Pattern Discovery and Stylometric Analysis in English Literature and Literary Translation Through State Integration in Markovian Representations

C. H. C. Leung
School of Science and Engineering &
Guangdong Provincial Key Laboratory of Future Networks of Intelligence
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
clementleung@cuhk.edu.cn

C. J. Zeng
School of Humanities and Social Science
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
chenjiezeng@link.cuhk.edu.cn

***Abstract*—In analysing English literary work, the distinct aims and objectives are to determine the authorship, period, style, motif, and purpose. Here, the proper evaluation of results in English Literature is: first, place the known literary work in a machine learning model and discover their patterns and styles; second, compare the corresponding metrics with an unknown literary work. Since obtaining such knowledge from human experts is laborious and highly subjective, we align a data analysis method with extensions of the Markovian representations, which can be generalized to more versatile descriptions as the context develops. In particular, we consider the simple Markovian model and more elaborate generalisations that aim to remove the limitations of the memoryless properties of the basic Markovian representations. The first generalisation extends the state space by using the Cartesian product to form the composite state space, while the second approach exploits the stanza structure to integrate the states. The first approach can incorporate arbitrary long-time steps but leads to a high-dimension transition matrix. In contrast, the second more preferable approach yields a relatively small dimension matrix, which is computationally much more efficient. In addition, the latter approach also leads itself to further state integration by judiciously analysing the purpose of each line of a passage and provides the scope for analysing much larger corpora. Through the appropriate use of Markovian representation generalisations, examining the pattern of probability entries in the transition matrix, and applying this characterisation to the vast body of English literature, much more scientific, objective, and reliable decisions can be arrived at concerning proper authorship, writing style and other literary qualities.**

Keywords - Victorian novels; English poems; multi-step Markov chain; Shakespearean plays; Brontës; Sparse Matrix.

I. INTRODUCTION

This paper extends our previous paper [1] that analysed English literature by using a Markov model from a pragmatic aspect, namely, defining authorship. Still, computer-aided authorship attribution can be divided into two approaches, pragmatic and philosophical, which are related [9]. To have additional contributions, this paper uses both pragmatic and philosophical approaches to further analyse English literature and its Chinese translation by using multi-step Markov chains and expanding memory through state integration. The additional contributions of this paper comprise a new philosophical approach, a second language in literary work, and a new integration method. The analysis of vast English poems and their Chinese translation aims to determine several attributes. Such tasks may be viewed from a machine learning perspective [11] [12] [16] [17] [20], where one learns the quantitative features of known passages and uses these to determine obscure passages' attributes.

Historically, a manual approach to determining authorship primarily relied on human experts. However, it could sometimes be faulty as experts must systematically process the vast volume of literary data. For example, one poem, 'Stanzas', supposedly written by Emily Brontë, beginning 'Often rebuked, yet always back returning', is of uncertain authorship. Although Emily's sister, Charlotte Brontë, included it with the seventeen poems by Emily that she published in her selection of 'Literary Remains' that accompanied her 1850 edition of 'Wuthering Heights and 'Agnes Grey' [23].

Fortunately, unsolvable problems can be identified using machine learning. One positiveness brought by technological development is that a Machine learning approach could interplay and assist humans in a more complicated and challenging synthesis of ascertainable attributes: authorship, style, pattern, purpose, theme, and function. Stylometric analysis may indicate that a particular author's works are structurally distinct.

By using stochastic analysis, more efficient and reliable decisions can be made. Here, we shall use the Markov model's most straightforward form of dependency. The advantage of the Markov model is that introducing some dependency allows the styles to be much more richly represented than the simple independent model.

The rest of the paper is organised as follows. Section II provides a literature review, limitations, and achievements. Then using English literature examples, Section III motivates studying literary work through Markov chains. In Section IV, we explain the expansion of memory through state integration. Experiments on English passages and their Chinese translations are carried out in Section V. Finally, the paper concludes in Section VI.

II. RELATED WORKS

Previously, scholars in many fields have argued and studied the authorship issue. The historical record shows that William Shakespeare of Stratford-upon-Avon was identified as the person who was a player, a Globe shareholder, and the author of the plays and poems that carried his name. No evidence indicates

that anyone from the Elizabethan and Jacobean periods doubted this attribution [35].

On the other hand, several studies focused on authorship analysis of English literature using Markov chain models. The following studies demonstrate the effectiveness of Markov chain models in authorship analysis of English literature and suggest that more complex models, such as multi-step Markov chains or neural networks, may lead to even better performance. Specifically, Koppel et al. [36] used statistical markers, complex multivariate methods, and machine learning-based classification methods to examine the authorship. The analysis shows that two of the most sophisticated machine learning methods, SVM and Bayesian regression, offer an effective and efficient solution to the problem of authorship attribution. Also, Malyutov [37] looked at the near-optimal method based on Kolmogorov conditional complexity, attributing the discovered works of Shakespeare and those allegedly written by M. Twain, as well as binary discrimination of the Federalist papers by using Naive Bayes and other classifiers. Segarra, Eisen, and Ribeiro [38] introduced a method of authorship attribution called function word adjacency networks (WANs), which uses function words as nodes and directional edges to represent the likelihood of finding one function word in proximity to another. These WANs can be interpreted as transition probabilities of a Markov chain and are compared using relative entropies. Since function words are independent of content, their use tends to be specific to an author, making a good summary of stylometric fingerprints. Our previous papers [1] [2] used iambic pentameter to characterize these dimensions using Markov chains. We adopt a machine learning approach, processing and extracting known passages to create a signature transition matrix. Then we use a multi-step Markov chain to characterize the evolution of stress levels over time. The model can incorporate an amount of previous stress level memory, making it a flexible approach.

