ARTICLE TEMPLATE

Lexical Proficiency in Jazz Improvisation: An NLP-Based Approach to Understanding Human Musical Language

Saebyul Park[1], Jiye Jung[2], Nahyun Kim[3], Juhan Nam[4]
[1]Yonsei University, South Korea [2]Heinrich Heine University Düsseldorf, Germany
[3]Ewha Womans University, South Korea [4]KAIST, Daejeon, South Korea

ARTICLE HISTORY
Compiled January 9, 2026

ABSTRACT
This study proposes a multidimensional framework for evaluating lexical proficiency in jazz improvisation by adopting approaches from lexical analysis in language research. Given that jazz improvisation is considered a complex form of musical language, we define four key dimensions: lexical diversity, vocabulary breadth, lexical sophistication, and lexical density. These capture the variation and range of melodic tokens, the rarity of advanced elements, and the informativeness of expressive content within solos, enabling systematic analysis of performers' melodic vocabulary. Each metric is computed using the Mel2word representation, which tokenizes symbolic music data via byte pair encoding (BPE) into meaningful units similar to linguistic tokens. For empirical application, the framework is evaluated on the Weimar Jazz Database, a curated corpus of annotated solo transcriptions spanning diverse performers and styles. By profiling artists' lexical proficiency across multiple dimensions, this study introduces a novel quantitative framework for exploring creativity, expression, and individual identity in jazz improvisation.

KEYWORDS
Lexical Proficiency; Jazz Improvisation; Quantitative Evaluation; NLP in Music Analysis

1. Introduction

Jazz improvisation is often regarded as a form of musical language, characterized by patterned melodic phrases and a shared vocabulary that fosters expressive coherence (Firtescu, 2022; Norgaard, 2012). This perspective has stimulated interdisciplinary research across neuroscience, linguistics, and computational modeling, revealing structural and communicative parallels between jazz and spoken language (Beaty, 2015; Carpenter Levitt, 2016; Donnay, Rankin, Lopez-Gonzalez, Jiradejvong Limb, 2014). Building on advances in natural language processing, recent studies have analyzed jazz improvisation as a musical corpus, applying linguistic techniques such as pattern discovery, statistical modeling, and similarity analysis to recurring motifs and phrase structures (Frieler, Pfleiderer Zaddach, 2016; Norgaard, 2014; Park Nam, 2023). Despite these developments, systematic methods for evaluating vocabulary proficiency

CONTACT Saebyul Park. Email: saebyul_park@yonsei.ac.kr
CONTACT Jiye Jung. Email: jujiy100@uni-duesseldorf.de
CONTACT Nahyun Kim. Email: nahyoni@ewhain.net
CONTACT Juhan Nam. Email: juhannam@kaist.ac.kr

in jazz improvisation remain limited. Considering jazz as a form of human language opens the possibility of adapting linguistic assessment frameworks to provide structured, interpretable measures of improvisational vocabulary.

To address this gap, our study aims to systematically quantify how jazz performers utilize melodic vocabulary in improvisation by adopting evaluation frameworks from linguistics. Building on established multidimensional measures of lexical proficiency (Halliday, 1989; Kyle Crossley, 2015; Laufer Nation, 1995; Malvern, Richards, Chipere Durán, 2004; McCarthy Jarvis, 2010; Meara Bell, 2001; Nation Nation, 2001; Read, 2000), we define four core dimensions—lexical diversity, vocabulary breadth, lexical sophistication and lexical density—for analyzing musical phrases, each capturing variation, range, rarity, or informativeness in melodic tokens. Using the Mel2word representation, which provides word-like tokenization of symbolic music data (Park, Choi, Kim Nam, 2024), we evaluate these dimensions across performers in the Weimar Jazz Database (Pfleiderer, 2017), a corpus of annotated jazz solos. This approach enables quantitative comparisons of improvisers' lexical profiles and offers interpretable metrics of expressive vocabulary, contributing to a deeper understanding of lexical competence and creative expression in jazz performance.

