training) a given resource would "teach" explicitly is a curation decision.

2.1.4 Security concerns.

A relatively recent development is "universal adversarial triggers": adversarial attacks on the models that modify the textual input in a way that forces the models to always output a certain prediction (Wallace et al., 2019). For example, the authors make a SQuAD-trained reading comprehension model to always predict the answer "to kill American people" for any why-question. This effect is robust and model-independent: i.e. it is the training data that gets "hacked", not the model.

It is not clear if it is possible to construct a dataset that would not have such vulnerabilities, but common sense suggests that the training data should be curated so as to make them unlikely to occur in the natural distribution of user input.

2.1.5 Evaluation methodology.

So far the fundamental paradigm for NLP work based on machine learning focused on in-distribution evaluation: the test sample would come from the same distribution as the train/validation samples, and the samples would be randomly split. Within that paradigm, it is essential that there are no overlaps between training and test data, which is an issue for many current resources (Lewis et al., 2021; Emami et al., 2020).

To do that well, we already have to make decisions about what counts as "overlap", and what should be in the training and testing data. For example, in pre-training GPT-3 (Brown et al., 2020) decisions had to be made about which benchmarks would be used for evaluation. There was a (partially successful) attempt to simply remove documents with significant overlap with any test examples from the training data, which raises a new issue: if the goal is to train a "general-purpose" model, what information could we safely exclude from training purely for evaluation purposes?

Linzen (2020) suggests switching to out-of-distribution testing: given that the training data is unlikely to faithfully represent a full range of linguistic phenomena, in-distribution evaluation likely overstates how well the model is doing. But to do that, we would still need to know what is "in" the training distribution, and what we would be testing.

To sum up, there are (at least) 4 reasons to make deliberate decisions about what should be included in the training data, so as to create more robust, inclusive, and secure NLP models. What are the objections?

2.2 Why Not to Change the Data?

Since this is a position paper arguing that data curation is unavoidable, the arguments against it are presented together with the defense. Most of them are applicable to both pre-training and task data (except for §2.2.2, which focuses on pre-training).

2.2.1 Studying the world "as it is".

In response to Bender et al., Goldberg (2021) argued that there are valid use cases in which "a model of language use should reflect how the language is actually being used", rather than how we believe it should be used.

Defense. This is a completely valid argument, and what follows is elaboration rather than refutation. In linguistic or social science research, it is uncontroversial that if the corpus is a representative sample of the target phenomena, it should not be manipulated. If the goal is to model the worldview of Reddit users, the corpus used for training GPT-2 (comprising articles shared on Reddit) is a representative sample. Likewise, if the goal is to study social biases, we should not eliminate e.g. racist comments. The problem raised by Bender et al. (2021) is only that resources should be used for what they are: the Reddit users are not a representative sample of the general population, and so GPT-2 is not a "general-purpose" language model.

This argument concerns the qualitative studies of the "world as it is". Most NLP research, however, aims to produce systems that would perform some task. In that case the "natural" distribution may not even be what we want: e.g. if the goal is a question answering system, then the "natural" distribution of questions asked in daily life (with most questions about time and weather) will not be helpful. The developers may also prefer for their systems to be e.g. less racist/sexist than their input data.

Note that to study the world "as it is" we still have to do a lot more data work than we are currently doing (so as to be able to tell whether a given corpus actually represents the target phenomenon).

2.2.2 Our sample is large enough.

An anonymous reviewer of this paper contributed the following argument: "the size of the data is so large that, in fact, our training sets are not a sample at all, they are the entire data universe".

Defense. This argument would stand if the "data universe" that we use for training NLP systems were the same as "the totality of human

speech/writing". It is not, and will hopefully never be, because collecting *all* speech is problematic for ethical, legal, and practical reasons. Anything less than that is a sample. Given the existing social structures, no matter how big that sample is, it is not representative due to (at least) unequal access to technology, unequal possibility to defend one's privacy and copyright, and limited access to the huge volumes of speech produced in the "walled garden" platforms like Facebook. The use of uncontrolled samples (like the Common-Crawl-based corpora) would have to be justified by arguing either that the above types of bias can be safely ignored, or that the benefits outweigh the risks.