While Markov chains have been effective in authorship analysis of English literature, several limitations and challenges are associated with using Markov models. First, Markov chain models assume that the underlying data follows a stationary probability distribution, meaning that the statistical properties of the data do not change over time. However, in real-world scenarios, authors' writing styles may change over time for various reasons, such as personal experiences, age, or exposure to literature. Second, Markov chain models rely on selecting appropriate features and parameters to accurately represent an author's writing style. Choosing the right features and parameters can be challenging and time-consuming, requiring domain expertise and experimentation. Moreover, most of the previous studies focused on monolingual literature work in English. This paper not only analyses the monolingual literature, but also compares bilingual texts and the Chinese translation of English literature using a multi-step Markov model. Lastly, Markov chain models may struggle to differentiate between multiple authors who have similar writing styles or who have collaborated on a piece of writing. Additional techniques, such as stylometry or deep learning algorithms, may need to be employed in such cases. Therefore, while Markov chain models have been helpful in authorship analysis of English literature, they are not without their limitations and further research is required to overcome these challenges.

We mainly went through the following steps to achieve authorship analysis of English literature using multi-step Markov chain models. First, we collected data collection and gather a corpus of written works from multiple authors, such as William Shakespeare, the Brontë sisters, and W.B. Yeats. Second, we clean and preprocess the data to remove unwanted elements such as punctuation, numbers, or the last words. Then we select the representative data and extract its feature that can be used to represent the writing style of the author. For example, one or two authors could use the same rhythming patterns as features. Third, we train the model using multi-step Markov to identify each author's writing style. Fourth, to test the Markov model on a validation dataset, we evaluate its ability to correctly identify unknown texts' authorship. Therefore, we conduct two lingual experiments with the Markov chain to optimise the performance of the model in both monolingual and bilingual literature. Finally, we can use the trained model to attribute authorship of unknown texts with similar features to the data and texts we test and analyse in this paper.

III. BASIC MARKOV REPRESENTATION

Here, we illustrate the use of our approach by making use of some well-known poems, which are typical of similar poems. Charlotte Brontë has written more than 200 extant English poems, which remain of great interest as evidence of her developing ability to express emotion, her fascination with exotic characters and scenery, and her absorption of the techniques, images, and vocabulary of the poets whose work excited her. The poets that inspired Charlotte Brontë include but are not limited to William Wordsworth, William Shakespeare, and Samuel Taylor. Thus, Charlotte Brontë's poems almost always have a rich verbal texture, but her control of their style and structure is often insecure, except when she fair-copies or revises them [23]. Fortunately, among the five poems that were written or edited in 1847, two were used in 'Jane Eyre' [26]: 'My feet they are sore' (p. 22) and 'The truest love' (pp. 265-266).

English literature and poems could be differentiated from poetic devices, rhyme patterns, themes, and motifs. Still, the works written by the Brontë sisters share a lot of common grounds. For instance, we can see a common theme at the heart of all of Emily Brontë's poetry: the desire for a unified sense of the self and a simultaneous awareness and fear of the self's diffusion and fragmentation [24]. Likewise, 'Jane Eyre', written by Charlotte Brontë, is a story of enclosure and escape of 'self' and a story of the movement for freedom within the economic-social and cultural context in the Victorian era [23] [25].

Such a literary theme is rather hard to capture through a machine-learning approach. For example, in the first chapter [26], little Jane is scolded by Mrs Reed, who labels Jane's actions as 'cavillers or questioners' and demands Jane sit and remain silent until Jane can 'speak pleasantly'. To quietly perform the rebellious act and not be submissive, upset as little Jane is, she buries herself in the book entitled Bewick's History of British Birds. In the introductory pages of Bewick's book, little Jane finds comfort through pictures that portray the sea fowl inhabited by "the solitary rocks and promontories" of the Norway coast.

However, it will become attainable if we break down human experts' complicated and intuitive judgement into small elements that comprise an overall appreciation of a poem and

literature to improve. Among all the attributes, iambic pentameter, rhythm pattern, and phonetic stresses could be computer-friendly and play a key role in analysing poetry using a machine-learning approach. Iambic pentameter consists of ten syllables per line, stressing every second syllable. This creates a rhythmic pattern of one unstressed syllable followed by one stressed syllable in each pair. The rhythm pattern of iambic pentameter can be compared to the sound of a heartbeat, with each stressed syllable acting as a 'beat' in the poem. Phonetic stresses emphasise a particular syllable or word, typically due to tone, pitch, or volume. In iambic pentameter, the stress is determined by the natural accentuation and pronunciation of the words.

'Where the Northern Ocean', the first poem in 'Jane Eyre', occurs here. This poem's imaginative world of literature allows Jane's mind to escape confinement [27].