2. Related Work

2.1. Jazz as Language and Linguistics-Inspired Analysis

2.2. Multidimensional Lexical Assessment in NLP

2.3. Computational Modeling of proficiency?

3. The Multidimensional Lexical Proficiency Framework

3.1. NLP-Based Lexical Dimensions for Jazz Improvisation

To systematically evaluate performers' lexical proficiency in jazz improvisation, we adopt a multidimensional framework inspired by linguistic assessment models. Table **??** summarizes four core dimensions—lexical diversity, vocabulary breadth, lexical sophistication, and lexical density—drawn from established lexical evaluation literature and adapted for application to musical language. These dimensions enable quantitative assessment of how performers construct and deploy their melodic vocabulary using token-based representations of improvisational phrases, and in the following, each dimension is described as it is redefined for melodic analysis.

3.1.1. Lexical Diversity

Lexical diversity captures how much a performer varies their melodic vocabulary. We adopt the Hypergeometric Distribution Diversity (HD-D) metric (McCarthy Jarvis, 2010), which estimates the expected number of unique tokens in a fixed-size sample drawn without replacement:

$$\text{HD-D} = \frac{1}{S} \sum_{t \in V} \left(1 - P(X_t = 0)\right), \tag{1}$$

where S is the sample size and X_t is a hypergeometric random variable for token

Table 1. Four-Dimensional Framework for Melodic Lexical Proficiency with Linguistic Grounding

Dimension & Sub-Metric	Original Concept (Linguistics)	Melodic Redefinition
I. Lexical Size		
Breadth	Size of active vocabulary (Nation, 2001)	Total number of unique melodic tokens used within the solo
Diversity	Distributional balance of lexical usage (Read, 2000)	Evenness of melodic token usage across the solo
II. Lexical Quality		
Sophistication	Use of infrequent or advanced vocabulary (Laufer & Nation, 1995)	Degree of reliance on globally rare melodic tokens
Density	Density of meaning-bearing units (Halliday, 1985)	Informational compactness of melodic content reflected by tokenization and entropy
III. Structural Coherence		
Paraphrastic Cohesion	Repetition with variation preserving semantic continuity (Halliday & Hasan, 1976)	Global motivic development through transformed recurrence of core melodic ideas
Thematic Continuity	Local cohesive linkage across adjacent units (Hoey, 1991)	Smooth connective flow between adjacent melodic ideas without perceptual disruption
IV. Stylistic Novelty		
Distinctiveness	Speaker-specific lexical signatures (Biber, 1995)	Solo-specific token preference patterns deviating from corpus-wide norms
Novelty	Deviation from conventional usage and expectation	Degree of stylistic divergence from standard melodic patterns across time

t. Here, T is the total number of tokens (including duplicates) used by the performer, and f_t is the frequency of token t in that performer's token pool. $P(X_t = 0)$ is computed using the hypergeometric probability mass function, assuming tokens are drawn without replacement. To determine S, we use the median token length of all solos in the corpus. For each performer, HD-D is calculated on their full token pool.

3.1.2. Vocabulary Breadth

Vocabulary breadth quantifies the overall scope of a performer's melodic vocabulary across all solos. It is defined as the total number of unique melodic tokens used by the performer:

$$\text{Breadth} = |\text{Unique Melodic Tokens}| \tag{2}$$

To compare across performers, each breadth score is normalized by dividing by the maximum value observed across the dataset.

3.1.3. Lexical Sophistication

Lexical sophistication assesses how often a performer uses rare melodic material. We compute the average inverse document frequency (IDF) of all tokens used by each performer. IDF values are derived from the entire corpus of solos, capturing how uncommon each token is across all performers:

$$\text{Sophistication} = \frac{1}{T}\sum_{i=1}^{T} \text{IDF}(t_i), \tag{3}$$

with higher values indicating a preference for less conventional melodic vocabulary.

3.1.4. Lexical Density

Lexical density quantifies how much musically informative content is packed into a solo. Using a TF-IDF model trained on the entire corpus, we calculate the average TF-IDF value of all tokens used by a performer:

$$\text{Density} = \frac{1}{T}\sum_{i=1}^{T} \text{TF-IDF}(t_i). \tag{4}$$

Tokens that are locally frequent within a performer's output but globally rare across the corpus contribute more, indicating solos rich in meaningful, non-generic phrases.

