DH 2015:
Multiple-paper session proposal

Panel title: New developments in quantitative metrics

Papers:

"Enabling the automated identification and analysis of meter and rhyme in Russian verse" (David J. Birnbaum, University of Pittsburgh and Elise Thorsen, University of Pittsburgh)

"Making visible the invisible: metrical patterns, contrafacture and compilation in a Medieval Castilian Songbook" (Gimena del Rio Riande, SECRIT-CONICET; Clara Martínez Cantón, UNED; and Elena González Blanco-García, UNED)

"Using bioinformatic algorithms to analyze the politics of form in modernist Urdu poetry" (A. Sean Pue, Michigan State University; Tracy K. Teal, Michigan State University; and C. Titus Brown, University of California, Davis)

Panel title: New developments in quantitative metrics

Panel organizer: David J. Birnbaum, University of Pittsburgh, djbpitt@gmail.com

Panel synopsis: This panel presents new primary research results in the formal study of poetry and poetics that have been made possible by the development and use of innovative digital technologies. The research questions underlying the presentations are varied, as are the linguistic and cultural traditions (early modern and modern Russian; medieval Spanish; and Urdu [in comparison with Hindi/Sanskrit and Persian/Arabic]). What unites the three presentations is 1) a focus on using digital technologies to create humanities knowledge that would not otherwise be possible; 2) the development of innovative methodologies that are able to address those research questions; and 3) the building of new digital tools that make it possible to address new research needs.

Our panel responds to the following specific areas of emphasis noted in the original call for paper:

- *Humanities research enabled through digital media, data mining, software studies, or information design and modeling.* The focus of our panel is on new types of research results that are made possible by the development of original digital tools and methods. In this sense, the research results are foremost, but the projects that produce that research are methodologically innovative and the research results would not have been attainable without that innovation.
- *Creation and curation of humanities digital resources.* The creation of plain-text poetry archives is relatively straightforward: the text can be generated through OCR or by repurposing digital files originally created to produce print editions. The creation of structured poetry archives is also relatively straightforward: the general hierarchical structure of poetry is represented by pseudo-markup layout in plain-text editions, and is amenable to autotagging with the aid of regular-expression parsing. The creation of poetry archives with metrical and rhyme annotation, however, is difficult because it requires linguistic knowledge, and the presentations on this panel describe new technologies that were developed in order to 1) facilitate the machine-assisted creation of these types of metrically annotated poetic corpora, and 2) undertake original research about formal metrical practice on the basis of large digital corpora.
- *Social, institutional, global, multilingual, and multicultural aspects of digital humanities … For the 2015 conference, we particularly welcome contributions that address 'global' aspects of digital humanities including submissions on in-*

terdisciplinary work and new developments in the field. Our panel includes three research reports from three diverse poetic traditions: early modern and modern Russian, medieval Spanish, and Urdu (in comparison with Hindi/Sanskrit and Persian/Arabic). The projects that produced these research results have operated independently, but within their highly varied cultural contexts they address similar types of research questions.

- *Digital humanities in pedagogy and academic curricula.* The projects that contribute to our panel, which were designed to create new research knowledge in the humanities, were developed in many cases with attention to pedagogical and curricular concerns. For example, some portions of the development for these projects was carried out in the context of digital humanities courses, with undergraduate and graduate students making contributions to authentic humanities research as a way of learning to be digital humanists.

Paper title: Enabling the automated identification and analysis of meter and rhyme in Russian verse

Paper authors: David J. Birnbaum (University of Pittsburgh, djbpitt@gmail.com, corresponding author), Elise Thorsen (University of Pittsburgh, enthorsen@gmail.com)

Paper synopsis: Quantitative metrics, and particularly the statistical study of meter and rhyme, has been a core research methodology in Russian verse theory and scholarship at least since the early twentieth century both among Russian scholars (e.g., Belyj, Taranovski, Gasparov) and abroad (e.g., Shaw, Scherr, Friedberg). Until recently, the methods have had to rely largely on the laborious human identification and tagging or recording of all individual stress and rhyme phenomena, which have then served as input into the (often computer-assisted) statistical analysis of synchronic patterns and diachronic trends in meter and rhyme. Not only is this sort of manual effort at corpus preparation not scalable, but more often than not the raw data underlying published scholarship have not been shared, which means that the results cannot be replicated and the conclusions cannot be verified. Almost the entire corpus of Russian classical verse is now freely accessible on the Internet in authoritative scholarly digital editions, and computational tools could therefore be used to relieve scholars of the human labor previously needed to prepare and collect the data needed for studies in quantitative versification, and to perform the analysis. To the extent that the data preparation and analysis proceeds algorithmically, intermediate results can be saved and examined and the entire process can be replicated and verified.