A D A D A D A D A D
^ | ^ | ^ | ^ | ^ |
*Whe**re** the **Nort**hern **Ocean**, in **va**st **whi**rls,*

A D A D A D A C A D
^ | ^ | ^ | ^ | ^ |
*Boils **round** the **nak**ed, **mel**an**cho**ly **isles***

A D A D A D A D A D
^ | ^ | ^ | ^ | ^ |
*Of **fart**hest **Thu**le; **and** the **Atlan**tic **surge***

A D A D A D A D A D
^ | ^ | ^ | ^ | ^ |
*Po**urs** in a**mong** the **stor**my **Heb**ri**des**.*

a. 'Where the Northern Ocean' in 'Jane Eyre', written by Charlotte Brontë. [26]

Stressed syllables vary in strength, while unstressed syllables vary in weakness [4]. In this paper, we notate the stressed sound with a "|" marking ictic syllables and a "^" marking unstressed syllables. In this notation, a standard line of iambic pentameter would look like "^ | ^ | ^ | ^ | ^ |", where each line of verse is made up of five two-syllable iambs for a total of ten syllables. As the meter is mainly about sound, not spelling, scansion adds numbers to indicate various stress levels to realize beats and offbeats (A = lightest stress, D = heaviest stress). Scansion is the analysis of regular patterns of accented, unaccented syllables.

The arrangement of words and phrases in poetic lines reflects our custom of speaking, and of hearing each other speak, in a succession of rhythmic units; if the lines are metrical, if they make patterns out of series of lightly or firmly stressed syllables, they reflect the fact that when we speak – we speak stressed syllables with greater and lesser degrees of stress [4]. It is possible to represent the stress structure by means of a Markov chain and so provide a characterization of the literary work. The Markov matrix of the above situation, when used as a sample, can be constructed as where the four states indicate the stress levels of A, B, C, and D, respectively,

$$\begin{pmatrix} 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Although the English syllables we speak can be spoken with many degrees or shades of emphasis on loudness, sharpness, duration, and other ways of signalling importance, it seems likely that in most English speech, we perceive mainly two significant levels of stress, and that we hear a continuous series of relatively stressed and relatively unstressed syllables [4]. Similarly, the following

A D A D A D A D A D
^ | ^ | ^ | ^ | ^ |
*His **spark**ling **eye**s, re**pleat** with **wrath**full **fire**.*

b. The Sonnets by William Shakespeare. [22]

can be constructed simply as

$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Another form of counter-lauréate authorship emerges in a more prominent place, Theseus's speech in the mid-1590s romantic comedy A Midsummer Night's Dream. Noticeably, Shakespeare's self-reflexive revision, such as inserting discourse about the 'poet' as a company for 'lunatic' and 'lover', turns a speech about the madness of love into one about the poet's role in forming an eternising state of consciousness [7].

A D A D A D A D A D
^ | ^ | ^ | ^ | ^ |
More **stran**ge **than** t**ue**. I **never** may be**lieve**

A D A D A D A D A D
^ | ^ | ^ | ^ | ^ |
These **an**tic **fa**bles, **nor** these **fai**ry **toy**s.

A D A D A D A D A D
^ | ^ | ^ | ^ | ^ |
Lo**vers** and **madmen** have **su**ch **see**thing **brains**,

A D A D A D A D A D
^ | ^ | ^ | ^ | ^ |
Such **shapi**ng **fan**tasies, **tha**t a**ppre**h**end**

c. The Sonnets by William Shakespeare. [22]

Most of the poems written by the Brontës are ended with the rhythm of 'abab' pattern, such as, 'My Feet They are Sore' (p. 22) written by Charlotte Brontë in her novel 'Jane Eyre' [26]. These two poems are fair copies or revised by Charlotte Brontës [26], thus we can see regular poetic patterns in them. Specifically, in the first stanza, the last words of each line are 'weary, wild, dreary, child'. If each of the last words is given a name using alphabetic order, we name the word of the same rhythm with the same alphabetic letter. Here in this stanza, 'weary' and 'dreary' are given a random letter 'a' because they

end with the same pronunciation of '-eary', whereas 'wild' and 'child' are given another name 'b' for ending with the '-ild' sound. The same logic applies to the remaining stanza of this poem: cdcd (the second stanza); efef (the third stanza). ghgh (the fourth stanza), ijij (the fifth stanza).

Stanza 1:

*My feet they are sore, and my limbs they are w**eary**; **(a)***

*Long is the way, and the mountains are w**ild**; **(b)***

*Soon will the twilight close moonless and dr**eary** **(a)***

*Over the path of the poor orphan ch**ild**. **(b)***

d. 'My Feet They are Sore' in 'Jane Eyre', written by Charlotte Brontë. [26]

To further analyse the poem with attributes, we can now look at the stresses of each word in each line. In a line of a poem, a foot refers to either a stressed syllable or an unstressed syllable. Both syllables then form distinct pairs, as a musical measure consists of a certain number of beats. Delimitation of the spoken chain sounds can be based on auditory impressions, but describing these sounds is an entirely different process. Description can be carried out based on the articulatory act, for it is impossible to analyze the sound units in their chain [16].