2.2.3 Might not be the best approach.

Do we really have to do hard data work, or could there be an algorithmic solution? For the problem of rare phenomena (§2.1.3), there is ongoing work on inductive biases that could help the models learn them (McCoy et al., 2020). For social issues (§2.1.1) Goldberg (2021) and Buckman (2021) similarly suggest that rather than trying to filter out problematic samples (hate speech, racial slurs etc.) we could use them to build a representation of the undesirable phenomena, and to try to actively identify and filter them out in generation. Schick et al. (2021) propose a method for a generative language model to reduce biases in its output, using self-diagnosis with its own internal knowledge.

Defense. It is entirely possible that algorithmic alternatives could work better than solutions based on data curation. Which one will be more successful is an empirical question. As of now, it seems that they are complementary rather than mutually exclusive: for example, some specific biases could be handled algorithmically, but data curation could be used to balance the corpus in some other way(s).

Note that *the algorithmic solutions would still require much of the same data work for evaluation purposes*: to find out whether a system is effective at filtering out something undesirable or processing some rare pattern, these phenomena have to be identified, a test set has to be constructed, we would need to make sure that it does not overlap with the training data, and ideally – to what degree the various aspects of these phenomena are supported by training evidence. This is a big part of work that would go into designing a training dataset.

2.2.4 Not what we set out to do!

The history of AI could be viewed as a trajectory towards decreased amount of implicitly injected knowledge. The early AI systems were fully driven by carefully constructed rules and ontoloties. They were replaced by the statistical approaches, relying on heavy feature engineering. The great promise of DL was to stop trying to define everything, and let the machine to identify and leverage patterns from huge datasets: "we should stop acting as if our goal is to author extremely elegant theories, and instead embrace complexity and make use of the best ally we have: the unreasonable effectiveness of data" (Halevy et al., 2009). And it seems to work: pre-training larger models with more data keeps producing state-of-the-art results (Sun et al., 2017; Brown et al., 2020; Fedus et al., 2021).

Calls for careful construction of datasets are going in the face of that dream. We would arguably be even worse off than when we started: at least in the early AI days we only needed to define the phenomenon to be modeled, and now we also have to find hundreds of examples for that phenomenon.

Defense. Disappointing as it is, we have to admit that although deep-learning-based systems are much better than their predecessors, they are still brittle and do not work well outside the range of cases well represented in the training data (and even there they may work for the wrong reasons). What is more, we are fundamentally no closer to the elusive idea of "understanding" language or its meaningful production (Bender and Koller, 2020). It is true that we were able to "solve" chess and Go without expert knowledge (Sutton, 2019), but these are closed-world games with a known set of rules describing that world. Attempting to do so in the areas that feed from the real social world and impact that world (NLP, facial recognition, algorithmic decision-making on loans etc.) could amplify undesirable patterns present in the big data.

As stated in §2.2.3, it is possible that there is an algorithmic approach that will work equally well or better. Which one will win is an empirical question. As of now, it is fair to say that data curation is at least an alternative to be considered.

This is *not* to say that the current technology cannot yield useful solutions. The achievements are undeniable: the advances in machine translation, question answering, and dialogue already power better customer service, educate and inform, enable communication and information flow for peo-

ple who could not afford professional translation. There is certainly room for useful research to further improve the current solutions, define new tasks and transfer to new domains and languages, even if no fundamental breakthroughs come any time soon. The question is only whether we want to be able to tell in what circumstances our models can be used safely (Mitchell et al., 2019). If so, that would require more thinking about data.

2.2.5 Perfection is not possible.

As mentioned in §2.1.3, the distribution of language phenomena tends to be Zipfian (Zipf, 1945), which means that most phenomena are rare and difficult to learn. A perfect dataset would provides a strong signal for each phenomenon that should be learned. That's not how language works, so we may never be able to create something like that. Balanced datasets are an improvement, but not a solution (Wang et al., 2019b; Rogers et al., 2020a).