The principal limitation to using available poetic texts for this purpose has been that the place of stress in words, which in Russian is not predictable without linguistic knowledge and which is not part of standard orthographic practice, must be known before the orthographic representation can be converted algorithmically to a useful phonetic representation, which is, in turn, a prerequisite for identifying meter and rhyme automatically. To address this need, the authors have built a network of tools, all freely available as web services, that automate as much as possible this process. The system begins with a full-text dictionary of Russian, consisting of approximately 100,000 headwords with all inflected forms, including information about the place of stress, which can be accessed through an API to add stress information to Russian-language texts in natural, native orthography. The stressed texts can be used for other purposes (such as in readers for language learners), but our principal goal was that they could then serve as input into API-driven web services that are capable of producing descriptive statistics and visualizations of the metrical patterns in individual poetic texts and in corpora. These integrated resources thus serve as a workstation for the formal and quantitative exploration of Russian versi-

fication in a way that is consistent with current best practice for research data management. All web services are publicly accessible and all data and other materials are available under a Creative Commons license.

There are other tools and services that address some of the same issues as our system, but none that enables users to process their own texts by means of a convenient pipeline of API-enabled web services that can convert corpora of texts in native Russian orthography into tagged output, descriptive statistical reports, and visualizations. Furthermore, our innovative two-pass methodology enables us to use regularities in the poetic structure to compensate for lacunae in the dictionary; the ambient meter that emerges on a first pass in the case of metrically regular poetry can be used, in a second pass, to infer the place of stress for words that either are not in the dictionary or return ambiguous results. The feedback portion of our system that corrects for lacunae and ambiguities does presuppose largely regular metrical and rhyme patterns, which means that this component of the system performs most reliably, effectively, and usefully with the largely regular syllabotonic verse that dominates Russian poetry from the beginning of the nineteenth well into the twentieth century. We also note, in response to early reader comments, that although the Russian National Corpus (RNC) provides a poetry subcorpus that purports to be able to return metrical information, in fact the RNC output merely layers the predetermined dominant metrical scheme of a poem on top of the text without regard to actual linguistic stress phenomena, which means that the output is of no value for determining which potential (metrical) stresses are realized (linguistically) and which were not.

Our conference presentation will illustrate the use of the system to answer types of research questions that are common in Russian versification studies, involving synchronic and diachronic properties and regularities of poets, periods, and forms and styles of poetry. In particular, metrical analyses have produced the concept of "semantic halo," a term coined by Mikhail Gasparov to describe intertexual meanings conventionally associated with a given meter. This is an observation enabled by a form of close reading performed numerous times, and because these observations of verse have by necessity been made by counting by hand, sample sizes have thus far been limited. The poetic canon is small enough that counting manually is feasible; however, scansion by hand could not begin to account for the large body of amateur and pulp poetry written from the mid-nineteenth century onward, or the extent to which this poetry does or does not reflect conclusions drawn from analyses of major nineteenth- and twentieth-century poets.

While the obscurity of most of these authors means their works remain undigitized, there is an ever-growing body of poetry published digitally. For example, there are

more than 30,000 poems self-published by members of the online journal *Poeziia.ru*. A corpus of this size offers the opportunity to elucidate the relationship of the larger population of those who consume and produce poetry with the poetic canon and contemporary critically acclaimed poets. With large-scale data about the use of meter, rhythm, and characteristic vocabulary, our research questions address the extent to which metrical semantic halos operate in this corpus and the extent to which they reflect patterns of readership and influence from the canon. What kinds of semantic halos are rarified phenomena, and which are more universally accepted as verse norms?

References

- Belyj, Andrej. 1929. "Ritm kak dialektika i 'Mednyj vsadnik.'" Moscow: Federacija.
- Friedberg, Nila. 2011. *English rhythms in Russian verse: On the experiment of Joseph Brodsky.* (Trends in Linguistics. Studies and Monographs 232.) Berlin: Mouton de Gruyter.
- Gasparov, Mixail Leonovič. 1984. *Očerk istorii russkogo stixa. Metrika, ritmika, rifma, strofika.* Moscow: Nauka.
- Scherr, Barry P. 1986. *Russian poetry: Meter, rhythm, and rhyme.* Berkeley: University of California Press.
- Shaw, J. Thomas. 1993. *Pushkin's poetics of the unexpected: The nonrhymed lines in the rhymed poetry and the rhymed lines in the nonrhymed poetry.* Columbus, OH: Slavica.
- Taranovski, Kiril. 1953. *Ruski dvodelni ritmovi.* Belgrade: Srpska akademija nauka.

