Dr. John W. Jones

SUNY Cortland

Cortland, NY

John.jones02@cortland.edu

AI, Language, and Education: Bias and Prejudice in Text and Corpus Analysis

Various linguistic disciplines make use of corpora (singular, corpus), which are “bodies” or collections of texts. For example, a corpus in literary studies may refer to the collected works of single author, and in linguistic anthropological fieldwork it may refer to the collection of data for linguistic research. In corpus linguistics, linguistic corpora are large collections of digitized texts, which are sampled to ensure authenticity and representativeness of a particular language variety. To say that the sample texts collected in a corpus are “authentic” means that they are examples of language that are produced for the purposes of actual communication and not artificial constructions produced by researchers. Corpora most often contain textual—or alphanumeric—data but can also consist of audio or video files. Data in corpora are often annotated with comments by researchers and tagged to highlight parts of speech, intonation, font style, syntactic formations, etc.

Although corpora have been used for a long time, Corpus Linguistics—the branch of linguistics whose primary research method is the use and analysis of data from corpora—has only recently regained popularity after the mid 20th century dominance of linguistic approaches that focused more on native speaker intuition and introspection of sentences to gauge grammaticality and ascertain the deep structural rules of human language. Corpus linguistics research really became feasible and popular after the widespread availability of personal computers. Software corpora made the searching for and comparing textual examples less time-consuming. The use of corpora has made possible new discoveries about how human languages work. One way this is possible is how they can display different instances of the same word or construction that appears in different texts. The samples are arranged in a grid-like fashion displaying the word or construction along with the context in which it is embedded. This allows researchers to compare different usages of the same word, and to make observations about the context in which a combination or word is used and the frequency of the combination or word.

Linguistic Corpora are also a key development that makes possible powerful new AI utilities, particularly those that generate human language, that rely on Large Language Models (LLMs) which are in turn “trained” on these large corpora. The training consists of analyzing large corpora of texts which can either be pre-existing corpora created for other purposes, corpora created for training purposes, or corpora gathered from textual activity on the world wide web. Related to AI and LLMs, Natural Language Processing (NLP) is heavily dependent on corpus linguistics and corpora. NLP can variously be defined as a discipline, set of methods, or even a type of AI technology. The ambiguity in its definition is a result of its overlap with many fields such as linguistics, AI, computer science, and speech processing. The goal of NLP is to analyze large datasets like corpora using a variety of statistical and probabilistic methods. Corpus linguists use various NLP tools such as syntactic parsers and Part-Of-Speech taggers (POS) to quickly analyze corpora and test theories.

Related to their centrality to Artificial Intelligence (AI) applications, is concern regarding bias which affects both AI and Linguistic Corpora. Being based on human behavior, which is naturally in large part an outgrowth of human beliefs, prejudices, ideologies, and interests, these corpora can contain traces of real-world bias which are then passed on to any AI application which may rely on them.

The problem of bias in computer systems more broadly was identified at least as early as the mid-1980s. Friedman and Nissenbaum (1996) propose a framework consisting of three categories of bias in computer systems: preexisting, technical, and emergent. Pre-existing bias comes from social institutions, practices, and attitudes, such as de facto or de jure systems of racial, ethnic, or gender bias. Technical biases are a result of technical limitations. An example of a technical limitation could be a low-resolution computer display that leads to "ranking bias" because the screen can only display a small number of search results at a time, therefore necessitating the use of another screen in order to see more results at once, or the use of a "next page" function to display a different page of results. Different ranking schemes, whether alphabetical or otherwise, will lead to people whose names are displayed earlier being seen (and maybe chosen) more often. Emergent biases can happen if technological systems are designed to reflect for a particular population or for a particular set of social values. The use of the system within different populations or changes to social values can lead to mismatches between the design of the system and the populations or social values.

