# Critique of Pure Digital Reason:

Validating Distant Reading from a Comparative Perspective

## Abstract

Since the beginning of the current century, the field of digital humanities has been described as an innovative and promising branch that will change the field of the humanities, pave new ways for understanding human society, and build a strong bridge between different disciplines. This development, which goes hand in hand with other cultural, technological and academic developments, raises many hopes, but also critical responses: does it not entail a dangerous renunciation of the 'good old' humanistic tradition? Can computers 'understand' phenomena as humans understand them? What is really the value of 'distant reading', as it is called following Moretti? Twenty years after the great and optimistic breakthrough in the field it can be said that these questions do not stem from mere conservatism. Although almost no one questions the analytical power of the computer anymore, or the potential contribution of computationality to the field of the humanities, it seems that this potential is still a long way from being realized. Various scholars point to problems in the research conducted in the digital humanities. The achievements already made in the field also suffer from what Adam Hammond called "the double bind of validation," i.e. the fact that on the one hand many computational studies prove things we already knew in advance, without any advanced aids, and on the other hand, more groundbreaking computational studies produce such surprising results that we have no ability to confirm or deny them.

The project proposed here follows the call of Hammond, Underwood, Meister and other leading researchers for the *validation* of the digital humanities, that is, the beginning of a new and more mature era in the field, where we will not only speak in the future-tense about what could be done theoretically, but in the present tense, about what is actually being done: an era in which we put to a balanced and careful critical test the successes against the failures, and develop tools for more effective, fruitful and meaningful research progress.

It should be emphasized that the Israeli context gives this call a unique opportunity: distant reading studies on Hebrew texts are still in their infancy (among other things, due to the difficulty of adapting computational analysis tools to the Hebrew language), and the willingness of Israeli academia to try the new directions is at its peak. Therefore, if we learn from the experience of others - and especially, if we properly address the question of validation - we can establish a good and stable starting point for further

development of the field in its local context, while offering an internationally unique test case. It is difficult to think of a more necessary move at this time for the establishment of digital humanities from here on.

To this end, we intend to examine, from a comparative and critical standpoint, conventional practices of distant reading in three Hebrew and Israeli corpora representing three different disciplines - prose, law and poetry – with each having its own unique characteristics and challenges. In the field of prose (Ben Gurion University team), we will find out how computers model modern and post-modern short stories, what phenomena they emphasize, how that differs from the interpretive process which literary readers experience when reading those stories and giving them meaning, and the extent to which a connection can be made between the two modes of reading, not just describing them as competing or contradictory perspectives. In the field of law (Hebrew University team), we will seek to examine how algorithms read and interpret legal evidence on two different narratological levels. The first level concerns the identification of the thematic content that testimony deals with. The second concerns the identification of the narrative sequence and plot events narrated in legal testimony. In law, unlike literature, testimony is not merely a narrative act; it serves as a factual basis for the entire legal process. As such, the legal field poses an extraordinary challenge to existing distant reading algorithms. In the field of poetry (Open University team), we will focus on the identification of poetic models by humans and by algorithm. Poetry, by its nature, is rich in structural characteristics (such as meter, rhyme, strophic division) and therefore repeating patterns are very dominant in it and in its research. The initial phase of the study will be based on manual tagging of poetic and prosodic elements of the corpus of liturgical poetry (Piyyut) and then distant reading of the tags in order to identify patterns. Human distant reading will be done on the basis of smart visualizations and the computational distant reading will be conducted based on algorithms. At the end of the process, human reading will be compared to computer reading in an attempt to produce a computational model for identifying poetic models.

The three corpora will be analyzed using one common strategy in two main phases: **(a)** the critical cross-examination of human reading with computational reading in its various iterations, and **(b)** the development of a conceptual framework and computational toolkit for informed and balanced validation of distant reading, including NLP models, visualization and additional tools which are necessary for the analysis of the Hebrew text. The deep and continuous synergy between the three teams, and between each of them and their respective partners in computer and data sciences, will give the project a significant advantage, because it will make the validation differential - that is, sensitive to different types of texts and textual problems in each of the fields. Also, from an integrative perspective of the research and its application to the research community and the general public, the research team will maintain a blog throughout the project years, describing each step in real time - its successes and challenges - and will enable a better understanding of the analytical process of computational textual analysis in the digital age.