Paper title: Making visible the invisible: metrical patterns, contrafacture and compilation in a Medieval Castilian Songbook

Paper authors: Gimena del Rio Riande (SECRIT-CONICET, gdelrio.riande@gmail.com), Clara Martínez Cantón (UNED, cimartinez@flog.uned.es), Elena González Blanco-García (UNED, elenagbg@gmail.com, corresponding author)

Paper synopsis: ReMetCa, Digital Repertoire on the Metrics of the Medieval Castilian Poetry (*Repertorio Métrico Digital de la Poesía Medieval Castellana*) is an online, free-access digital tool designed to undertake simultaneous complex searches on the metrical and rhyming patterns of Medieval Castilian poetry (starting from the late twelfth-century epics to the poetry of the sixteenth-century Castilian *Cancioneros*). ReMetCa is part of the corpus of online digital resources on Medieval Romance poetry, alongside such others as the ones related to Galician-Portuguese (*MedDB*, *The Oxford Cantigas de Santa Maria Database*), French (*BedTrouveres*, *Nouveau Naetebus*), and Occitan and Catalonian poetry (*BedT*, *Corpus des Trobadours*). The research project that sustains ReMetCa aims to integrate traditional studies of philology (especially those pertaining to metrics) with Digital Humanities, revising and classifying the Castilian corpus in a hybrid digital framework that embeds TEI-Verse module tags in a relational database that works altogether with a controlled vocabulary on Medieval Castilian Poetry.

One interesting case of study that illustrates the topic of the panel is our digital approach to the *Cancionero de Baena*, a large songbook containing almost six hundred poems transcribed and compiled in the first half of the fifteenth century by Juan Alfonso de Baena, scribe of the court of King Juan II of Castile. The data retrieved from the tagging and classification of the metrical and rhyming patterns of a large section of Baena's songbook—the one regarding the *antiquiores* or the eldest poets, and those that composed their texts mainly in the second half of the fourteenth century—yielded interesting results in the area of the Hispanic Studies devoted to metrics. On the one hand, we discovered that the *antiquiores* composed a large part of their poems using a pattern almost unknown by their predecessors, the Galician-Portuguese troubadours: the octosyllabic *octava* (eight-lined stanza, lines of eight syllables that may sometimes be isometric or heterometric). This pattern was shaped in a body of four stanzas (*glosa*) with rhymes structured in a singular pattern (*rimas singulares*) and words stressed on the penultimate syllable (*rima grave* or *femenina*). Furthermore, we were able to identify some groups or cycles of poems composed on the basis of the metrical and rhyming imitation or *contrafacture* (→ 4x8@8) (Spanke 1928: 73–117, Marshall 1980: 289–335, Rossell 2000: 149–156).

It was the theoretical organization of the XML markup as a discursive system (Jockers-Flanders 2013), which we used to shape an ontology framework (available at:

http://www.purl.org/net/remetca and http://datahub.io/dataset/remetca-ontology), that in practice helped us to move from the descriptive to the connotative dimension, thus making visible the invisible. Apart from using the expected TEI-verse attributes such as @met for our schemes based on the number of syllables of each line and on the number of lines (e.g.: 8,8,8,8), and @rhyme for the rhyming structure of the stanzas (e.g.: abba), there were some new attributes that did not exist in the TEI-Verse module and that we decided to add to our XML schema: @asonancia, an attribute that indicates the two possible values of the rhyming typology: "asonante" (which means that only vowel sounds are repeated) and "consonante" (which means that every sound after the stressed syllable is repeated), @unisonancia, which takes the values of "unisonante" or "singular" and shows whether the same rhyming scheme (e.g.: abba) is repeated in different stanzas or not, and @isometrismo, which states whether all the stanzas have the same number of syllables (isométrico) or not (heterométrico). It was the joint work of all these attributes that retrieved the metrical and rhyming patterns and led us to those new results. In addition to this, the whole analysis of the corpus resulted in an unexpected fact: Juan Alfonso de Baena may have organized and compiled the texts of the different poets in his songbook guided not only by chronological order, but also following metrical patterns and types of stanza.

With the help of our digital tool we will illustrate possible contrafactures, common metrical and rhyming patterns, and cycles of poems in the *antiquiores*' corpus, and give an account of more complex definitions. The examples will also serve as an opportunity to cast our eyes again on the macro-microanalysis (Jockers 2013, Jockers-Flanders 2013, Liu 2014) and data-text (Marche 2012) debates in the field of the literary studies and the concepts of close-distant reading (Moretti 2013, Latour 2014) and relate them to a subject of study that is interested in formal patterns (and not as much in content) and acquires new meaning when compared through large corpora: metrics.