These types of bias can affect software systems, algorithms, datasets used to "train" AI, and other aspects of computer technology. The biases that affect algorithms can be categorized in ways that resemble Friedman and Nissenbaum's (1996) framework. Biases in data that result from preexisting institutional and individual biases in measurement will affect downstream machine learning applications that rely on these datasets. A technological bias exists in that the very goal of algorithms is to minimize the overall prediction errors. Therefore, these systems focus more on data from categories with majority representation, since the behavior and experiences of these categories give the best probability of identifying statistically "normal" behavior. This can lead to a bias in favor of majority groups over minorities (Pessach and Shmueli 2021). A third type of bias occurs when legitimate criteria for the decision-making that were not consciously utilized for the purpose of disadvantaging any group, serve as proxies to sensitive attributes by their correlation or co-occurrence with these other attributes. In these cases, the bias could still happen even if the dataset does not directly include sensitive attributes. Realizing the importance of this issue, courts, institutions, and states have begun to take preliminary steps to mitigate any harm that might come from biased AI and machine learning (ML) systems. The European Commission's 2021 Proposal for Regulation Laying Down Harmonized Rules on AI (AI Act) would require that "high-risk AI systems" should be trained on models that use "high quality" data sets which are relevant to the populations in which the AI systems are deployed, and that the statistical properties of the data sets are themselves properly representative of the people in the populations that might be affected by the AI systems. This proposal thus recognizes the real harm that can be done by the implementation of AI systems that rely on statistical data which is itself conditioned by biases and assumptions at the point of data collection.

A famous case of this kind of algorithmic bias is the 'criminal risk assessment' tool called COMPAS, or 'Correctional Offender Management Profiling for Alternative Sanctions. COMPAS is used by courts of law when determining sentencing and 'predicts' the chance of recidivism, or the probability that an offender will reoffend, using 137 'features', such as employment status or income level. Although race is not one of features used to create its predictions, the program consistently judged Black offenders as higher risks of recidivism than Whites. Angwin et al. (2016) analyzed the COMPAS results of more than 7000 individuals arrested in Broward County, Florida between 2013 and 2014. They determined that COMPAS's overall accuracy for White defendants was only slightly higher than its accuracy for Black defendants. However, the systems results did not affect Black and White defendants in the same way, and COMPAS incorrectly predicted that Black defendants who actually did not go on to reoffend would do so at a rate nearly twice as high as their White counterpart, while White defendants who did reoffend were incorrectly predicted to not commit further crimes at a rate nearly twice as high as their Black counterparts. While racial and ethnic identify were not among the features used to make predictions, other features that in the datasets used by COMPAS were statistically correlated with higher risks of recidivism may be serving as "proxies" for these categories, and therefore may have implicitly indexed race and ethnicity.

Language Varieties, Corpora, and Bias

Similar issues regarding fairness affect the large linguistic corpora upon which a utility like OpenAI's ChatGPT rely. Although linguists in general understand the deep effects social systems have upon language, and effect even something as seemingly basic as delineating a "language" from a "dialect—a difficulty enshrined in the phrase "a language is a dialect with an army and a navy". Much of what demarcates a language from a dialect depends largely on non-linguistic factors. Linguists who both create and work with large corpora may still succumb to biases of the institutional, technological, and emergent varieties. For example, in a study which examined computational methods for deriving word meanings from co-occurrence (the appearance of words in proximity in a text) statistics drawn from linguistic corpora like the British National Corpus (BNC), the researchers noted that "corpus quality" could affect semantic representations produced by their computational methods (Bullinaria and Levy 2007). The BNC draws textual examples from a variety of sources, so "corpus quality" could be affected by examples drawn from a source in which word frequency distributions deviate from those of average texts, or by examples drawn from nonstandard English (the authors use the example "picture window" vs. "pitcher winder"). The authors of the study probably did not mean to disparage the speakers of nonstandard English varieties, but they identify a very important connection between linguistic variety, the aims of technologies that rely on corpora, and bias. For many applications, linguistic variety is not simply a theoretical curiosity, it is a reality that may impede the intended functioning of the technology and therefore must be accounted for, as an error, an exception, or special case. As in the case of AI, the presence of the minority, in this case nonstandard English varieties, may vitiate the goals of the technology.