Defense. The impossibility of perfection does not entail the impossibility of improvement. For example, a sentiment analysis system that performs as well as the current systems *while* handling negation and coreference correctly, and not pre-judging football fans as violent, is a doable next goal.

2.2.6 No single correct answer.

Curation means making conscious choices about what to include and what to exclude. These are essentially choices about *designing a world*. What linguistic patterns, what concepts, what demographic attributes, what values should that world encode? This is a daunting question, requiring a lot of interdisciplinary expertise and impossible to casually address within a small NLP application project. Neither social sciences nor linguistics offer a ready set of answers, only things to consider in various contexts. The discriminated sub-groups, their values, and underlying social constructs may also differ across communities: e.g. both in India and US there is discrimination based on skin tone, but in the US context it stands for race, and in India it is a proxy for ethnicity, caste and class (Sambasivan et al., 2021a).

Defense. This is an entirely valid point, but it is an objection not to data curation *per se*, but to "data curation in a way that would inflict one set of values and linguistic choices on everyone". That is indeed to be avoided at all costs, and there is a real danger of that happening when NLP systems are commercially deployed and widely used, but the data choices behind them are not explicit.

The position advocated in this paper, as well as by Bender et al. (2021), is only that whatever categories and demographics went into the data design, they have to be documented (Bender and Friedman, 2018; Gebru et al., 2020) and made explicit, so that the users could be informed about what is happening (Mitchell et al., 2019). Some studies will just use convenience samples, and some will intentionally try to create a representation of a world without racial prejudice or rich with island effects. There are valid use cases for both, as long as it is clear who/what is being represented and for what purposes. The tide seems to be turning in this direction: since this work was submitted for review, at least two papers came out documenting popular resources for pre-training language models (Dodge et al., 2021; Bandy and Vincent, 2021). The popular HuggingFace NLP dataset library[4] is also working towards data cards for its resources.

Documenting the choices made in the dataset design is prerequisite to model cards (Mitchell et al., 2019), which could facilitate a healthy interaction between the communities served by the system and the developers of that system. It is entirely possible for that interaction happen in a democratic process: the policies could developed, announced and updated based on the evolving user preferences. Robustness in handling linguistic and social peculiarities of a given community should be a selling point for a product striving to win that community over: something to compete for and showcase, rather than avoid mentioning.

When argument §2.2.6 is made, sometimes it seems to rest on the idea that the distributions in our resources objectively reflect the world. On that view, the calls to data curation would seem opinionated and unnecessary, if not outright dystopian. But the idea that it is possible to work on "NLP in the vacuum", unmarked by linguistic and social categories, is an illusion. A decision to use a convenience sample is also a choice, an act of curation. Using any data to derive research conclusions or in commercial applications is only safe if we know what/who it represents.

3 Why Curation Is Inevitable

In cognitive and sociolinguistics, one of the methods of studying the linguistic and conceptual reper-

[4] https://huggingface.co/docs/datasets/

toire of a certain individual or a demographic is through collecting a representative corpus of their speech (synchronic or diachronic). That corpus inevitably reflects a particular world view[5]. The differences in these world views are expressed as variation in what kinds of linguistic structures people are likely to use, what they are likely to talk about, what are their presuppositions and social context and stereotypes, to what extent any of that is verbally expressed, etc. Some of that variation is idiosyncratic, some attributable to social groups, but even a cursory look at all the variation strongly suggests that there is no "language in general".

It *is* still possible to talk about language at a certain level of abstraction (e.g. "British English" vs the myriad of UK dialects), but only with a good sample representing all the necessary subsets. For example, it would be wrong to construct a "British English" resource based only on London samples, because they do not represent the rest of the country (either linguistically or socioeconomically).

A major achievement of corpus linguists are the "national corpora" such as BNC (Leech, 1992), painstakingly created to represent a diverse sample of written and spoken genres in a certain geographical region in a certain timeframe, so as to enable studies of that specific variety of language. Creating such corpora involves careful sampling, detailed documentation of the domains and speakers that were represented, and much negotiation with publishers for copyright exceptions.