References

- Asperti, Stefano, Fabio Zinelli, et al. *BedT, Bibliografia Elettronica dei Trovatori*. www.bedt.it
- Brea, Mercedes, et al. *MedDB: Base de datos da Lírica profana Galego-Portuguesa*. http://www.cirp.es/bdo/med/meddb.html.
- González-Blanco, Elena, et al. *ReMetCa: Repertorio Métrico Digital de la Poesía Medieval Castellana*. www.remetca.uned.es
- Jockers, Matthew L. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

- Jockers, Matthew L. and Julia Flanders. "A matter of scale", 2013. http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1106&context=englishfacpubs
- Latour, Bruno, Opening plenary, Digital Humanities 2014 (DH2014). http://dh2014.org/videos/opening-night-bruno-latour/
- Liu, Alan, *The laws of cool: Knowledge work and the culture of information*. Chicago: University of Chicago Press, 2014.
- Marche, Stephen, "Literature is not data: against Digital Humanities". *Los Angeles Review of Books*, 2012. http://www.lareviewofbooks.org/article.php?id=1040&fulltext=1
- Marshall, J. H., "Pour l'étude des *contrafacta* dans la poésie des troubadours," *Romania* CI, 1980, pp. 289–335.
- Moretti, Franco, *Distant reading*. London: Verso, 2013.
- Parkinson, Stephen, et al. *The Oxford Cantigas de Santa Maria Database.* http://csm.mml.ox.ac.uk/.
- Rossell, Antoni, 2000, "Intertextualidad e intermelodicidad en la lírica medieval." In: Bagola, Beatrice (ed.) *La Lingüística española en la época de los descubrimientos: Actas del Coloquio en honor del profesor Hans-Josef Niederehe, Tréveris 16 a 17 de Junio de 1997*, Hamburg: Helmut Buske, pp. 149–156.
- Seláf, Levente. *Le Nouveau Naetebus. Répertoire des poèmes strophiques non-lyriques en langue française d'avant 1400*. www.nouveaunaetebus.elte.hu.
- Spanke, Hans, "Das öftere Aufreten von Strophenformen und Melodien in der altfranzösichen Lyrik", *Zeitschrift für französichen sprache und Literatur* 51, 1928, pp. 73–117.

Paper title: Using bioinformatic algorithms to analyze the politics of form in modernist Urdu poetry

Paper authors: A. Sean Pue (Michigan State University, pue@msu.edu, corresponding author), Tracy K. Teal (Michigan State University, tkteal@msu.edu), C. Titus Brown (University of California, Davis, ctb@msu.edu)

Paper synopsis: This paper has two aims. First, it shows how the authors—a humanist and two computational biologists—adapted graph-based algorithms used in genome assembly and multiple sequence analysis to scan the meter of Urdu poetry. Second, applying these techniques to modernist free-verse poetry of the early 1940s, the paper argues that data-rich analysis of poetic meter offers humanistic insights into the politics of literary form.

1. Adapting Bioinformatic Algorithms to Scan Classical Urdu Poetic Meters

Metrical Overview. The meter of Urdu poetry is quantitative, based on length rather than qualitative stress. The traditional theory of prosody, called *ʻurūz*, describes these meters following the metrical system devised for Arabic by al-Khalīl Ibn Ahmad of Basra (b. 718CE). The meter can be described as moraic, consisting of "long" and "short" metrical "syllables" arranged, in classical poetry, into metrical feet in particular combinations. Pronunciation plays a primary role in determining these metrical syllables, though orthography must also be taken into account in certain situations. The determination of "long" and "short" syllables follows particular rules. There is also considerable flexibility, as certain word-final long vowels can be long or short, metrical syllables can be formed across word boundaries, and additional short syllables can be inserted at particular locations.[i]

Interdisciplinary Analogy. The authors found a productive analogy, which reached across their respective disciplines, between the workflow of computational scansion and the central dogma of biology. In this sequential transfer of information, DNA *replicates* and *transcribes* into messenger RNA (mRNA), which *parses* into RNA codons, and finally *translates* into particular proteins.

