Prosodic Clustering via Cosine Similarity of Sound Sequence Inventories

Christopher Hench
chench@berkeley.edu
UC Berkeley, United States of America

Much of the discourse on music and rhythm before the time of the Minnesänger ('love singers', composers of German love poetry and songs around 1200) considers Latin chant. Alberic of Monte Cassino, active one century earlier than the Minnesänger, was the first to address rhythmic poetry in a music theory context in his treatise *De rithmis*, but he crucially does not mention rhyme, preferring to distinguish differing treatments of syllable length (Davis, 1966; Fassler, 1987). Not until the 12th and 13th centuries do treatises emerge elevating rhyme as integral to rhythm (Mari, 1899). A clear trend emerges in the period directly preceding MHG's Blütezeit (intense period of literary production)—theorists deemphasize syllable length in favor of a greater emphasis on syllable count and rhyme. After the Minnesänger, the Meistersänger ('master singers') of the German early modern period believed they were continuing the famed tradition of the medieval Minnesänger by focusing their art on syllable count and placement (März, 2000). For the Meistersänger, this focus was less a reflection of the increasing emphasis on music in poetic texts beginning in the 14th century, and more a method to reproduce the work of the medieval poets as closely as possible.

Unfortunately, we know very little about rhythm and sound in vernacular German poetry between these Latin treatises and the songs of the Meistersänger. Yet as both surrounding periods point to the syllable as fundamental to the composition of rhythmic poetry, we suggest the syllable as a rich source of formal information, both aesthetic and stylistic, which can serve as a formal alternative to common lexical methods in quantitative analyses, and which can help disambiguate the tension between form and content when subsequently compared to lexical methods. Although the broader project encompasses several analyses of sound and rhythm, this paper focuses on one application in particular—contrasting and visualizing the different uses of prosodic sound in the medieval German corpus (written in a stage of the language referred to as Middle High German (MHG)) through cosine similarity measurements of prosodic feature sequence tfidf (term frequency inverse document frequency) matrices.

Fortunately, although no unambiguous accounts of MHG rhythm survive, the manuscripts do provide phonologic evidence. We know the language of MHG had vowels of varying sonority, was structured in syllables, and was composed of words, regardless of the dialect or orthography (Paul et al., 2007). While we know that long vowels, and thus varying heavy syllables, did exist, we cannot be sure how the authors intended them in verse. Metrical systems have also been theorized, but orthographic variation in long vowels and theoretical disputes allow only limited, though productive, investigations (Estes and Hench, 2016).

To consider rhythm and sound broadly across the entire MHG corpus we turn to a relatively simple NLP technique in calculating the cosine similarity between tfidf inventories of texts. Yet our 'terms' are not words, but rather syllable features. For guidance, we turn to biology and new methods for clustering DNA sequences (Volkovich et al. 2005; Tomović et al. 2006; Maetschke et al. 2016). A technique employed in this scholarship takes n-gram samples of DNA strands. Similarly, we propose taking n-gram samples of prosodic features, primarily syllable features (closed 'C', open 'O', and word boundaries '-'), and constructing something resembling a tfidf matrix. Because our syllabification methodology is a combination of the sonority sequencing principle (Jesperson, 1904) and onset maximization and legal initials (Venneman, 1995), our syllabification method does not bias a dialect or orthography, but is accurate for most variants of MHG (Estes and Hench, 2016). Our features are coded as below:

Ein ritter sô gelêret was, (*Der Arme Heinrich*, l. 1-2)
C-CC-O-OOC-C-X

daz er an den buochen las,
("There was a knight so learned, that he read in the books")
C-C-C-C-OC-C-1

Where 'C' is a closed syllable (ends in a consonant), and 'O' is an open syllable (ends in a vowel). Hyphens for word boundaries account for the stress-initial tendency of MHG. Numbers at the end of a line mark end-rhyme; the number is how many lines back the

rhyme was seen and an 'X' stands for the beginning of a rhyme pair (it was not seen in the past lines). All sequences for a text are then joined:

C-CC-O-OOC-C-X-C-C-C-C-OC-C-1 [...]

N-grams with an *n* of 10 are taken from this long string, going between lines for coherency, resulting in a tfidf feature matrix with feature strings of length 10. Most choices of *n* yield similar results; a lower *n* simply results in a higher degree of similarity between every text. With every increase of *n*, this similarity inevitably decreases. An *n* of 10 allows for sequences of around three to four words to be compared across texts.