Bias against minority language use can have profound effects on the lives of the speakers of such languages. The use of minority languages can be highly political and emotionally charged. In France, where the prevailing ideology since the French Revolution has promoted the unity of identity of all French citizens, the use of minority languages like Occitan has historically been attacked in institutions like schools and in the public. In the USA, the nonstandard English variety, African American Vernacular English (AAVE), caused a national controversy in the 1990s when the Oakland Unified School District (OUSD) in Oakland, California, followed the recommendations of a task force convened to find ways to overcome the disparities in educational attainment between Black and White students. The task force decided that recognition and understanding of AAVE would enhance Black students' learning of Standard American English (SAE). The school district's decision was based on consultations with linguists who explained the benefits of using students' nonstandard home languages to help those students to learn SAE.

The failure to categorize speech as a legitimate, though non-standard, variety can lead those who either compile or analyze the linguistic data in large corpora to designate such a variety as an error or a "low quality" sample. The failure to acknowledge these speech varieties as legitimate can lead to institutional biases, manifesting the categorization of such varieties as errors or low-quality samples with a possibility that there may not be enough examples or records of these nonstandard varieties in existence in the first place. A lack of political recognition and social status prevents many of these varieties from being studied or recorded. Another related problem is that there simply may be a lack of speakers of some varieties, making the compilation of examples from these languages difficult.

Variation and change are important topics for corpus linguistics, and both indicate an area where bias may arise. Linguists have long realized that languages can vary in a number of ways. Generally, linguists accept four dimensions of linguistic variation: diaphasic, diastratic, diachronic, and diatopic. Diaphasic variation depends on changes to the context in which linguistic communication occurs and can affect choices such as style of speech or register. Diastratic variation occurs during communication between different social groups, like ethnic or age groups, for example. Diachronic variations happen when languages change over time, and diatopic variations happen when languages vary from place to place. Distinguishing between language varieties and the different dimensions of variation within the same language is a technological problem for the field of Natural Language Processing (NLP) and algorithms have problems correctly identifying languages from within the same family, like Italian and Portuguese, for example. Potential solutions to these difficulties exist by leveraging the insights of researchers who focus more on the sociolinguistic aspects of language variation.

There are other technical problems presented by nonstandard varieties for the compiling and usage of large corpora. One problem is the accuracy of transcription of the spoken language. The variability of pronunciation of spoken language is inherently highly unpredictable, and this may be even more true of nonstandard language varieties. If systems are produced which narrowly define the range of sounds, the authenticity and accuracy of analysis may be increased but the number of successful searches from within a corpus may be limited.

Corpora, Education, and Bias

Corpus linguistics has increasingly been recognized as a strategy for language education. The use of corpora can be beneficial for helping second language learners acquire a new language, but speakers of nonstandard varieties and also native speakers of standard varieties can benefit from instruction using corpora. Languages can be thought of as rule-based systems that follow some kind of internal “logic”, but they can also be thought of as communicative behavior in which word choices are statistically constrained by factors such as previous words in a string, social context, speaker desire, and more. By this second view, easy access to information about which words are more statistically “acceptable” within a structure can help language learners to develop more natural-sounding speech, or to better understand grammatical norms and conventions. The necessity of sensitivity to possible unfairness exists in this new field as well. This is especially true for speakers of nonstandard varieties. If sufficient examples of the nonstandard variety have not been collected and no corpora exist, helpful comparisons between this variety and the standard are almost impossible. Text-to-speech programs that cannot account for the sound system of the nonstandard learner will cause confusion for words that have identical meanings across varieties but are pronounced differently (“that” vs. “dat” in SAE). If grading criteria are applied incorrectly to tasks—for example, if a student is graded on pronunciation when the specific task involved learning vocabulary or word order—students who are in fact performing at a standard level can be punished unfairly.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks. ProPublica, 23 May 2016; https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. *Behavior Research Methods, 39*(3), pp. 510-526.

European Commission. (2021). *Laying Down Harmonised [sic] Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206> (last visited January 14, 2024).

Friedman, B. and Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems, 14*(3), pp. 330 –347.