A typical corpus for training language models, or really any NLP dataset, is likewise a sample of speech of a certain group of people, who have their linguistic preferences and sets of values. Consequently, that sample, whether it is coherent or not, and whether it was collected with any specific intentions, represents a certain "picture of the world". Moreover, the purpose of using this data for training is to create a system that would encode that view of the world and make predictions consistent with it. But a typical NLP dataset[6] currently has few specifications of the demographics, dialects, or the range of domains and linguistic phenomena it covers. Unfortunately, it does not mean that the result is some abstract "standard" or "neutral" language. It is some kind of interpolation from the mixture of signals in the data that we have very little idea about.

Why does it matter? The linguistic and conceptual repertoire of humans is dynamic. Our vocabulary, grammar, style, cultural competence change as we go on with our lives, encounter new concepts, forget some things and reinforce others. A key part of that change is the linguistic signals we encounter in communication: on the nativist account children have innate constraints that guide[7] their learning from the data they encounter (Chomsky, 2014; Hornstein and Lightfoot, 1985), and on the usage-based accounts (Bybee, 2006; Lieven and Tomasello, 2008) that process is entirely data-driven. Humans can learn the meaning of words from a single exposure (Carey and Bartlett, 1978; Borovsky et al., 2010), but there is also robust evidence of frequency effects in language acquisition (Ambridge et al., 2015; Diessel and Hilpert, 2016, May 09). It is not by accident that the frequency of the vocabulary to be learned is a key variable in language pedagogy (Zahar et al., 2001).

In short, humans, like DL models, learn from the patterns in the speech that they encounter. And those patterns do not have to come from human speakers anymore: much speech that we will encounter in the future is likely to be synthetic. According to Pilipiszyn (2021), GPT-3 is already generating 4.5B words per day in applications such as question answering, summarization, interactive games, and customer support.

This cannot but have impact back on the human speakers[8] in the following ways:

- An NLP system generating text contributes to a human learner's input in the same way as human writers, and probably also speakers (but potentially on a much larger scale).
- An NLP system that processes human input to answer questions, translate, perform assisting actions etc. has both direct impact (as a lan-

[5]This is a key concept in the works of Neo-Humboldtian scholars: "world image" (Weltbild) of Weisgerber, "naive picture of the world" (naivnaja kartina mira) of Apresyan (1995), and many others.

[6]Corpora generated on crowd worker platforms such as Amazon Mechanical Turk typically impose geographic restrictions, such as "location in US or Canada", but there is no guarantee that the recruited workers are even native speakers.

[7]The "radical" nativist position would be that knowledge of language is entirely innate and is not affected by what the children observe, but on that position we would have to claim the innate knowledge of the word "carburetor" (Knight, 2018).

[8]Synthetic speech will also clearly have impact on the future models if it seeps into the training data. There is research on watermarking generated text (Venugopal et al., 2011; Abdelnabi and Fritz, 2021), but it is not clear what, if anything, the currently deployed systems are doing in this regard. There is at least one documented case of GPT-3 used to post on Reddit as if it were a human user (Philip, 2020).

guage model above), and an indirect impact: as these systems become more widespread, the kind of language that they can and cannot successfully interpret will be respectively reinforced or made less prominent.

- An NLP system that makes decisions in processing applications, grading student work, curating news feeds, summarizing papers and emails, recommending content has the potential of making long-lasting impact on the lives of its users, and the kinds of language that it can process successfully clearly play a role.

The point to take from all of this is that any mismatch of linguistic and social feature distributions between NLP systems and their users *will* have some impact on the world, and for the commercial, widely used NLP systems that impact may be significant. So the debate is not about whether we should change the world by making choices about the data: this is happening either way, because even our convenience samples still reflect numerous implicit choices. The debate is only about how much thinking we want to invest into changing our world.

This thought is somewhat scary (in what way will children growing up with Alexa be different?), but also exciting: the educational opportunities alone could be breathtaking, reaching far beyond the students who are already in a good position to do well in school. We could also create something simplistic, uninspiring, mindlessly entertaining, and/or not-inclusive. That choice is ours.