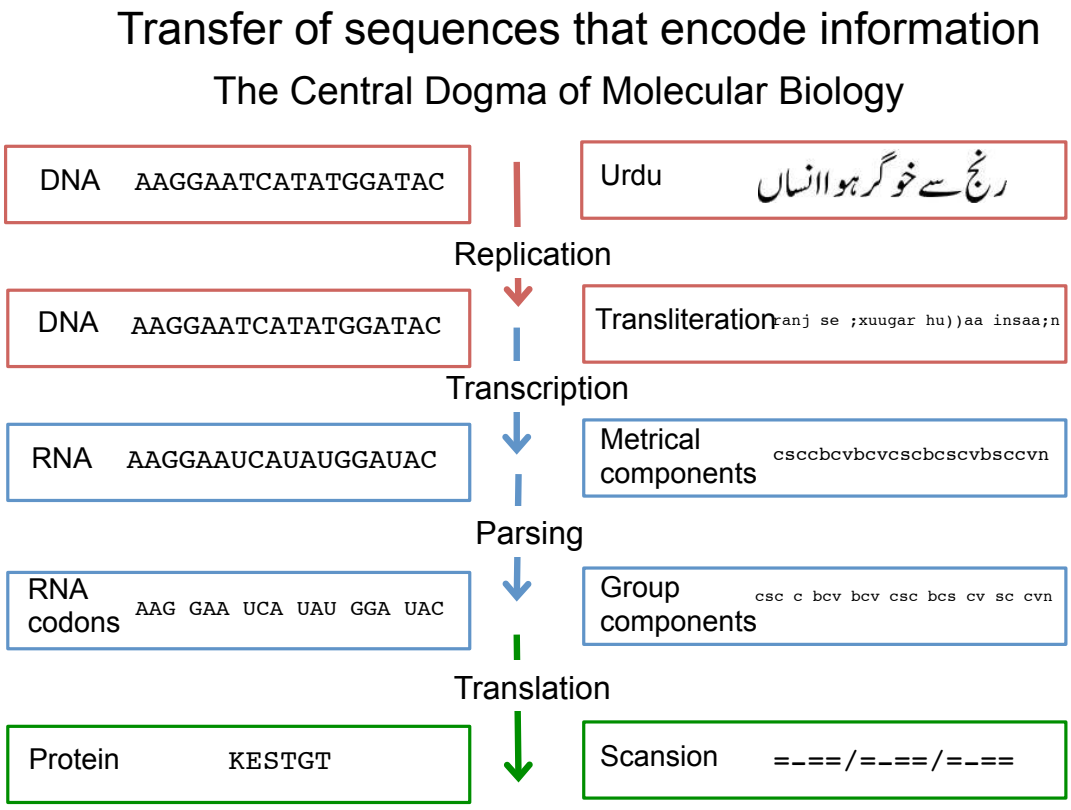

Figure 1. Interdisciplinary analogy

Computational scansion involves a similar process. The first stage is the *replication* of the Urdu text into its transliterated form, which shows additional information about the source text, most notably short vowels, required for metrical scansion. Next, in *transcription* the transliteration converts into metrical components, such as consonants, short vowels, and long vowels. Through *parsing*, the metrical components are grouped into short and long metrical syllables. Finally, *translation* renders those grouped components as a particular scansion, which can be described as a combination of long (=) and short (-) metrical syllables divided into feet or as a meter (*bahr*) named according to traditional Urdu prosody.

Adapted Graph Theory. Urdu meter is combinatorially explosive, leading quickly to hundreds or thousands of possibilities. The authors adapted graph theory, commonly used in "next generation" sequence analysis and genome assembly, in the transcription, parsing, and translation stages of this workflow.