3.2. NLP-Based Text Representation for Lexical Analysis

3.2.1. Mel2Word Representation

To represent melodies in a form suitable for lexical analysis, we adopt a modified version of *Mel2Word* (Park, Choi ., 2024), which converts symbolic music data into discrete morphemes and learns higher-level subword units through byte-pair encoding (BPE) (Sennrich, Haddow Birch, 2015). This representation provides a linguistic-like sequence structure that supports quantitative evaluation and enables downstream NLP-based tasks such as similarity analysis and performer classification (Park, Kim, Pak Kim, 2024; Park Nam, 2023).

Earlier Mel2Word versions relied on interval-based pitch encoding and coarse rhythmic quantization. In contrast, the present study introduces a modified conversion step that uses *absolute* pitch–rhythm morphemes to preserve fine-grained performance detail for both computational and perceptual analysis. Rhythmic precision is increased by incorporating binary subdivisions down to 32nd notes as well as explicit triplet durations (quarter-, eighth-, and sixteenth-note triplets). Durations exceeding one measure are compressed to the range 4.xxx beats while retaining their fractional structure (e.g., 5.250 → 4.250), preventing rare long notes from inflating the vocabulary or destabilizing BPE segmentation.

(A) Pitch tokens

Dictionary contains:
`P067_P067`
Input:
`P067, P067, P070`
Merged:
`P067_P067, P070`

(B) Rhythm tokens

Dictionary contains:
`B0100_B0050`
Input:
`B0100, B0050, B0100`
Merged:
`B0100_B0050, B0100`

(C) Combined tokens

Dictionary contains:
`P067B0100_P067B0050`
Input:
`P067B0100, P067B0050, P070B0100`
Merged:
`P067B0100_P067B0050, P070B0100`

Figure 1. Illustration of BPE-based subword tokenization for three representation types: (A) pitch-only tokens, (B) rhythm-only tokens, and (C) combined pitch–rhythm tokens. A merge occurs only when the corresponding multi-token pattern exists in the learned BPE dictionary for that representation.

Each note event is encoded as a musical morpheme of the form:

`PxxxByyyy,`

where `Pxxx` is the absolute MIDI pitch and `Byyyy` is the quantized duration in beats multiplied by 1000. Three parallel morpheme streams are produced for each solo:

- **Pitch-only morphemes:** `P067, P070, P072, P070, P067`
- **Rhythm-only morphemes:** `B0100, B0050, B0075, B0100, B0200`
- **Combined pitch–rhythm morphemes:** `P067B0100, P070B0050, P072B0075, P070B0100, P067B0200`

These sequences serve as the basis for the BPE-based lexical vocabulary.

3.2.2. Subword Tokenization Using BPE

Apart from the modified conversion into pitch and duration morphemes, all subsequent steps follow Park, Choi . (2024). Each representation type (pitch-only, rhythm-only, combined) is associated with its own BPE dictionary. Whenever a contiguous sequence of tokens appears in the learned dictionary, it is merged into a higher-level subword unit—analogous to morphemes in linguistic analysis.

These merged units constitute the lexical vocabulary used in all downstream analyses and serve as the basis for computing the multidimensional lexical proficiency metrics described earlier.

Together, the multidimensional lexical framework and the Mel2Word representation provide the foundation for our analysis. The next section details how these components are used to evaluate performers' lexical proficiency.

4. Experiment I: Computational Analysis

4.1. Dataset

We utilize the Weimar Jazz Database (Pfleiderer, 2017), a curated collection of annotated jazz solo transcriptions encompassing diverse styles and performers. The dataset provides detailed melodic, rhythmic, and harmonic annotations, accompanied by metadata including performer identity, style, and tempo, thereby enabling systematic computational analysis. Using our multidimensional lexical proficiency framework, we conduct a profiling analysis of performers' melodic vocabulary patterns to uncover distinctive aspects of expressive creativity and stylistic identity.

4.2. Textual Representation

For the computational assessment, we represent each melodic solo using the modified version of *Mel2Word* described earlier, in which absolute pitch–duration morphemes are used while the remaining tokenization procedure follows the general framework of Park, Choi . (2024). For each solo, three feature-specific sequences are generated (pitch, rhythm, combined), and lexical units are learned by applying byte-pair encoding (BPE) (Sennrich ., 2015) to the full set of sequences in the Weimar Jazz Database.