A match between the MHG epics *Parzival* and *Tristan* illustrates these features:

Ist zwîvel herzen nâchgebûr
("If the heart lives with doubt",) (*Parzival* l. 1)

C-OC-CC-COC-Xvon sinen schulden ungemach
([which had] suffered due to him) (*Tristan* l. 769)
C-OC-CC-COC-X

This match implies that the number of words and syllables per word are the same, the syllable quality patterning is the same, and importantly, the rhyme is the leading rhyme (in a pair). Because the sequence matches for 13 features including word boundaries, this will create three additional matches in the tfidf inventories when the 10-gram samples are taken. While these two lines also happen to share the same scansion in the Heusler tradition, we cannot assume that every match also bears the same scansion, as more than phonology dictates metrical value (Heusler, 1956). A visualization of these cosine similarity relationships between 595 verse texts from the *Mittelhochdeutsche Begriffsdatenbank* is available at http://mhg-sound.appspot.com (MHDBDB, 2016).'

One may object: what if this method does not abstract enough? What if the syllable sequences texts share are exact lexical matches? In order to determine to what degree this prosodic sequencing approach is lexically driven, we undertake two separate measures: 1) correlation and rank correlation between the suggested formal method and a traditional lexical method, and 2) on the basis of two sample texts, we remove every possible lexical match in the inventory of sound DNA matches via a Levenshtein distance threshold, and recalculate the cosine similarities.

To account for genre intertexualities or formulas affecting this measure we take a prototypical Arthurian romance *Iwein* and its 10 most similar texts. When considering texts between an oral and written culture it is important to recognize formulaic language as influential on genre, a field pioneered by Adam Parry and Milman Parry for Homeric verse (Parry and Parry, 1971); in the Germanic tradition Franz Bäuml (Bäuml, 1972; 1976). For each n-gram sequence match in *Iwein* to each of the 10 most similar texts, we evaluate the corresponding lexical strings of the matching prosodic sequences with the Levenshtein ratio (the Levenshtein ratio is defined by the Levenshtein distance (edit distance) divided by the alignment length), if the ratio is > .85, the prosodic sequence is removed from *Iwein*, e.g., 'wîp unde man â' ≈ 'wîp unde man ze' has a Levenshtein ratio > .85, so all sequences of '-C-CO-C-XO' are removed from *Iwein*. Removing sequences of close lexical matches in *Iwein* removes 40.65% of the prosodic sequence feature strings, yet correlation of text cosine similarities before and after removal remains high at .991, and top 10 and top 20 overlap are 60% and 70% respectively, implying that text similarities are not primarily lexically driven, though lexical similarities still account for many of the mutual sequences.

To investigate this further and disambiguate the relationship between form and content, the correlation (Pearson's and Spearman's) is calculated between the similarity ranks of the prosodic sequencing method and ranks of a traditional lexical method using lemmata unigrams (*r* .624 (rank .640)), bigrams (*r* .799 (rank .801)), and trigrams (*r* .834 (rank .839)). While these coefficients are high, the main concern are the nearest neighbors, a top 20 overlap of the two methods reveals a slightly different picture: unigrams (21.8%), bigrams (32.6%), and trigrams (36.2%). These results suggest that while form and content in MHG together contribute to what one may call 'genre', a large share of this grouping may be similarities in form derived from the prosodic sequencing features.

Which texts exhibit the most similar and most different rank similarities between the two methods? The top 20 overlap argues that the best-matched texts in form and content are Heinrich von Veldeke's *Eneide*, Konrad von Würzburg's *Herzmaere*, and Gottfried von Straßburg's *Tristan*. Interestingly, most texts in the top 10 for being best-matched are those most studied by scholars historically, and are broadly considered founders of the genre. In contrast, the most mismatched texts in form and content are Konrad's *Der Ritterspiegel*, *Die Klage der Kunst*, and the anonymous *Lohengrin* (none of the top 20 most similar texts measured by form are the same as the top 20

texts measured by content). Severe mismatch is often an intentional aesthetic strategy, made famous by Wolfram in *Willehalm* and *Titurel*, producing what Christoph März calls a "Verfremdungseffekt", or defamiliarization effect, per Shklovsky (März, 1999: 327).