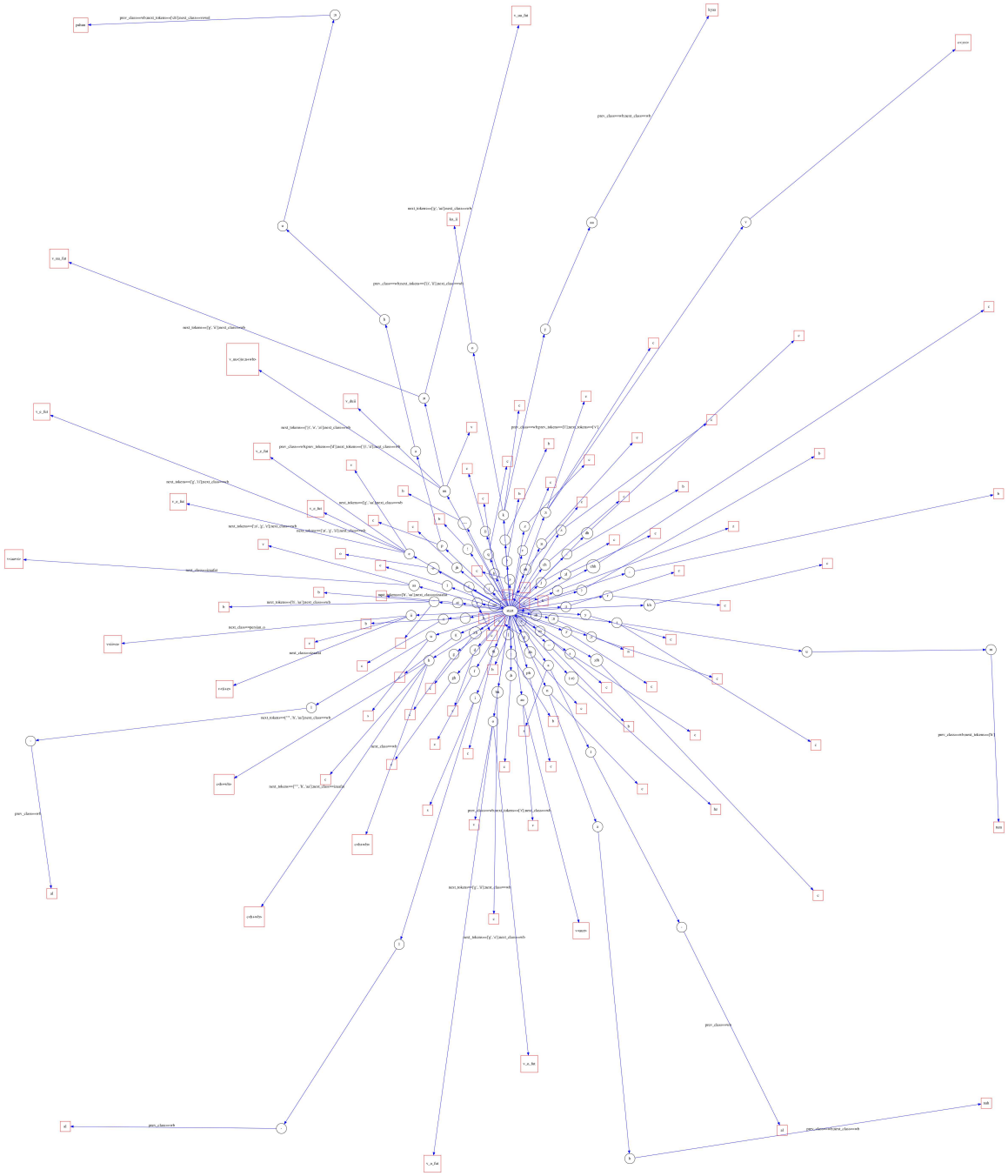

Figure 2. Transcription graph

During *transcription*, the transliteration is tokenized, and each token assigned a class, e.g. “consonant” or “short vowel.” A Depth-first Search (DFS) through a directed graph determines metrical components from these tokens. Nodes correspond to tokens to be consumed. Certain edges contain constraints that must be met before reaching a terminal node, corresponding to a classified metrical component.

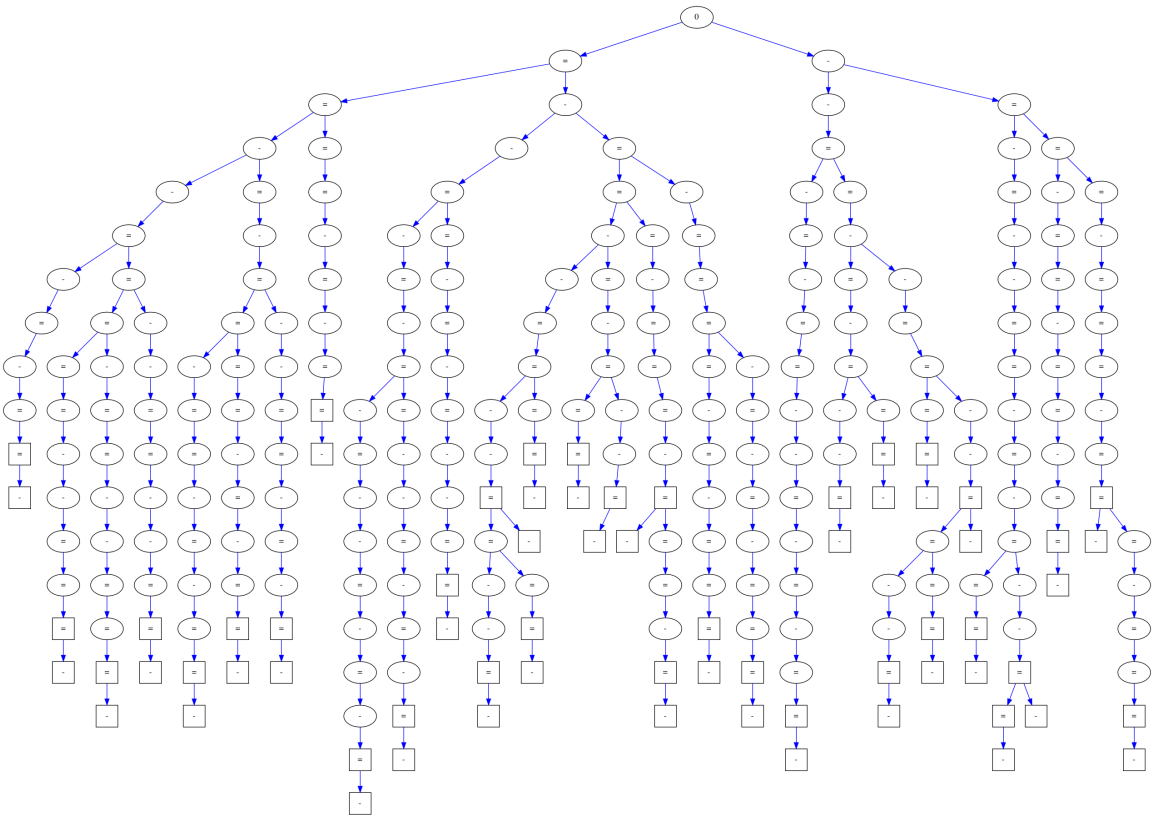

Figure 3. Parsing graph of a subset of classical meters

The *parsing* and *translation* stages utilize another graph, traversing which metrical components are consumed. Again, there are constraints on the edges, here to prevent certain invalid metrical combinations. This graph can consist of only acceptable, standard meters based on genre, or it can be adapted to allow for variations, as described below in the specific application to free verse. Multiple acceptable scansions may exist. The common meter can be found by requesting a union of the multiple sets of the lines' scansions.