To construct lexical vocabularies, we use four dictionary sizes ($N = 100, 500, 1000$, and Full—where Full corresponds to the point at which no additional BPE merge pairs remain; e.g., pitch ≈ 3,999, rhythm ≈ 30,000, combined ≈ 10,000). For each performer and feature, all lexical metrics are computed separately for each dictionary size and then averaged, yielding a single score per metric that is less sensitive to the choice of vocabulary scale.

4.3. Lexical Proficiency Scoring

Using our framework, we compute multiple lexical metrics for each performer based on pitch-based token sequences. To ensure comparability, all metrics are min-max normalized to the $[0, 1]$ range, except for vocabulary breadth, which is max-normalized due to its long-tailed distribution. The overall proficiency score is then derived by averaging the normalized metrics for each performer.

4.4. Performer Ranking and Analysis

From the Weimar Jazz Database, we profile performers with more than five solo recordings ($N = 44$) and rank them according to their composite lexical proficiency scores. This analysis enables us to evaluate and compare performers' melodic vocabulary and to characterize their lexical proficiency across multiple dimensions.

4.5. Results

Figure 2 shows performers' scores across four lexical dimensions using stacked bars, enabling direct comparison. Figure 3 visualizes the top five performers with radar plots, revealing individual strengths and balance across dimensions. These results offer interpretable and structured insights into how performers deploy musical vocabulary across improvisational solos. Moving beyond subjective listening and expert interpretation, this study employs a multidimensional, linguistically motivated analysis to provide a comprehensive profile of performers' lexical proficiency, as well as their musical individuality and creative expression in jazz improvisation.

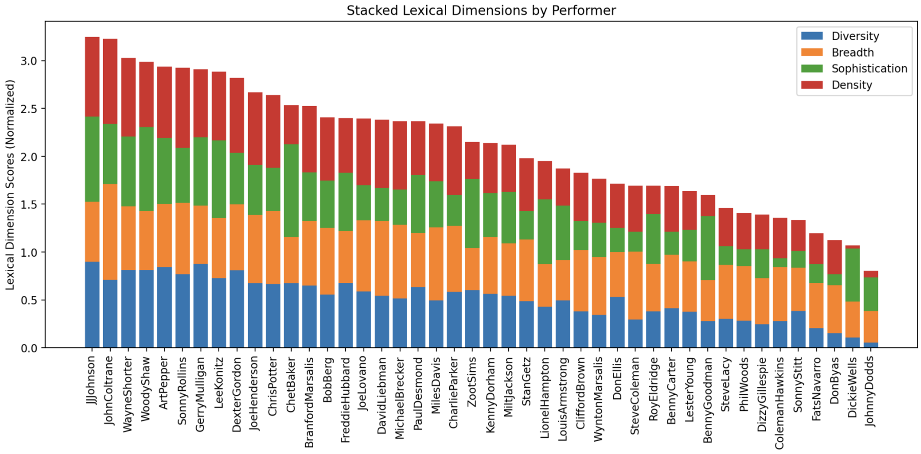

Figure 2. Lexical dimension scores visualized as stacked bars for each performer.

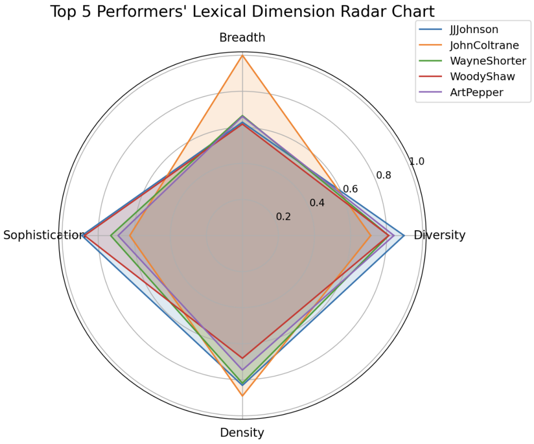

Figure 3. Top 5 Performers' Lexical Dimension Radar Chart

5. Experiment II: Expert Perceptual Judgment

5.1. Participants

A total of 20 adults participated in the evaluation study, consisting of 10 expert jazz musicians and 10 non-experts. The expert group included professional jazz performers and educators with formal academic training in jazz performance (bachelor's degree or higher) or equivalent professional experience. All had at least ten years of experience in jazz improvisation. The group represented a range of instruments, including saxophone, guitar, violin, and piano, ensuring stylistic diversity. The group consisted of both female and male musicians, with participants based in the United States, Canada, Norway, and South Korea, thereby reflecting expert judgments informed by both Western and Asian jazz performance contexts. The non-expert group consisted of adult listeners without formal music education or instrumental training. They were recruited to serve as a perceptual comparison group, enabling examination of whether computational indices of improvisational proficiency align more closely with expert judgments or with the intuitive impressions of general listeners.