Stanza 2:

A B A C A D A B A D
^ | ^ | ^ | ^ | ^ |
*Why **did** they **send** me **so** far **and** so **lonely**, **(c)***

A B A C A B A C A C
^ | ^ | ^ | ^ | ^ |
*Up **where** the **moor**s spr**ead an**d grey **rocks** are **piled**? **(d)***

A B A D A C A D
^ | ^ | ^ | ^ |
*Men **are** hard-**hearted**, and **kind** angels **only** **(c)***

A B A C A D C D
^ | ^ | ^ | ^ |
*Watch **o'er** the **steps** of a **poor** orphan **child**. **(d)***

Stanza 3:

A D A C A C A B A C
^ | ^ | ^ | ^ | ^ |
*Yet **distant** and **soft** the **night** bree**ze is** blo**wing**, **(e)***

A B A C A B A C B C
^ | ^ | ^ | ^ | ^ |
*Clouds **there** are **none**, and **clea**r st**ar**s bea**m mild**, **(f)***

A B A C A C A D
^ | ^ | ^ | ^ |
*God, **in** His **mercy**, pro**tection** is **showing, (e)***

A B A C A D C D
^ | ^ | ^ | ^ |
*Comfor**t and** **hope** to the **poor** orphan **child**. **(f)***

Stanza 4:

A D A D A D A D
^ | ^ | ^ | ^ |
*Ev'n **should** I **fall o**'er the **broken** bridge **passing**, **(g)***

A D A D A D A D
^ | ^ | ^ | ^ |
*Or **stray** in the **marshes**, by **false** lights **beguiled**, **(h)***

A D A D A D A D A D
^ | ^ | ^ | ^ | ^ |
*Still **will** my **Fa**ther, **wi**th **promise** and **blessing**, **(g)***

A D A D A D A D A D
^ | ^ | ^ | ^ | ^ |
*Take **to** His **bos**om **the** poor **or**phan **child**. **(h)***

Stanza 5:

A D A D A D A D
^ | ^ | ^ | ^ |
*There **is** a **thought** that for **strength** should **avail** me, **(i)***

A D A D A D A D
^ | ^ | ^ | ^ |
*Though **both** of **shelter** and **kindred** de**spoiled**; **(j)***

A D A D A D A D A D
^ | ^ | ^ | ^ | ^ |
*Hea**ven** is a **home**, and a **rest** will **not** fail **me**; **(i)***

A D A D A D A D
^ | ^ | ^ | ^ |
*God **is** a **friend** to the **poor** orphan **child**. **(j)***

e. 'My Feet They are Sore' in 'Jane Eyre', written by Charlotte Brontë. [26]

The advantage of Markov models [13] [18] for analysing sequential data is that segmentation and classification are performed simultaneously in an integrated procedure. Using efficient and robust training and decoding algorithms, Markov model recognition systems can effectively be realized for large-scale classification systems [3].

The poem reappears in a different context when Jane and Rochester declare their love for each other on Midsummer Eve [26]. It serves a new purpose: to symbolise the to-be Mrs Rochester's potentialities for freedom and happiness, as well as the natural affinity between her and Rochester. The poem that starts with 'the truest love that ever heart' is narrated from Rochester's perspective, indicating a potential marriage and freedom afterwards.

Stanza 1:

A D A D A D A D
^ | ^ | ^ | ^ |
*The **tru**est **love** that **ev**er **heart***

A D A D A D
^ | ^ | ^ |
*Fel**t at** its **kin**dled **core**,*

A D A D A D A D
^ | ^ | ^ | ^ |
Did **through** each **vein**, in **quicken**ed **start**,

A D A D A D
^ | ^ | ^ |
The **tide** of **bei**ng **pour**.

A D A D A D A D
^ | ^ | ^ | ^ |
Her **com**ing **was** my **hope** each **day**,

A D A D A D
^ | ^ | ^ |
Her **part**ing **was** my **pain**;

A D A C A D A D
^ | ^ | ^ | ^ |
The **chance** that **did** her **steps** de**lay**

A D A D A D
^ | ^ | ^ |
Was **ice** in **ev**ery **vein**.

Stanza 2:

A D A D A D A D
^ | ^ | ^ | ^ |
I **dreamed** it **would** be **name**less **bliss**,

A D A D A D
^ | ^ | ^ |
As **I** loved, **loved** to **be**;

A D A D A D
^ | ^ | ^ |
And **to** this **object** did I **press**

A D A D A D
^ | ^ | ^ |
As **blind** as **ea**ger**ly**.

Stanza 3:

A D A D A D A D
^ | ^ | ^ | ^ |
But **wide** as **path**less **was** the **space**

A D A D A D
^ | ^ | ^ |
That **lay** our **lives** bet**ween**,

A D A C A D A D
^ | ^ | ^ | ^ |
And **dange**rous **as** the **foa**my **race**

A D A D A D
^ | ^ | ^ |
Of **oce**an-**sur**ges **green**.

Stanza 4:

A D A D A D A D
^ | ^ | ^ | ^ |
And **haun**ted **as** a **rob**ber-**path**

A D A D A D
^ | ^ | ^ |
Through **wild**er**ness** or **wood**;

A D A D A D A D
^ | ^ | ^ | ^ |
For **Might** and **Right**, and **Woe** and **Wrath**,

A D A D A D
^ | ^ | ^ |
Bet**ween** our **spir**its **stood**.

Stanza 5:

A D A D A D A D
^ | ^ | ^ | ^ |
I **dan**gers **dared**; I **hind**rance **scorned**;

A D A D A D
^ | ^ | ^ |
I **om**ens **did** de**fy**:

A D A D A D A D
^ | ^ | ^ | ^ |
Wha**te**ver **men**aced, **hara**ssed, **warned**,

A D A D A D
^ | ^ | ^ |
I **pass**ed im**petu**ous **by**.