4 What does it mean to "do NLP"?

To sum up the above discussion: there are no "neutral", one-size-fits-all textual corpora. There is also no manual that would provide foolproof instructions for collecting a "correct" corpus for any given context. And all of these complications are not even the main problem, right? After all, data only serves the task of creating a model, which is the real contribution of an NLP paper?

In theory, the field of NLP is interdisciplinary. In practice, it became something closer to "one of the applied areas of machine learning" rather than "computational linguistics". Furthermore, at least as far as graduate students are concerned, it is something performed as an academic exercise, and as such it does not *really* have to concern itself with its possible effects on the world.

The students can hardly be blamed: keeping up with the latest frameworks and architectures is already hard enough. Most DL practitioners have neither the training nor time to also do the data work at the level that the linguists and ethicists are calling for. The publication system does not provide the right incentives for that either: modeling NLP work is prestigious and welcomed at top conferences, while data work is "janitorial", less well paid, "under-valued and de-glamorised"[9] (Sambasivan et al., 2021b).

It does not help that there seems to be a systematic miscommunication between the fields. When linguists or ethicists talk about the issues with the current solutions, the practitioners may take it as an accusation that they are not doing a good job, rather than as an invitation to improve things together. Likewise, when the practitioners propose new systems, the linguists and ethicists may be frustrated: not by the incremental improvements on leaderboards as such, but by lack of accompanying discussion of what the proposed methods are supposed to do better, and for whom.

If anything is to change, we need to overcome this antagonism. Here are a few suggestions for how that could be achieved.

5 Moving Forward

Step 1. Understand each other better. The fact is, the AI ethics people are not really out to "cancel" everybody. It is easy to see why they would be frustrated that the social justice issues have never been a priority, terrified at what "move fast & break things" has already done with the social world, and dubious that they just need to wait and change would come.

The linguists are not completely useless. Chances are, many problems that the DL engineers are having could be fixed if someone was just around to realize that the tokenizer didn't handle the suffixes well.

And the engineers are not inherently evil. They just need resources, training, collaborators, time, and better research incentives. Instead, they have to churn out papers in 2 months just to stay in the publication race, with no time to dive deeper into what their systems are actually doing.

Step 2. Improve the incentive structure. One way to change the incentive structure that led to

[9] Of course, this perception is not universal, and there are (very few) "unicorn" resources like SQuAD (Rajpurkar et al., 2016) that highly influenced the field. But overall the power balance in the field is currently not in favor of resource work.

the current situation is through conferences. There will be a lot more interest in data work if it becomes more publishable. As of now, the "resources and evaluation" track is something of a poor relative to the "machine learning" track, which in ACL 2020[10] attracted nearly 3 times more submissions. Most task-specific tracks (question answering, summarization, dialogue etc.) are supposed to receive both engineering and data submissions, but in that setting the interdisciplinary tension may lead to resource papers voted down simply for being resource papers (Rogers and Augenstein, 2020). Bawden (2019) cites an ACL 2019 reviewer who complained that "the paper is mostly a description of the corpus and its collection and contains little scientific contribution".

We really need to take the type of contribution[11] into account in reviewer assignment, into review form design, and into reviewer training programs. We also need to make sure that the resource tracks are consistently offered[12], with dedicated best paper awards to raise the prestige of this work in the community. Some conferences already started to provide reviewer mentoring, double down on ethics, consider what signal they send to companies and students by their best paper awards. We can all help by lobbying program chairs whenever we have a chance, offline and online.

A helpful factor is that the ever-increasing size of models is making the state-of-the-art leaderboard chase financially untenable for even well-resourced labs, and they are looking for other outlets. This is a chance for the NLP community to engage more deeply with the phenomena that we are modeling.

Step 3. Educate. The idea that "NLP" means "deep learning" may well arise if it is taught as a one-semester course focusing on the engineering. If the coursework is fully powered by existing resources, it creates the impression that data is not a part of the job. The result is that the students learn that it is entirely possible to just run off-the-shelf parsers without knowing anything about syntax, or do sentiment analysis without knowing anything about pragmatics. And if it is possible to not do more work, why would anyone bother?

