The potential of c2xg's unsupervised learning for metaphor extraction in African American novels

Kamal Abou Mikhael
University of British Columbia
Vancouver, Canada
kamalabm@student.ubc.ca

Abstract

This paper presents a pilot study of metaphors of motion in African American literary language (AALL) in two sub-corpora of novels published in 1920-1925 and 1926-1930. It assesses the effectiveness of Dunn's (2024) unsupervised learning approach to computational construction grammar (c2xg) as a basis for searching for constructional metaphors, a purpose beyond its original design as a grammar-learning tool. This method is chosen for its statistical orientation and employed without pre-trained models to minimize bias towards standard language; its output is also used to choose a target search term. Focusing on the verbal phrase 'come to', the study analyzes argument-structure constructions that instantiate conceptual metaphors, most prominently experiencer-as-theme (e.g., 'he came to know') and experiencer-as-goal (e.g., 'thoughts came to her'). The evaluation compares c2xg coverage against a manually annotated set of metaphors and examines the uniformity of metaphor types extracted. Results show that c2xg captures 52% and 63% of metaphoric constructions in the two sub-corpora, with variation in coverage and uniformity depending on the ambiguity of the construct. The study underscores the value of combining computational and manual analysis to obtain outcomes that are both informative and ethically aware when studying marginalized varieties of English.

1 Introduction

Developments in Construction Grammar (CxG), Conceptual Metaphor (CMT), Natural Language Processing (NLP), and the study of African American English (AAE) call for linguistic inquiry that goes beyond the vernacular (AAVE) and uses computational tools to explore the construal of the African American experience in metaphoric constructions.[1] Accordingly, this paper documents the initial phase of a longitudinal study of metaphors of motion in the literary language of African American novels published between 1920 (Harlem Renaissance) and 1975 (Black Arts). Here the data consists of novels from 1920-1925 and 1926-1930. Given that a marginalized variety should be studied without bias toward dominant varieties of English, Dunn's (2024) unsupervised learning approach to computational construction grammar (henceforth c2xg) is considered a suitable candidate for such a project because it can learn a CxG grammar without prior training or pre-existing biased models. However, in addition to the ethical criteria, the grammar it generates must be assessed for its effectiveness in searching for constructional metaphors, a purpose beyond its original design as a grammar-learning tool. The evaluation compares its coverage against a manually annotated set of metaphors and examines the uniformity of metaphor types extracted.

The paper focuses on argument structure constructions containing metaphoric usages of 'come to' whose meaning is distinguished by constraints on pairing and positioning of arguments. The results show that most such constructions in the corpus are experiencer-as-theme (e.g., 'he came to know/realize/think') and experiencer-as-goal (e.g., 'thought/suspicion/love came to her'), along with idioms such as 'come to think of it', and 'what is coming to you'.

2 Background

Study of AAE began with a focus on African American Vernacular English (Labov, 1972), but has come to include African American Standard English, African American Middle Class English,

[1] Some use the term *Language* instead of *English*, thus AAE and AAVE are also referred to as AAL and AAVL, respectively.

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 202–207

African American Church Language, and various regional and demographic varieties (Bloomquist et al., 2015). The study of the language of African American literature has mainly focused on the representation of vernacular speech (Holton, 1984; Bailey, 1965; Williamson, 1970; Minnick, 2010; Green, 2002). Moreover, CMT based studies in African American literature have focused on a single work or a single author (Levinson, 2012; Mensah, 2011). This project focuses on the entirety of the language of literary works, a variety of AAE described as African American Literary Language (AALL). The language of novels is of interest because it evolved over time in two aspects: the representation of vernacular speech in dialogue, and the integration of vernacular forms into narration (Wideman, 1977). To approach AALL without assumptions, the study uses c2xg's unsupervised learning to discover metaphoric constructions that characterize AALL (Dunn, 2024).