Stanza 6:

A D A D A D A D
^ | ^ | ^ | ^ |
On **sped** my **rain**bow, **fast** as **light**;

A D A D A D
^ | ^ | ^ |
I **flew** as **in** a **dream**;

A D A D A D A D
^ | ^ | ^ | ^ |
For **glor**ious **rose** up**on** my **sight**

A D A D A D
^ | ^ | ^ |
That **child** of **Shower** and **Gleam**.

Stanza 7:

A D A D A D A D
^ | ^ | ^ | ^ |
Still **bright** on **clouds** of **suffer**ing **dim**

A D A D A D
^ | ^ | ^ |
*Shines **that** soft, **sol**emn **joy**;*

A D A D A D A D
^ | ^ | ^ | ^ |
*Nor **care** I **now**, how **dense** and **grim***

A D A D A D
^ | ^ | ^ |
*Di**sas**ters **gat**her **nigh**.*

Stanza 8:

A D A D A D A D
^ | ^ | ^ | ^ |
*I **care** not **in** this **mom**ent **sweet**,*

A D A D A D A D
^ | ^ | ^ | ^ |
*Though **all** I **ha**ve **ru**shed **o'er***

A D A D A D A D
^ | ^ | ^ | ^ |
*Should **come** on **pin**ion, **strong** and **fleet**,*

A D A D A D
^ | ^ | ^ |
*Pro**claim**ing **vengea**nce **sore**:*

Stanza 9:

A D A D A D A D
^ | ^ | ^ | ^ |
*Though **haug**hty **Hate** should **strike** me **down**,*

A D A D A D
^ | ^ | ^ |
*Right, **bar** app**roach** to **me**,*

A D A D A D A D
^ | ^ | ^ | ^ |
*And **grin**ding **Might**, with **furi**ous **frown**,*

A D A D A D
^ | ^ | ^ |
*S**wea**r **end**less **enmity**.*

Stanza 10:

A D A D A D A D
^ | ^ | ^ | ^ |
*My **love** has **plac**ed **her** little **hand***

A D A D A D
^ | ^ | ^ |
*With **no**ble **faith** in **mine**,*

A D A D A D A D
^ | ^ | ^ | ^ |
*And **vowed** that **wed**lock's **sac**red **band***

A D A D A D
^ | ^ | ^ |
*Our **nat**ure **sha**ll en**twine**.*

Stanza 11:

A D A D A D A D
^ | ^ | ^ | ^ |
*My **love** has **sworn**, with **seal**ing **kiss**,*

A D A D A D
^ | ^ | ^ |
*With **me** to **live**—to **die**;*

A D A D A D A D
^ | ^ | ^ | ^ |
*I **have** at **last** my **name**less **bliss**.*

A D A D A D
^ | ^ | ^ |
*As **I** love—**loved** am **I**!*

f. 'The Truest Love' in 'Jane Eyre' by Charlotte Brontë. [26]

The above passage may be characterized by the following matrix

$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Thus, the authorship may be evaluated by examining the pattern of entries in the transition matrix.

IV. Extending Memory Through State Integration

The basic Markovian representation, while useful, suffers from the limitations that plague all basic Markovian models in that the memory of any previous states is completely erased.

However, it is possible to inject limited memory by forming the Cartesian product of the basic states, but this tends to be computationally prohibitive. In our situation above, incorporating one-step memory requires augmenting the state space to 16, and incorporating two-step memory requires augmenting the state space to 64, resulting in a 64 × 64 transition matrix. Thus, incorporating k-step memory requires augmenting the state space to 4^{k+1} with a rather unwieldy $4^{k+1} \times 4^{k+1}$ transition matrix:

$$\begin{pmatrix} e_{11} & e_{12} & e_{13} & \dots & e_{1n} \\ \dots & \dots & & & \\ e_{i1} & e_{i2} & \dots & & e_{in} \\ \dots & \dots & & \dots & \dots \\ \dots & \dots & & & \\ e_{n1} & e_{n2} & \dots & & e_{nn} \end{pmatrix} \quad (1)$$

where each $e_{ij} \geq 0$, and

$$\sum_{j=1}^{n} e_{ij} \leq 1, \qquad i = 1, 2, \ldots, n.$$

Here, we allow the possibility of

$$\sum_{j=1}^{n} e_{ij} < 1,$$

which can sometimes occur, especially when absorbing states may be suitably identified [34]. Techniques for dimension reduction of the above matrix have been proposed in [1], where, for example, a 16 × 16 transition matrix may be reduced to a 6 × 6 matrix.

Instead of the above approach, however, we shall exploit the line structure in each stanza. Consider Stanza 4 of 'My Feet They Are Sore' in 'Jane Eyre', written by Charlotte Brontë above. We observe that line 1 and line 2 have the same structure, which we can represent by ξ, while line 3 and line 4 have the same structure, which we can represent by ζ. Thus, in Stanza 4, we have the transition structure

$\xi \rightarrow \xi$ (probability ⅓)
$\xi \rightarrow \zeta$ (probability ⅓)
$\zeta \rightarrow \zeta$ (probability ⅓)
$\zeta \rightarrow \xi$ (probability 0).

This gives the 2 × 2 transition matrix

$$\begin{pmatrix} ⅓ & ⅓ \\ 0 & ⅓ \end{pmatrix}$$

where the first state corresponds to ξ, and the second state corresponds to ζ.