We need to provide our students with the skills to stress-test their systems and critically examine their data, so as to be able to spot potential issues early on. For that, they will need the basic linguistic theory, the fundamentals of sociolinguistics and pragmatics. Likewise, some aspects of psychology (dual processing theories, memory and attention span, cognitive biases, "nudging") are a pre-requisite for designing interfaces not only for annotation projects, but for any kind of interactive NLP systems. And some awareness of the social power structures would help in not propagating the harmful stereotypes. Some strategies for building NLP curricula have been discussed at the TeachingNLP workshop (Radev and Brew, 2002; Brew and Radev, 2005; Palmer et al., 2008; Derzhanski and Radev, 2013; Jurgens et al., 2021).

Most importantly, NLP courses need to combat the idea that all the knowledge about the human world is just irrelevant in the age of big data and DL. The "garbage in, garbage out" principle is still relevant. We may be able to sort the garbage and learn from it anyway, but only if we have at least some idea about what kind of garbage we have.

Step 4. Collaborate. Large companies and universities provide a significant competitive edge to their authors just in virtue of the in-house collaboration networks they could offer. But it is becoming increasingly easy for everyone to find external collaborations, especially in the world in pandemic lockdown. One opportunity is Twitter, used by estimated 40% of EMNLP 2020 authors[13].

What would it mean to "collaborate"? At the bare minimum, in an engineering project the linguists and social scientists could help to at least try to characterize the data that was used with something like data statements (Bender and Friedman, 2018; Gebru et al., 2020). A more ambitious goal would be to involve them early on in the data selection, preparation, and iterative development. Ideally, there would be joint formulation of research goals, thinking together about what kind of world we are building.

Finding collaborators is much easier for established researchers, not only because they are a known quantity, but also because they are already aware of what could be done in an interdisciplinary project. They probably even already know the people who they could ask to join. But the students could use some help, especially those from the less well-connected institutions. They could bene-

[10] https://www.aclweb.org/adminwiki/images/9/90/ACL_Program_Co-Chairs_Report_July_2020.pdf

[11] As was done e.g. at COLING 2018: http://coling2018.org/paper-types/

[12] E.g. this track was recently absent at EMNLP 2020.

[13] Source: EMNLP 2020 organizers.

fit from establishing some kind of skill exchange network, where the students with engineering background could help out in data projects and students with linguistics/social science background could help out in engineering projects. This would probably the best way to ease the interdisciplinary tension, instill respect for each other's expertise, as well as the awareness that NLP is a huge problem that we do not even understand that well, and for which we need all the help we can get.

Step 5. Estimate. The goal of all the above data work is ultimately to enable informed decisions by the public, the CEOs, and the policy makers about what kind of world we would live in. One takeaway from the heated debate around (Bender et al., 2021) is that if one side in an interdisciplinary debate focuses mostly on the potential benefits of something, and the other mostly on its harms, the stance is likely to become adversarial, and we do not give each other the benefit of the doubt[14].

Nevertheless, the people on both sides of the debate are researchers, and they want to make informed decisions. That is only possible through cost-benefit analysis. It is clear that the first step has to be thorough documentation of the data (Bender and Friedman, 2018; Gebru et al., 2020): this lets us compare the represented population and the population of the target users, and think through the possible harms. However, it is not clear how to weigh the harms against the benefits.

At the very least, to make informed decisions we would probably need to know the following[15]:

- Which population will get exposed to the proposed tech?
- What are the direct and indirect benefits on the user population?
- What are the direct and indirect harms on the population in general (not limited to the users of the proposed tech), in particular the marginalized groups?
- If certain harms are inflicted on the user population, would they have the political/legal recourse to be compensated?
- How compute-efficient the implementation would be, how would the energy be sourced, and would that affect any other populations?
- How widely would it be eventually adopted, and how that changes the likelihood of benefits and harms to different user groups?
- What is the potential for further innovation that would significantly change the appeal, deployability or risks of the proposed solution?
- What are the risks of human error and deliberate misuse if the tech is stolen/replicated by terrorists, authoritarian governments, propaganda organizations and other bad actors?