3 Data

The data consists of the literary corpus and the output of the c2xg Python package. The corpus consists of novels from a list curated by the History of Black Writing Project (The Project on the History of Black Writing, 2024, 1987). For each time period (1920-1925 and 1926-1930) the works are narrowed down to ten based on the availability of the digital text and a strong element of realism which allows one to access metaphors that construe the African American experience.[2] The 1920-1925 sub-corpus contains 573,113 tokens and 1926-1930 contains 654,918. For each sub-corpus, a CxG grammar is generated using c2xg's Python implementation (Dunn, 2025). The CxG grammars are the generated from a single round of learning using the default parameters based on 500,000 words of each sub-corpus. Each grammar is a list of computationally derived constructions that can be augmented with examples from the corpus which are instantiations of the construction. These are referred to as *examples* in the c2xg documentation and they are henceforth referred to as c2xg examples to distinguish them from standard numbered linguistic examples listed in the paper. In this study c2xg is run to list all of the c2xg examples from which each construction are derived.

[2]This eliminates works of satire and historical fiction.

4 Methodology

Given the status and history of marginalization of the language being studied, the methodology prioritizes minimizing algorithm and human bias. To minimize algorithm bias, it uses the unsupervised learning of c2xg instead of language models and POS taggers that are skewed towards dominant varieties of English (Jørgensen et al., 2016; Hovy and Prabhumoye, 2021; Ziems et al., 2022) ; in addition, c2xg is run without any of its pre-trained models. To mitigate human bias, the target motion verb is chosen based on its frequency within the corpus and the metaphoric meanings it exhibits in the output of c2xg. The frequency is calculated from a word frequency list containing various inflections of 'come' (i.e., 'come', 'comes', 'came', 'coming', 'comin', and 'comin').

Although 'go' is more frequent than 'come' in the corpus, the latter is chosen because in the c2xg examples the verbal phrase 'come to' exhibits greater metaphoric variety in terms of argument structure. Metaphoric uses of 'go to' mainly consist of 'X going to Y' constructions in which subject X intends to take action Y (e.g., 'I'm going to quit', 'she was going to sleep'). On the other hand, 'come to' exists in a variety of metaphoric schematic (e.g., 'X comes to Y, Y=VP or NP: 'he came to love', 'he came to a decision') and idiomatic (e.g., 'come to think of it') constructions.

In order to assess the usefulness of c2xg for finding constructional metaphors two data sets are created. First, for each sub-corpus, an *evaluation set* is created using *verb-based* search (i.e., 'come to'). The results of the search are manually formatted to create a set of key-word-in-context (KWIC) entries where each 'come to' construction is annotated. Second, the evaluation sets are searched using corresponding c2xg examples (e.g., 'come to love') to create a pairs of search terms and matches (i.e., c2xg example and corresponding KWIC entry). The result is a *retrieved set* for each sub-corpus. These two sets are the basis for measuring *coverage* and *uniformity*.

4.1 Evaluation Set

The corpus is searched for 'come to' construct ('come to', 'came to', 'coming to', 'comin to', 'cominto', 'coming to', and 'comes to') to create a set of KWIC entries. Each 'X comes to Y' construction is delimited and marked according to its type: experiencer-as-theme, experiencer-

c2xg Example	KWIC Entry
['came', 'to', 'know']	Avey and my real relation to her, I thought I [came to know]+Verb.
['came', 'to', 'realize']	And although Peter [came to realize it]+Verb later it was many years before he told her so.

Table 1: c2xg examples and corresponding annotated KWIC entries.

as-goal, goal-as-physical-part (e.g., 'A determined look came to his face.'), other metaphor, and non-metaphor. Furthermore, the theme and goal in the construct are given semantic labels to distinguish them from other types of arguments (e.g., in 'come to love', the goal 'love' is labeled as 'Verb', and in 'peace came to her', the theme 'peace' is labeled as 'State').