Applying the same framework to the first stanza of Rochester's poem, we introduce a further state λ corresponding to the second line of Rochester's poem, which then yields the transition structure

$\lambda \rightarrow \xi$ (probability 3/7)
$\xi \rightarrow \xi$ (probability 0)
$\xi \rightarrow \lambda$ (probability 4/7)
$\lambda \rightarrow \lambda$ (probability 0)
$\lambda \rightarrow \zeta$ (probability 0)
$\zeta \rightarrow \lambda$ (probability 0).
$\zeta \rightarrow \xi$ (probability 0)
$\xi \rightarrow \zeta$ (probability 0)
$\zeta \rightarrow \zeta$ (probability 0)

This will give a 3 × 3 transition matrix with all diagonal elements equal to 0.

$$\begin{pmatrix} 0 & 0 & 4/7 \\ 0 & 0 & 0 \\ 3/7 & 0 & 0 \end{pmatrix}$$

While this is a more complete representation, it will be computationally less efficient and is only necessary when we need to combine the above two passages. In analyzing individual passages, using a 2 × 2 transition matrix may sometimes suffice, which on suitably defining the states, yields the following simpler transition matrix

$$\begin{pmatrix} 0 & 4/7 \\ 3/7 & 0 \end{pmatrix}$$

Following the simplified approach, the second stanza of the same poem yields the transition structure

$\zeta \rightarrow \xi$ (probability ⅔)
$\xi \rightarrow \xi$ (probability 0)
$\xi \rightarrow \zeta$ (probability ⅓)
$\zeta \rightarrow \zeta$ (probability 0).

This gives the transition matrix

$$\begin{pmatrix} 0 & ⅓ \\ ⅔ & 0 \end{pmatrix}$$

Another way to avoid dealing with large transition matrices, judgment can be exercised to integrate the states further. In this way, considering large passages will be less unwieldy. From the recital perspective, it may be possible to merge the states ζ and λ and regard them as serving the same purpose. Let us map both states to η. Then we have the states ξ and η. In so combining the two stanzas will yield the following transition structure

$\eta \rightarrow \xi$ (probability 6/11)
$\xi \rightarrow \xi$ (probability 0)
$\xi \rightarrow \eta$ (probability 5/11)
$\eta \rightarrow \eta$ (probability 0),

which yields the combined transition matrix

$$\begin{pmatrix} 0 & 5/11 \\ 6/11 & 0 \end{pmatrix} \quad (2)$$

Evidently, this matrix is computationally much more efficient than the matrix (1), while the amount of memory in the underlying Markovian model (2) is no less substantial.

V. EXPERIMENTATION

In this section, we conduct two experiments: monolingual and bilingual experiment. In a monolingual experiment, we analyse English stanzas written by Emily Brontë. Whereas in a bilingual experiment, we first analyse an English poem, 'Leda and the Swan' written by W. B. Yeats, then compare its Chinese versions translated by Mu Yang and Guangzhong Yu.

A. Monolingual Experiment

Ranked with Elizabeth Barrett Browning, Christina Rossetti, and Emily Dickinson, Emily Brontë is one of the pre-eminent women poets of the Victorian period [23]. As mentioned in the introduction of this paper, the poem, 'Stanzas' beginning 'Often rebuked, yet always back returning', is of uncertain authorship. Supposedly, it was written by Emily Brontë. Charlotte substantially revised and retitled most of these poems, and editors now print Emily's version from her manuscript (except in the case of 'Often rebuked'). Charlotte clarified in her prefatory note to 'Selections from the Poems by Ellis Bell' that 'it would not have been difficult to compile a volume out of the papers left by my sisters' [23].

Stanza 1:

*Often rebuked, yet always back retur**ning** (a)*

*To those first feelings that were born with **me**, (b)*

*And leaving busy chase of wealth and lear**ning** (a)*

*For idle dreams of things which cannot **be**: (b)*

Stanza 2:

*To-day, I will seek not the shadowy re**gion**; (c)*

*Its unsustaining vastness waxes dr**ear**; (d)*

*And visions rising, legion after le**gion**, (c)*

*Bring the unreal world too strangely n**ear** (d)*

Stanza 3:

*I'll walk, but not in old heroic tr**aces**, (e)*

*And not in paths of high morali**ty**, (f)*

*And not among the half-distinguished f**ace**s, (e)*

*The clouded forms of long-past histo**ry**. (f)*

Stanza 4:

*I'll walk where my own nature would be lea**ding**: (g)*

*It vexes me to choose another gu**ide**: (h)*

*Where the grey flocks in ferny glens are feed**ing**; (g)*

*Where the wild wind blows on the mountain s**ide**. (h)*

Stanza 5:

*What have those lonely mountains worth revea**ling**? (i)*

*More glory and more grief than I can t**ell**: (j)*

*The earth that wakes one human heart to feel**ing** (i)*

*Can centre both the worlds of Heaven and H**ell**. (j)*

g. 'Stanzas' begins with 'Often rebuked' by Emily Brontë, edited by Charlotte Brontë. [28]

Then, before a further elaboration of the remaining poem, Charlotte writes, 'The following are the last lines my sister Emily ever wrote.' [28]

No coward soul is mine,

No trembler in the world's storm-troubled sphere:

I see Heaven's glories shine,

And faith shines equal, arming me from fear.

O God within my breast,

Almighty, ever-present Deity!

Life—that in me has rest,

As I—undying Life—have power in thee!

Vain are the thousand creeds

That move men's hearts: unutterably vain;

Worthless as withered weeds,

Or idlest froth amid the boundless main,

To waken doubt in one

Holding so fast by thine infinity;

So surely anchored on

The stedfast rock of immortality.