Bibliography

Aue, H. von (2004). Der Arme Heinrich. *Gregorius ; Der Arme Heinrich ; Iwein*. 1. Aufl. (Bibliothek Des Mittelalters Bd. 6). Frankfurt am Main: Deutscher Klassiker Verlag.

Bäuml, F. H. (1986). The Oral Tradition and Middle High German Literature. *Oral Tradition*, **1**: 398–445.

Bäuml, F. H. and Bruno, A. M. (1972). Weiteres zur mündlichen Überlieferung des Nibelungenliedes. *Deutsche Vierteljahrsschrift Für Literaturwissenschaft Und Geistesgeschichte*, **46**: 479–93.

Davis, H. H. (1966). The "De rithmis" of Alberic of Monte Cassino: A Critical Edition. *Mediaeval Studies*, **28**: 198–227.

Estes, A. and Hench, C. (2016). Supervised Machine Learning for Hybrid Meter. *Proceedings of the Fifth Workshop on Computational Linguistics for Literature, NAACL-HLT 2016*: 1.

Fassler, M. E. (1987). Accent, Meter, and Rhythm in Medieval Treatises 'De rithmis'. *The Journal of Musicology*, **5**(2): 164–90.

Gottfried (1843). *Tristan und Isolt*. (Dichtungen Des Deutschen Mittelalters Bd. 2). Leipzig: Göschen.

Heusler, A. (1956). *Deutsche Versgeschichte: Mit Einschluss des Altenglischen und Altnordischen Stabreimverses*. 2 unveränderte Aufl. Vol. 2. (Grundriss Der Germanischen Philologie 8). Berlin: W. De Gruyter.

Jespersen, O. (1904). *Lehrbuch der Phonetik*. Leipzig, Teubner.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, **10**: 707.

Maetschke, S. R., Kassahn, K. S., Dunn, J. A., Han, S.-P., Curley, E. Z., Stacey, K. J. and Ragan, M. A. (2010). A visual framework for sequence analysis using n-grams and spectral rearrangement. *Bioinformatics*, **26**(6): 737–44.

Mari, G. (1899). *I Trattati Medievali Di Ritmica Latina*. . Vol. 11. U. Hoepli.

März, C. (1999). Metrik, eine Wissenschaft zwischen Zählen und Schwärmen?. In Müller, J.-D. and Wenzel, H. (eds), *Mittelalter: Neue Wege Durch Einen Alten Kontinent*. Stuttgart: Hirzel.

März, C. (2000). Der Silben Zall, der Chunsten Grunt. Die gezälte Silbe in Sangspruch und Meistergang. *Zeitschrift Für Deutsche Philologie*, **119**(2000): 73–84.

Mittelhochdeutsche Begriffsdatenbank (MHDBDB). Universität Salzburg. Koordination: Margarete Springeth. Technische Leitung: Nikolaus Morocutti/Daniel Schlager. 1992-2017. URL: http://www.mhdbdb.sbg.ac.at/ (2016).

Parry, M. and Milman, A. (1971). *The Making of Homeric Verse: The Collected Papers of Milman Parry*. Oxford: Clarendon Press.

Paul, H., Klein, T., Solms, H.-J. and Wegera, K.-P. (2007). *Mittelhochdeutsche Grammatik*. Tübingen: Niemeyer.

Tomović, A., Janičić, P. and Kešelj, V. (2006). N-Gram-based classification and unsupervised hierarchical clustering of genome sequences. *Computer Methods and Programs in Biomedicine*, **81**(2): 137–153.

Vennemann, T. (1995). Der Zusammenbruch der Quantität im Spätmittelalter und sein Einfluß auf die Metrik. *Quantitätsproblematik und Metrik*. (Amsterdamer Beiträge zur älteren Germanistik 42.1995). Amsterdam: Rodopi.

Volkovich, Z., Kirzhner, V., Bolshoy, A., Nevo, E. and Korol, A. (2005). The method of N-grams in large-scale clustering of DNA texts. *Pattern Recognition*, **38**(11): 1902–12.

Wolfram, Lachmann, K., Nellmann, E. and Kuhn, D. (1994). *Parzival*. 1 Aufl. (Bibliothek deutscher Klassiker 110). Frankfurt am Main: Deutscher Klassiker Verlag.