4.2 Retrieved Set

The c2xg examples containing the various inflections of 'come to' are used to search the evaluation set. The search terms and matching results are paired to form the retrieved set. Table 1 shows a sample of entries; the '+' is shorthand annotation to mark experiencer-as-theme constructions.

4.3 Coverage and Uniformity

The results of the these two steps are used to evaluate *coverage* and *uniformity*. *Coverage* is the percentage of metaphoric constructions in the evaluation set that are found in the retrieved set ($metaphors\_retrieved/metaphors\_identified$). *Uniformity* is measured for c2xg examples that retrieve more than one result, and is the percentage of result sets that contain the same type of metaphor ($uniform/uniform + varied$). It is measured after the result set of each c2xg example is assigned a label. A result set consisting of one match is *single* and not part of the measure. Otherwise, if a result set has multiple matches which consist of the same metaphor type, it is *uniform*; otherwise, it is *varied*. These two measures indicate the usefulness of *c2xg* for locating metaphoric constructs.

5 Results and Analysis

Although the goal of this study is to assess coverage and uniformity, such a discussion is informed by an overview of the linguistic findings in the evaluation set. The first subsection gives an overview of the metaphoric constructions identified and annotated in the evaluation set. The second subsection provides an account of the coverage and uniformity.

5.1 Metaphoric Constructions

The main linguistic findings consist of the experiencer-as-theme and experiencer-as-goal schematic constructions that also include idioms. These constructions constitute the majority of the metaphors in the verb-based search results: 71.34% in 1920-1925 and 81.01% in 1926-1930.[3] In experiencer-as-theme, the experiencer is the theme because it is the subject of 'come to' and it can be a character (1a), group (1b), or general referent in the novel (1c). The experience it undergoes is construed with the conceptual metaphor CHANGE OF MENTAL STATE IS CHANGE OF LOCATION. The goal argument is realized through verbs featuring mental verbs or the verb 'be', and nominals denoting a state, event, or result.

(1) a. Peter came to realize it later
b. the cubs came to know him
c. handwriting all had come to know
d. he came to think it possible that
e. having come to understand
f. he had come to feel

The verbs convey cognition or perception such as 'realize', 'know', 'think', 'understand', and feel (1a-1f). The verb 'think' is also part of the idiom 'come/came to think of it' in which the experiencer can be implicit (2a) or explicit (2b). Other mental verbs include verbs of emotion such as 'love' (3a) and 'hate' (3b). Another notable verb is 'be' which introduces a state or result (4).

(2) a. Come to think of it they were ...
b. How'd you come to think of it?

[3]These results are calculated over example types (unique sequences of one or more tokens), as the same example instance may occur under multiple constructions, and some constructions are duplicates that yield identical sets of example instances.

(3) a. Just how I came to love her ...
b. We might come to hate each other

(4) a. you might come to be ashamed of me
b. puzzled by how they came to be there

Nominals of state or result in experiencer-as-theme include idiomatic (5a-5c) and compositional (5d-5f) constructs. In 'came/coming to himself' (5a-5b) the experiencer and the goal may seem to be the same, but this analysis considers 'himself' to refer to an ideal state of sound judgment that the experiencer had to arrive at. 'Came to' does not have an explicit goal, but the experiencer is understood to regain (arrive to) consciousness (5c).

(5) a. he came to himself
b. coming to himself
c. when he comes to there'll be no ...
d. came to the conclusion
e. she had come to the parting of the ways
f. what we are coming to

(6) a. thoughts of his condition came to her
b. suspicion had come to her
c. the desires ... which had come to her
d. the peace which had come to her

In experiencer-as-goal, the goal can be a character, part of a character (e.g., ears, mind), or thought (e.g., 'her reception of him'). The theme mostly consists of mental phenomena such as thoughts (6a), perceptions (6b), desires (6c), and psychological states 6d. The `c2xg` examples often do not contain the theme ('X' in 'X came to Y'). For example, the construct 'visions of Lida came to him' is extracted with the `c2xg` example 'Lida came to him'. `c2xg` examples containing a fully formed 'X came to him/her' extract non-metaphoric constructs such as 'she/he came to him/her'.