With wide-embracing love

Thy spirit animates eternal years,

Pervades and broods above,

Changes, sustains, dissolves, creates, and rears.

Though earth and man were gone,

And suns and universes ceased to be,

And Thou were left alone,

Every existence would exist in Thee.

There is not room for Death,

Nor atom that his might could render void:

Thou—THOU art Being and Breath,

And what THOU art may never be destroyed.

h. 'Stanzas' begins with 'Often rebuked' by Emily Brontë, edited by Charlotte Brontë. [28]

The above poem is claimed to be written by Emily Brontë and edited by Charlotte Brontë. Nevertheless, the authorship is uncertain [26]. We analyse this piece of work as the dataset for comparison and evaluation. There is a possibility that the writers' writing style varies from time to time. Therefore, the loose verification of late Brontës may be significantly different from their early works [9].

Given a close look at the poem 'Stanzas', we can spot a similar pattern and rhythm in it. For example, the number of lines and the rhythm pattern in the above 'Stanzas' that begins with 'Often rebuked' written by Emily Brontë, edited by Charlotte Brontë is akin to the poem 'My Feet They are Sore' in 'Jane Eyre' written by Charlotte Brontë. They end with the 'abab, cdcd, efef, ghgh, ijij' pattern. More specifically, in the first stanza of the poem 'Stanzas', the rhythm of the first and third lines ends with '-ning', which is labelled by a random 'a', and the rhythm of the second and fourth lines end with '-e', which is marked as a sequential letter 'b'. The following rhythm in the 'Stanzas' edited by Charlotte Brontë is aligned with the pattern in her poems included in 'Jane Eyre'. Such identical authorship features can be easily found through Markov Model.

B. Bilingual Experiment

W. B. Yeats's poem 'Leda and the Swan' is based on a Greek story in which the god Zeus swooped down and hit Leda in the form of a swan, a human and ancient Greek queen. Consequently, such misconduct led to the Trojan War. Seemingly ironic, the poem could allude to the colonial relationship between Great Britain and Ireland, more specifically, to the Irish War for Independence.

Stanza 1:

Line 1: A sudden blow: the great wings beating still (k)

Line 2: Above the staggering girl, her thighs caressed (l)

Line 3: By the dark webs, her nape caught in his bill, (k)

Line 4: He holds her helpless breast upon his breast. (l)

Stanza 2:

Line 5: How can those terrified vague fingers push (m)

Line 6: The feathered glory from her loosening thighs? (n)

Line 7: And how can body, laid in that white rush, (m)

Line 8: But feel the strange heart beating where it lies? (n)

Stanza 3:

Line 9: A shudder in the loins engenders there (o)

Line 10: The broken wall, the burning roof and tower (p)

Line 11: And Agamemnon dead.

Line 12: Being so caught up, (q)

Line 13: So mastered by the brute blood of the air, (o)

Line 14: Did she put on his knowledge with his power (p)

Line 15: Before the indifferent beak could let her drop? (q)

i. 'Leda and the Swan' by W. B. Yeats. [31]

Analysing Yeats's poem through a human's close reading from a literature perspective, the poem consists of three concepts: defamiliarisation, paradox, and textual indeterminacies. The world would become strange due to the act of defamiliarisation. In the first stanza, 'sudden, still, staggering' suggests that the story takes place in *medias res*, defamiliarising the familiar. Through the angle of the third party, Yeats creates an impression to the readers who are not notified of anything that suddenly, everything is happening from nowhere. Because of the unawareness, audiences like the blind, who can only touch a part of an elephant at once, feel unfamiliar with this pornographic scene. Such a defamiliarised act 'presents objects or experiences from an unusual perspective or in unconventional and self-conscious language that our habitual, ordinary, rote perceptions of those things are disturbed [30]'. The estranging lines in the last stanza also indicate defamiliarisation that the poetic form and language are intentionally made strange after the 'Agamemnon dead', indicating the tragedy of such an irresponsible act of sexuality.

Power imbalance, *id est* who overpowers whom, remains an unsolved and ambiguous question for readers. The textual indeterminacy could lie in 'Leda's profound and provocative dramatisation of the ambiguities of sexual encounters' [29]. Conventionally speaking, 'helpless, terrified' could possibly imply Leda's despair because Zeus, in the form of a swan, rapes Leda. However, without pinpointing commas, it is not absolutely clear who is helpless. We may read that ambiguity as the basis of assuming interchangeable roles, both Leda and the swan being potentially the rapist [29]. Another indeterminacy example in stanza two leads to a question: whose 'body' feels whose 'strange heart beating', which resulting audiences' imagination of what suits their habitus.

The by-product of textual indeterminacy is the paradox regarding the definition of and separation between rightness and wrongness in Yeats' poem. The sexual intercourse between the helpless and the powerful may suggest the intertwining essence of the paradox between rightness and wrongness. The presumably mighty 'glory' and helpless Leda would never foresee that their sexual affairs would 'engender' the 'broken wall, burning roof' causing 'dead, blood'. By asking three questions (two questions begin with 'how' in lines 5-8, and one begins with 'did' in lines 14-15), Yeats implies that the natural bond and sexual desire from both parties are intertwined and inseparable, and so do the paradoxical rightness and wrongness in Zeus and Leda. As a human, Leda sees Zeus as an animal; thus, sex with such a 'feathered glory' is prohibited. She intends to 'push' him from 'her loosening thighs'. However, her feeling and, more importantly, Zeus's power takes over Leda, who is therefore subordinated by power, emotion, and sexuality.