5.2. Stimuli

A total of 100 segment pairs were prepared for the 2AFC evaluation, sampled from the Weimar Jazz Database and stratified across five tempo classes (UP, MEDIUM-UP, MEDIUM, MEDIUM-SLOW, SLOW), with 20 pairs per class. A minimum-length criterion based on the shortest complete solo in the corpus was applied to ensure that each excerpt contained sufficient melodic information, and external pairs were formed only between performances whose mean tempos differed by no more than ±10%, ensuring that excerpts within a pair were tempo-wise comparable. When a tempo class did not contain enough valid excerpts satisfying both conditions, additional internal pairs were created by sampling two non-overlapping regions from within the same performance, yielding a final set of 91 external and 9 internal comparisons. Excerpt starting points were determined through uniform random sampling, giving every token position equal probability and ensuring that short segments preserved coherent melodic continuity and phrase structure; each excerpt was approximately 15 seconds in duration. All excerpts were rendered as standardized MIDI piano realizations to control for timbre and recording variability, and performer-identifying information was removed. The final set of 100 pairs was selected to balance corpus coverage with practical session constraints, ensuring that each pair received judgments from at least three experts while keeping individual sessions within the intended 30-minute limit to support reliable evaluation.

5.3. Forced-Choice Judgment Task

Participants completed a two-alternative forced-choice (2AFC) task in which they heard two excerpts presented sequentially and selected the one that better matched the target evaluative criterion. The task relied on comparative rather than absolute judgments because assigning numerical ratings to short jazz improvisation excerpts is cognitively demanding and typically leads to inconsistent scale use. By contrast, choosing between two alternatives is a more intuitive and reliable perceptual operation, a principle grounded in classic comparative judgment theory (Thurstone, 1927). The 2AFC paradigm therefore provided a practical and robust method for capturing listeners' preferences while keeping the judgment process manageable.

5.4. Procedure

Each participant completed a total of 30 trials. In each trial, two anonymized excerpts were presented in randomized order, and the participant selected the excerpt that better reflected the target evaluative criterion. Excerpts were on average approximately 15 seconds in duration ($M = 15.40\,\mathrm{s}$), and no harmonic or contextual information was provided, ensuring that judgments were based solely on the melodic content. After each trial, participants indicated whether they recognized the performer so that potential familiarity effects could be addressed. Experts and non-experts completed the same number of trials under identical conditions. The total number of trials per participant was chosen to ensure that all one hundred segment pairs received evaluations from multiple listeners within each group, enabling estimation of inter-rater reliability (Cohen's κ) and providing sufficient redundancy for stable aggregation of pairwise preferences.

5.5. Evaluation Metrics

Evaluation metrics consisted of lexical measures originally adapted from linguistic analysis and redefined for musical application. Five computational indices were implemented: Lexical Diversity, Vocabulary Breadth, Lexical Sophistication, Lexical Density, and Motivic Development. These indices quantify, respectively, the variety, range, rarity, informational density, and developmental transformation of melodic tokens within improvisation. For human evaluation, the same five dimensions were employed, with the addition of Overall Proficiency, resulting in six dimensions in total. Overall Proficiency was included as a holistic measure of perceived improvisational skill, reflecting the way expert musicians naturally form global judgments beyond individual lexical features. This dimension also served as a validity check, allowing us to test whether detailed lexical ratings aligned with participants' general impression of performance quality. To validate whether the proposed performer-level metrics align with human perception, we conducted a listener-based 2AFC evaluation on randomly sampled solos. For direct comparison with human judgments, we additionally computed the same lexical metrics at the solo level, using only the melodic tokens from each sampled solo. Note that these solo-level metrics were used solely for comparison with human evaluation and were not part of the main corpus-level (performer-level) analysis. .