Although in general the goal in experiencer-as-goal refers to broad, external, and passive mental phenomena, there are a few instances where they refer to states or results which are the goal in experiencer-as-theme constructions. For example, the state 'senses' is the theme in (7a) whereas it is the goal in (7b). A similar example for the result 'decision' is found in (7c) and (7d).

(7) a. Then her senses came to her.
b. Just wait till you come to your senses!
c. A swift decision came to her.
d. before she could come to any decision

5.2 Coverage and Uniformity

`c2xg` example-based search has a total coverage of 51.79% and 62.5%. Table 2 shows a more detailed breakdown for each type of metaphor. It is evident that the individual coverage for each of these two constructions are not consistent across the corpora. Experiencer-as-theme has higher coverage in 1920-1925 whereas experiencer-as-goal's coverage is higher in 1926-1930. Certain metaphors are not extracted due to frequency of the search token sequence on the `c2xg` examples. For example, the phrase 'came to her' appears 13 times in 1920-1925, 7 of which are metaphors; in 1926-1930, it appears 42 times, 37 of which are metaphors. As a result, 'came to her' is represented in the grammar of 1920-1925 and not in 1926-1930.

Table 3 summarizes the measure of uniformity across metaphor types and non-metaphors. For the experiencer metaphors (-as-theme and -as-goal), uniform sets have almost double the uniformity in 1920-1925 compared to 1926-1930 (93% vs 42%). This is partially explained by the size of the result set: on average, a `c2xg` example that matches more than one result extracts 4.85 in 1920-1925, and 9 in 1926-1930. The `c2xg` examples that account for the majority of the varied results are incomplete phrases such as 'come to the' and 'come to her'. In the case of 'the', the type of metaphor is determined by what follows, resulting in experiencer-as-theme (8a), experiencer-as-goal (8b), and other metaphors (8c and 8d). In the case of 'her', the pronoun can be objective (9a) (experiencer-as-goal) or possessive (9b) (experiencer-as-theme), but possessive does not necessarily predict the type of metaphor (experiencer-as-goal).

(8) a. she ... came to the conclusion
b. sorrow ... come to the singers
c. she had come to the parting of the ways
d. the ... pessimist ... came to the front

(9) a. the thought had come to her
b. reluctant to come to her journey's end
c. phrases of thanks came to her mind

These preliminary observations of coverage and uniformity show that the evaluation of a tool like `c2xg` can inform how it is used and what is expected of it. In the case of coverage, one can expect `c2xg` to reduce the problem space by learning constructions of higher frequency, and a larger sample

Period	Source	Experiencer as Theme	Experiencer as Goal	Goal is Physical Part	Other Metaphor	Total
1920-1925	c2xg	34	24	0	9	69
	Corpus	56	56	6	39	157
	Coverage	61%	43%	0%	23%	43%
1926-1930	c2xg	21	59	3	8	91
	Corpus	48	80	5	25	158
	Coverage	44%	74%	60%	32%	58%

Table 2: Frequencies of 'come to' metaphors extracted by verb vs. c2xg examples.

Period	Source	Experiencer as Theme	Experiencer as Goal	Goal is Physical Part	Other Metaphor	Non-Metaphor
1920-1925	Single	28	6	0	2	45
	Uniform	13	4	0	3	23
	Varied	1	0	0	2	1
	Uniformity	93%	100%	0%	60%	96%
1926-1930	Single	15	3	0	3	79
	Uniform	5	2	0	0	32
	Varied	7	6	1	5	14
	Uniformity	42%	25%	0%	0%	70%

Table 3: Uniformity of metaphoricity in result sets extracted by c2xg examples.

size may ensure that key metaphors are not omitted. In the case of uniformity, c2xg examples consisting of incomplete phrases may reduce uniformity, but also may increase variety. Thus, the goals of the project would determine whether such phrases are used in the search process or whether they would need to be manually expanded in order to ensure greater uniformity in the results.