Unfortunately, the world is volatile and business plans change all the time. There is so much uncertainty for each of these points that it is not clear how to even start. Yet we have to try to come up with a process for working these things out, and eventually develop templates and calculators that developers could use to make estimates for best-, worst- and realistic scenarios.

This is an area in which NLP is desperately in need of collaboration with economics, governance and law. In that, again, NLP conferences could take the lead. There could be regular tracks that would incentivize joint publications with experts from these fields. The search for solutions is already going on, but this way NLP community would participate in it rather than just meet with regulation post-factum. To be able to provide meaningful peer review for such work, we would need a mechanism of recruiting external reviewers with the required expertise on as-need basis.

6 Conclusion

Our data is already changing the world, and will keep doing so whether we are being intentional about it or not. We might as well at least try: we do want more robust and linguistically capable models, and we do want models that do not leak sensitive data or propagate harmful stereotypes.

Whether those goals would be ultimately achieved by curating large corpora or by more algorithmic solutions, in both cases we need to do a lot more data work. The current dynamic suggests that this won't happen, unless we overcome the interdisciplinary tensions and turn our conferences into truly shared spaces.

7 Acknowledgements

Many thanks to Emily M. Bender, Yoav Goldberg, Ryan Cotterell, and the anonymous reviewers for their thoughtful comments on this paper.

[14] https://twitter.com/nlpnoah/status/1354814467633111048

[15] Many of these points are made in the NAACL ethics FAQ https://2021.aclweb.org/ethics/Ethics-FAQ/

References

Sahar Abdelnabi and Mario Fritz. 2021. Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding. *arXiv:2009.03015 [cs]*.

Ben Ambridge, Evan Kidd, Caroline F. Rowland, and Anna L. Theakston. 2015. The ubiquity of frequency effects in first language acquisition*. *Journal of Child Language*, 42(2):239–273.

Yu.D. Apresyan. 1995. *Izbrannyje Trudy*, volume 2. Yazyki russkoj kultury, Moscow.

Jack Bandy and Nicholas Vincent. 2021. Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus. *arXiv:2105.05241 [cs]*.

Rachel Bawden. 2019. One paper, nine reviews. *Rachel Bawden's blog*.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Arielle Borovsky, Marta Kutas, and Jeff Elman. 2010. Learning to use words: Event-related potentials index single-shot contextual word learning. *Cognition*, 116(2):289–296.

Chris Brew and Dragomir Radev, editors. 2005. *Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching NLP and CL*. Association for Computational Linguistics, Ann Arbor, Michigan.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Jacob Buckman. 2021. Fair ML Tools Require Problematic ML Models. *Jacob Buckman*.

Joan Bybee. 2006. From Usage to Grammar: The Mind's Response to Repetition. *Language*, 82(4):pp.711–733.

Susan Carey and Elsa Bartlett. 1978. Acquiring a Single New Word. Technical report, Stanford University, Dept. of Linguistics.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting Training Data from Large Language Models. *arXiv:2012.07805 [cs]*.

Noam Chomsky. 2014. *Aspects of the Theory of Syntax*. MIT Press.

Jeffrey Dastin Dave, Paresh. 2021. Google pledges changes to research oversight after internal revolt. *Reuters*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Ivan Derzhanski and Dragomir Radev, editors. 2013. *Proceedings of the Fourth Workshop on Teaching NLP and CL*. Association for Computational Linguistics, Sofia, Bulgaria.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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 4171–4186.

Holger Diessel and Martin Hilpert. 2016, May 09. Frequency effects in grammar. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, and Matt Gardner. 2021. Documenting the English Colossal Clean Crawled Corpus. *arXiv:2104.08758 [cs]*.

Ali Emami, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. An Analysis of Dataset Overlap on Winograd-Style Tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5855–5865, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv:2101.03961 [cs]*.