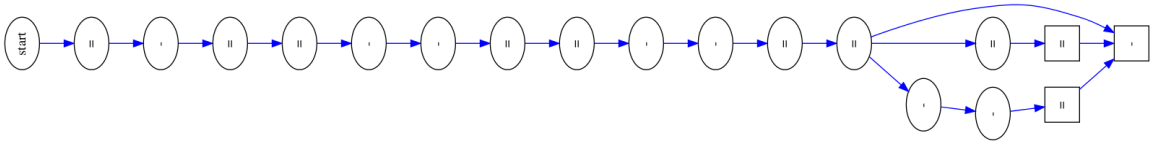

Figure 4. A Single Classical Meter

In adapting the algorithms designed for classical Urdu poetry to address free verse poetry, the authors also took a cue from bioinformatics, employing a score-based progressive algorithm commonly used in multiple sequence alignment, which seeks patterns despite frame shifts, substitutions, and mismatches to determine the *parsing* graph appropriate for particular free-verse poems.

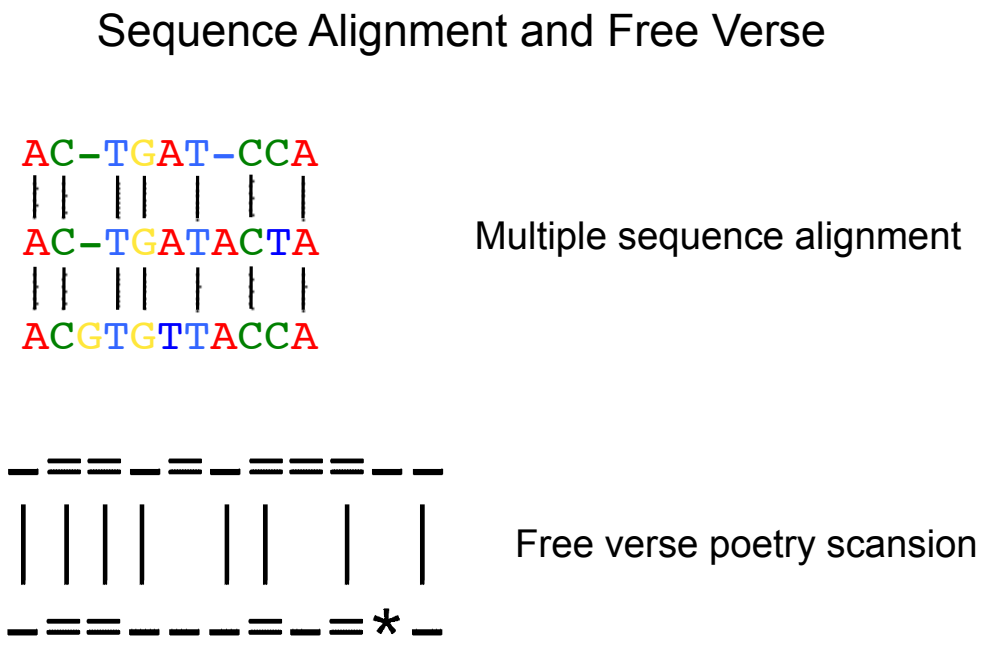

Figure 5. Sequence alignment and free verse

2. Metrical analysis and the politics of modernist literary form

Urdu poetry can usually be understood by speakers of the closely related, and mutually intelligible, South Asian language of Hindi, even while it retains a distinct vocabulary and meter as well as script. As noted above, poetry in Urdu is built primarily on the meters of Persian/Farsi, which took on the prosody of Arabic, thereby participating in a translingual Muslim poetic tradition. Borrowing moraic meters from poetry in the early-modern Braj Bhasha and classical Sanskrit, Hindi poets aligned themselves with the poetry of the Hindu devotional tradition.[ii] In the twentieth century, many Urdu poets were eager to modernize a literature criticized as foreign or too closely associated with the so-called misrule of the erstwhile Mughal Empire. They chose between retaining the treasured and melodious form of the Urdu and Indo-Persian ghazal, or looking away from that form and toward either English models or the meters embraced by Hindi.[iii] In the effort of anti-colonial nationalist movements to establish India and Pakistan in 1947, the choice of meter took on a political meaning: writers addressed different publics through the sound of their poetry. To some adherents of the "Two-Nation Theory," which argued that Indian Muslims were a separate and distinct nation, the *āhang* (poetic melody) of Indo-

Muslim civilization was external to the Indian soundscape. Other writers, including advocates of secular nationalism, insisted that Urdu poetry fell within a shared and distinct "Hindustani" literary culture that embraced both Hindu and Muslim elements.