6 Conclusion

This paper described an ongoing-research project that is in its initial phases. It outlined a process for the use and evaluation of c2xg which was used to establish a statistical basis to identify potential metaphoric constructions. The metaphor analysis and argument structure analysis were fully manual. However, the process of extracting potential metaphors was done both computationally and manually which allowed for c2xg to be evaluated for uses beyond its intended purposes.

The main contribution of this study was insights on how the output of c2xg affects the extraction of metaphoric constructions so that it can be used in a manner that serves a project's objectives. Still, the usefulness of its data was not exhausted. Additional c2xg runs and rounds of learning are necessary, and there remains analysis of the computationally derived constructions and their translation into human-legible argument structure CxNs.

In the context of studying marginalized varieties of English, the unsupervised learning approach of c2xg presents a relatively safe start. However, given the presence of General American English (GAE) in the data encountered, at least for the period observed (1920-1930), there may be potential for the use of existing English NLP tools whose output can be monitored and evaluated so that they can be modified or that their output can be used in a manner that is less biased. The mindful and vigilant interplay between the computational and manual analysis of constructions and metaphor analysis is key for obtaining outcomes that are informative and ethically aware.

Acknowledgments

I would like to acknowledge Jonathan Dunn whose approach to computational CxG and the Python package he implemented were central in this paper along with his generous feedback and support. I also would like to acknowledge the feedback of my advisor Dr. Elise Stickles at the University of British Columbia.

References

Beryl Loftman Bailey. 1965. Toward a new perspective in Negro English dialectology. *American Speech*, 40(3):171–177.

Jennifer Bloomquist, Lisa J. Green, and Sonja L. Lanehart, editors. 2015. *The Oxford Handbook of African American Language*. Oxford University Press.

Jonathan Dunn. 2024. *Computational Construction Grammar: A Usage-Based Approach*. Elements in Cognitive Linguistics. Cambridge University Press.

Jonathan Dunn. 2025. jonathandunn/c2xg. Original-date: 2016-05-22T21:03:06Z.

Lisa J. Green. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press, Cambridge.

Sylvia Wallace Holton. 1984. *Down Home and Uptown: The Representation of Black Speech in American Fiction*. Fairleigh Dickinson University Press ; Associated University Presses, Rutherford : London.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2016. Learning a POS tagger for AAVE-like language. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1120, San Diego, California. Association for Computational Linguistics.

William Labov. 1972. *Language in the Inner City: Studies in the Black English Vernacular*. University of Pennsylvania Press, Philadelphia.

Julian Levinson. 2012. All the metaphors you are: Conceptual mappings of bebop in James Baldwin's Sonny's blues and Jack Kerouac's On the road. *Jazz research journal*, 6(1):69–87.

Eric Opoku Mensah. 2011. The metaphor: A rhetorical tool in some selected speeches of Martin Luther King, Jr. and Kwame Nkrumah. *Language in India*, 11(4):155–172. Publisher: Language in India.

Lisa Cohen Minnick. 2010. *Dialect and Dichotomy: Literary Representations of African American Speech*. The University of Alabama Press, Tuscaloosa.

The Project on the History of Black Writing. 1987. History of Black writing novel collections.

The Project on the History of Black Writing. 2024. History of Black writing (hbw) corpus.

John Wideman. 1977. Defining the Black Voice in Fiction. *Black American Literature Forum*, 11(3):79–82.

Juanita V. Williamson. 1970. Selected features of speech: Black and White. *CLA Journal*, 13(4):420–433.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding Dialect Disparity in NLU. ArXiv:2204.03031 [cs].