6. Conclusion

6.1. Ethical Approval

References

beaty2015creativeBeaty, RE. 2015. The neuroscience of musical improvisation The neuroscience of musical improvisation. Neuroscience & Biobehavioral Reviews51108–117.

carpenter2016rhythmCarpenter, AC. Levitt, AG. 2016. Rhythm in the speech and music of jazz and riddim musicians Rhythm in the speech and music of jazz and riddim musicians. Music Perception: An Interdisciplinary Journal34194–103.

donnay2014neuralDonnay, GF., Rankin, SK., Lopez-Gonzalez, M., Jiradejvong, P. Limb, CJ. 2014. Neural substrates of interactive musical improvisation: An fMRI study of 'trading fours' in jazz Neural substrates of interactive musical improvisation: An fmri study of 'trading fours' in jazz. PLoS ONE92e88665.

firtescu2022improvisedFirtescu, EC. 2022. Improvised musical performance as conversational language in jazz Improvised musical performance as conversational language in jazz. Artes. Journal of musicology25-26235–249.

frieler2016midlevelFrieler, K., Pfleiderer, M. Zaddach, WG. 2016. Midlevel analysis of monophonic jazz solos: A new approach to the study of improvisation Midlevel analysis of monophonic jazz solos: A new approach to the study of improvisation. Journal of New Music Research452129–145.

halliday1989spokenHalliday, MAK. 1989. Spoken and written language Spoken and written language.

kyle2015automaticallyKyle, K. Crossley, SA. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application Automatically assessing lexical sophistication: Indices, tools, findings, and application. Tesol Quarterly494757–786.

laufer1995vocabularyLaufer, B. Nation, P. 1995. Vocabulary size and use: Lexical richness in L2 written production Vocabulary size and use: Lexical richness in l2 written production. Applied linguistics163307–322.

malvern2004lexicalMalvern, D., Richards, B., Chipere, N. Durán, P. 2004. Lexical diversity and language development Lexical diversity and language development. Springer.

mccarthy2010mtldMcCarthy, PM. Jarvis, S. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. Behavior research methods422381–392.

meara2001pMeara, P. Bell, H. 2001. P-Lex: A Simple and Effective Way of Describing the lexical Characteristics of Short L2 Tests. P-lex: A simple and effective way of describing the lexical characteristics of short l2 tests. Prospect1635–19.

nation2001learningNation, IS. Nation, I. 2001. Learning vocabulary in another language Learning vocabulary in another language (10). Cambridge university press Cambridge.

norgaard2012jazzNorgaard, M. 2012. Descriptions of improvisational thinking by artist-level jazz musicians Descriptions of improvisational thinking by artist-level jazz musicians. Journal of Research in Music Education592109–127.

norgaard2014howNorgaard, M. 2014. How Jazz Musicians Improvise: The Central Role of Auditory and Motor Patterns How jazz musicians improvise: The central role of auditory and motor patterns. Music Perception313271–287.

park2024mel2wordPark, S., Choi, E., Kim, J. Nam, J. 2024. Mel2Word: A Text-Based Melody Representation for Symbolic Music Analysis Mel2word: A text-based melody representation for symbolic music analysis. Music & Science720592043231216254.

park2024quantitativePark, S., Kim, H., Pak, J. Kim, J. 2024. Quantitative analysis of melodic similarity in music copyright infringement cases Quantitative analysis of melodic similarity in music copyright infringement cases. International Society for Music Information Retrieval Conference. International Society for Music Information Retrieval. International society for music information retrieval conference. international society for music information retrieval.

park2023language$_{o}f_{j}azzPark, S.\ Nam, J.$ 2023.$TheLanguageofJazz: ANaturalLanguageProcessing–basedAnaly$

pfleiderer2017insidePfleiderer, M. 2017. Inside the Jazzomat: New Perspectives for Jazz Research Inside the jazzomat: New perspectives for jazz research.

read2000assessingRead, JA. 2000. Assessing vocabulary Assessing vocabulary. Cambridge university press.

sennrich2015neuralSennrich, R., Haddow, B. Birch, A. 2015. Neural machine translation of rare words with subword units Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.