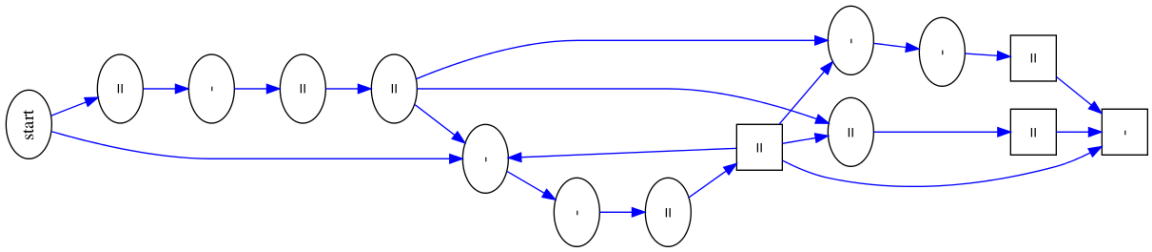

Figure 6. Free-verse meter (Abstracted from classical meter in Figure 4)

Among the latter group of authors was Sana Ullah Dar 'Miraji' (1912–49), who is often described as the progenitor of a modernist Hindustani literary culture that moves beyond older Persian forms to embrace South Asian folk culture.[iv] Hindustani displaced the Persian of the Mughal Court as the official language of colonial north India. It is commonly understood as both the origin language of contemporary Urdu and Hindi and a still-existing language of common address with a primarily indigenous vocabulary.[v] To describe Miraji's poetry as Hindustani is to evade casting his allegiance to either of the nations founded just before the poet's death. His most prominent critics prize how Miraji's poetry defies the solidification of language and nation; they base their claim primarily on the poet's choice of vocabulary common to both languages.

By applying the computational techniques described above, this paper explores how the poet negotiates these language politics less in terms of word choice than in the sound of his poetry. It argues that Miraji's innovative uses of meter are both fully transparent to his audiences and lie outside of the ideological construct of Hindustani as a "composite" language.

The paper examines the metrical variation in Miraji's free-verse poetry by pitching its analysis at a formal level independently from the linguistic register. It shows how, in addition to abstracting from classical Urdu poetry, as did his modernist contemporaries, Miraji also attempted to provide an indigenous meter borrowed from poetry written before Hindi and Urdu audiences were separated by religious community. He does so by translating into Urdu prosody poetic forms from Braj Bhasha, which require certain combinations of *guru* (heavy) and *laghu* (light) metrical syllables, and by adapting into free verse the so-called Hindi meter associated most

strongly with the eighteenth-century Urdu poet Mir Taqi Mir. This paper argues that meter itself carries meaning, and examines the rhythmic quality of language, which is among the most prized aspects of poetry in the Indian Subcontinent.

3. Source and Data Dissemination

Along with the data, the source code for metrical parsing and visualizations will be available at http://github.com/seanpue/dh2015.

---

[i] G.D. Pybus, *Urdu prosody and rhetoric* (Lahore: Rama Krishna and Sons, 1924); Frances Pritchett and Kh. A. Khaliq, *Urdu meter: A practical handbook* (Madison, WI: South Asian Studies, University of Wisconsin at Madison, 1997); Ashwini Deo and Paul Kiparsky, "Poetries in contact: Arabic, Persian, and Urdu," in *Frontiers in comparative prosody*, ed. Mihhail Lotman and Maria-Kristiina Lotman (Bern: Peter Lang, 2011), 147–172.

[ii] Vasudha Dalmia, *The nationalization of Hindu traditions: Bhāratendu Hariśchandra and nineteenth-century Banaras* (Delhi: Oxford University Press, 1997); Christopher R. King, *One language, two xcripts: The Hindi movement in nineteenth-century North India* (New York: Oxford University Press, 1994); Lucy Rosenstein, *New poetry in Hindi* (New Delhi: Permanent Black, 2002).

[iii] Frances W. Pritchett, *Nets of awareness: Urdu poetry and Its critics* (Berkeley: University of California Press, 1994).

[iv] Geeta Patel, *Lyrical movements, historical hauntings: On gender, colonialism, and desire in Miraji's Urdu poetry* (Palo Alto: Stanford University Press, 2002).

[v] Bernard Cohn, "Command of language and the language of command," in *Colonialism and its forms of knowledge: The British in India* (Princeton, NJ: Princeton University Press, 1996), 16–57; David Lelyveld, "Colonial knowledge and the fate of Hindustani," *Comparative studies in society and history* 35, no. 4 (October 1993), 665–682.