# Humanistic NLP: Bridging the Gap Between Digital Humanities and Natural Language Processing

**Tasovac, Toma**
ttasovac@humanistika.org
Belgrade Center for Digital Humanities

**Ermolaev, Natalia**
nataliae@princeton.edu
Center for Digital Humanities at Princeton

**Janco, Andrew**
apjanco@upenn.edu
University of Pennsylvania

**Lassner, David**
lassner@tu-berlin.de
Technische Universität Berlin

**Budak, Nick**
budak@stanford.edu
Stanford University

There was a time — in the sixties and seventies, during the early days of humanities computing and the heyday of Chomskian linguistics — where natural language processing (NLP) more or less made sense to your average humanities scholar. This was because NLP was to a large extent rule-based. Computer scientists and linguists built complex systems on top of explicit rules (and exceptions) that would, in theory at least, cover all the grammatical and syntactical structures of a natural language. The dominant ap-proach in NLP these days has nothing to do with explicit rules; instead, it is based on statistical models. Statistical NLP infers ru-les from existing texts and annotations by converting words into vectors in a multidimensional space. These — to the human mind— impenetrable models are used to make predictions on new data. This works fairly well for certain types of tasks. But they also feel a bit like magic (Tasovac and Ermolaev 2022).

Most humanists come to quantitative, statistical and machine-learning methods with a healthy dose of skepticism to begin with: we are trained to recognize that context is everything, that meaning is always irreducibly complex, that texts are often inherently contradictory, and that there is no such thing as ideology-free space. So it should come as no surprise that humanists are especially sensitive to the challenges of power dynamics, data availability, and domain specificity, as well as structural and representational bias in statistical language models based on large datasets (see Bender et al 2021; Bamman et al. 2019; del Rio Riande 2022; Klein 2022)

In this paper, we present Humanistic NLP as an interdisciplinary framework which aims to articulate the multiple challenges facing present-day interactions between NLP and humanities research, with the goal of finding solutions to bridge this gap. We will first analyze the epistemological, conceptual, technical and socio-cultural obstacles that stand in the way of a wider use of NLP methods in Digital Humanities: the fundamental clash of positivist and empirical traditions in linguistics with the hermeneutic traditions typical for the humanities; the differences in rule-based and statistical approaches to NLP and how they relate to humanistic perspectives on language, structure and meaning; as well as the relationship between textual formats (plain text, XML), tools that can process them and the kinds of questions that can be asked about such texts, while paying special attention to the difficulties presented by the black-box machine-learning models (i.e. models which are created directly from data by an algorithm) to the humanistic production of knowledge.

Taking into account the significant contributions of other research projects in this domain (Bamman 2020; Bamman 2022; Wilken et al. 2022; Mcgillivray et al. 2020; Underwood 2019) we'll introduce the concept of Humanistic NLP as an area of applied, translational research aimed at the development of theories, tools, and processes that enable the use of NLP frameworks by humanist scholars in specific use cases. We'll argue that to support humanists working on domain-specific (for instance, literary) datasets and/or language varieties (historical or dialectal) which have no robust NLP support, we need to 1) raise their awareness about both the potential and the pitfalls of machine-learning in NLP; and 2) teach them very specific skills and workflows so that they can independently collect and annotate data, train and evaluate language models and experiment with transfer learning in their own work.

We tested this approach in a series of workshops under the banner of the NEH Institute for Advanced Topics in the Digital Humanities "New Languages for NLP" (Janco and Ermolaev 2022). Between the summer of 2021 and the spring of 2022, we worked with 10 language teams (Classical Arabic, Old Chinese, Kanbun, Kannada, Ottoman Turkish, Quechua, Russian literature of the 19th century, Tigrinya, Yiddish and Yoruba) to help them create language models for lesser-resourced and/or domain-specific languages using the spaCy NLP framework (spaCy 2022).

In our paper, we will analyze the pedagogical approaches we adopted and practical solutions we implemented in order to make our students' learning process as effective as possible. We will discuss the major challenges we have encountered, including text acquisition and corpus compilation, OCR quality, data curation, unreliability of linguistic terminology (such as "sentences" and "paragraphs") as well as morphological complexity, all of which have impacted the nature and the success of the teams' language models. We will also address the importance of team building in such interdisciplinary enterprises, where individual disciplinary expertise needs to be matched with social and communication skills especially when it comes to managing and supporting annotators.

Humanistic NLP shouldn't be thought of as merely an application of NLP tools in the humanistic domains, but as an opportunity to engage more humanist scholars in the development and critical appraisal of NLP language resources. Substantive discussions across the disciplinary boundaries of the humanities and data and computer science may at times feel insurmountable. But, in the spirit of this year's topic of DH2023 — collaboration as opportunity *and* revolutions — we hope to contribute with some ideas on how to communicate and learn from each other in the age of intense academic overspecialization in order to prevent our segregated disciplines from turning us into methodological, epistemological and ideological loners.

# Bibliography

**Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S.** (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* (pp. 610-623).

**Bamman, D., Popat S., and Shen S.** (2019) An annotated dataset of literary entities. 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 2138–2144, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1220. URL https://www.aclweb.org/anthology/N19-1220.

**Bamman, D.** (2020) "LitBank: Born-Literary Natural Language Processing." In *Debates in the Digital Humanities: Computational Humanities*, edited by Jessica Marie Johnson et al.

**Bamman, D.** (2022) Book NLP https://github.com/dbamman/book-nlp

**del Rio Riande, Gimena.** (2022). On Spanish-Speaking Parrots. *Startwords*, 3. https://doi.org/10.5281/zenodo.6567850

**Gebru, T., Morgenstern J., Vecchione B., Wortman Vaughan, J., Wallach H., Daumeé III, H., and Crawford, K.** (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010

Janco, A. and Ermolaev, N. (2022) New Languages for NLP. https://newnlp.princeton.edu/

**Klein, L.** (2022) "Are Large Language Models Our Limit Case?" *Startwords*, 3. https://doi.org/10.5281/zenodo.6567985

**Jo, E.S. and Gebru T.** (2020) "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning." In *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 306–316. doi: 10.1145/3351095.3372829.

**Mcgillivray, B., Poibeau, T. and Ruiz, P.** (2020) "Digital Humanities and Natural Language Processing: 'Je t'aime... Moi non plus.'" *Digital Humanities Quarterly*, 14(2).

**spaCy** (2022) spaCy: Industrial-Strength Natural Language Processing in Python. https://spacy.io/

**Tasovac, T. and Ermolaev, N.** (2022) "Introduction." In *Startwords 3.* https://startwords.cdh.princeton.edu/issues/3/

**Underwood, T.** (2019) "Do humanists need BERT? Neural models have set a new standard for language understanding. Can they also help us reason about history?" https://tedunderwood.com/2019/07/15/do-humanists-need-bert/

**Wilkens, M., Mimno, D., Walsh, M., Thalken, R.** (2022) BERT for Humanists Project. http://www.bertforhumanists.org/