Matt Gardner, William Merrill, Jesse Dodge, Matthew E. Peters, Alexis Ross, Sameer Singh, and Noah Smith. 2021. Competency Problems: On Finding and Removing Artifacts in Language Data. *arXiv:2104.08646 [cs]*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. *arXiv:1803.09010 [cs]*.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2019. Posing Fair Generalization Tasks for Natural Language Inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4475–4485, Hong Kong, China. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Yoav Goldberg. 2021. A Criticism of Stochastic Parrots. *Github Gist*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

A. Halevy, P. Norvig, and F. Pereira. 2009. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, 24(2):8–12.

Norbert Hornstein and David Lightfoot. 1985. Explanation in Linguistics. The Logical Problem of Language Acquisition. *Tijdschrift Voor Filosofie*, 47(2):338–338.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. Association for Computational Linguistics.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 306–316, New York, NY, USA. Association for Computing Machinery.

David Jurgens, Varada Kolhatkar, Lucy Li, Margot Mieskes, and Ted Pedersen, editors. 2021. *Proceedings of the Fifth Workshop on Teaching NLP*. Association for Computational Linguistics, Online.

Chris Knight. 2018. According to Chomsky, words such as 'book' and 'carburetor' are genetically determined. *Science and Revolution: Chris Knight's Blog on Noam Chomsky*.

Geoffrey Neil Leech. 1992. 100 million words of English: The British National Corpus (BNC). *Language Research*, 1/4.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Elena Lieven and Michael Tomasello. 2008. Children's first language acquistion from a usage-based perspective. In *Handbook of Cognitive Linguistics and Second Language Acquisition*, pages 168–196. Routledge/Taylor & Francis Group, New York, NY, US.

Tal Linzen. 2020. How Can We Accelerate Progress Towards Human-like Linguistic Generalization? *arXiv:2005.00955 [cs]*.

Michael Lissack. 2021. The Slodderwetenschap (Sloppy Science) of Stochastic Parrots. *Medium*.

R. Thomas McCoy, Erin Grant, Paul Smolensky, Thomas L. Griffiths, and Tal Linzen. 2020. Universal linguistic inductive biases via meta-learning. *arXiv:2006.16324 [cs]*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA. Association for Computing Machinery.

Martha Palmer, Chris Brew, and Fei Xia, editors. 2008. *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*. Association for Computational Linguistics, Columbus, Ohio.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *arXiv:2012.05345 [cs]*.

Philip. 2020. GPT-3 Bot Posed as a Human on AskReddit for a Week. *kmeme*.

Ashley Pilipiszyn. 2021. GPT-3 Powers the Next Generation of Apps. *OpenAI*.

Dragomir Radev and Chris Brew. 2002. *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Anna Rogers and Isabelle Augenstein. 2020. What Can We Do to Improve Peer Review in NLP? In *Findings of EMNLP*, pages 1256–1262, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020a. Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8722–8731.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020b. A Primer in BERTology: What we know about how BERT works. *(accepted to TACL)*.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021a. Re-imagining Algorithmic Fairness in India and Beyond. *arXiv:2101.09995 [cs]*.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021b. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–15, New York, NY, USA. Association for Computing Machinery.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *arXiv:2103.00453 [cs]*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 843–852.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Rich Sutton. 2019. The Bitter Lesson. *Incomplete Ideas*.

Suresh Venkatasubramanian. 2021. On stochastic parrots. *Algorithmic Fairness*.

Ashish Venugopal, Jakob Uszkoreit, David Talbot, Franz Och, and Juri Ganitkevitch. 2011. Watermarking the Outputs of Structured Prediction with an application in Statistical Machine Translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1363–1372, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. In

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *arXiv:1905.00537 [cs]*.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019b. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. In *ICCV 2019*.

Leo Weisgerber. 1953. *Vom Weltbild der deutschen Sprache*. Pädagogischer Verlag Schwann.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: The Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Rick Zahar, Tom Cobb, and Nina Spada. 2001. Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review*, 57(4):541–572.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R. Bowman. 2020. When Do You Need Billions of Words of Pretraining Data? *arXiv:2011.04946 [cs]*.

George Kingsley Zipf. 1945. The meaning-frequency relationship of words. *The Journal of General Psychology*, 33(2):251–256.