From a machine learning perspective, the above-mentioned three key concepts can be captured and simplified by analysing the form of the original and translated texts. Here we provide two different versions in Chinese for the bilingual experiment.

The original English version of 'Leda and the Swan' (also called source text) has distinct machine-learning-friendly features, especially rhythming. For example, in stanza one, the ending rhythm for lines one and three is '-ill', which we name the 'k' line; in contrast, the second and fourth lines end with '-ssed/st', which we call it the 'l' line. We can see the pattern regularly goes in the 'klkl' pattern. A similar pattern and feature can be seen in the second and third stanzas, which end with the rhythm of 'mnmn' and 'opqopq', respectively. The random letter 'm' represents the rhythm of '-ush' in lines five and seven; 'n' refers to the identical ending rhythm of '-ighs/ies'.

Using the Markov model, if such a feature can be primed, detected, and memorised, then this model can be used to predict the translated text (also known as the target text). In other words, such rhythming features in the English source text should ideally be translated by a faithful translator to achieve the same effect in Chinese-translated poems.

遽然的垂擊：巨翼猶拍打於
暈眩無力的女子之上，她的雙股
被黑色的腳蹼撫弄，頸為喙所擒，
他把她無依的胸乳緊納入懷。

那些恐慌猶疑的手指怎麼可能
將插翼的光輝自漸漸鬆弛的股間推開？
而身體，在那白色的疾撞之下，
如何不覺察一奇異的心在那裡跳動？

腰際一陣戰慄於焉產生
是毀頹的城墻，塔樓熾烈焚燒
而阿加梅儂死矣。
被如此攫獲著，
如此被蒼天一狂猛的血力所制服，
她可曾利用他的威勢奪取他的洞識
在那冷漠的鳥喙廢然鬆懈之前？

j. 'Leda and the Swan' translated by Mu Yang. [32]

However, in the first piece of Chinese translation written by Yang [32], it is unlikely to see such features of rhythming if we look at the last words of each line. For example, in the first stanza, we have to end words pronounced respectively 'yu, gu, qin, huai', which do not comply with the source text's rules and writing convention, largely distinct from Yeats's writing habits. In the second stanza, the ending pronunciations are 'neng, kai, xia, dong', which are significantly different from the rhythm format in the source text. As the different rhythms cannot be memorized by a simple step of the Markov chain, the undetectability of the rhythm in Yang's translation could fail and an undesired translation from a machine learning perspective.

Stanza 1:

Line 1:猝然一攫：巨翼猶兀自拍動，(k')
Line 2:扇著欲墜的少女，他用黑蹼(l')
Line 3:摩挲她雙股，含她的後頸在喙中，(k')
Line 4:且擁她捂住的乳房在他的胸脯。(l')

Stanza 2:

Line 5:驚駭而含糊的手指怎能推拒，(m')
Line 6:她鬆弛的股間，那羽化的寵倖？(n')
Line 7:白熱的衝刺下，那撲倒的凡軀(m')
Line 8:怎能不感到那跳動的神異的心？(n')

Stanza 3:

Line 9:腰際一陣顫抖，從此便種下(o')
Line 10:敗壁頹垣，屋頂和城樓焚毀，(p')
Line 11:而亞加曼儂死去。
Line 12: 就這樣被抓，(q')
Line 13:被自天而降的暴力所凌駕，(o')
Line 14:她可曾就神力汲神的智慧，(p')
Line 15:乘那冷漠之喙尚未將她放下？(q')

k. 'Leda and the Swan' translated by Guangzhong Yu. [33]

On the contrary, Yu's version of translation [33] aligns with Yeats's rhythming format. Such rhythming features in Yu's translation could be memorized by the Markov model. For instance, in stanza one of Yu's translation, lines one and three end with the rhythm '-ong', and the rhythm of lines two and four are identical 'pu'. Likewise, the rhythming pattern in stanza two goes 'u' (lines five and seven) and 'xin' (lines six and eight), which suggests that Yu has captured the structure feature and beauty in Yeats's poem and then strategically and equivalently translated them into Chinese.

VI. CONCLUSION

In carrying out the analysis of the works in English and other literature, there are definite requirements, which include determining the authorship, dating the period of the work, and establishing the passage's style, theme, and purpose. We have placed such tasks in a machine-learning context, where a learning phase involving known passages is followed by a testing phase involving unknown passages.

Since obtaining such knowledge from human experts is time-consuming, subjective, and error-prone, we combine a data analysis approach with Markov models. The Markov model can represent and encapsulate the sequential flow of writing characteristics in passages. In addition, we also proposed two strategies to overcome the memoryless properties of the Markov model. The first involves augmenting the state space of the Markovian representation by repeatedly forming the Cartesian product of the underlying space. However, while this is a mathematically versatile approach, it will lead to high computational costs. The second approach exploits a poem's stanza and line structure, which has shown to be much more efficient, yielding a much smaller dimension transition matrix. In addition, the latter approach also lends itself to further state integration by judiciously analysing the purpose of each line of a passage and providing the scope for analysing much larger corpora. Through the appropriate use of Markovian variants, examining the pattern of probability entries in the transition matrix, and applying this characterisation to the vast body of English literature, more scientific, objective, and reliable decisions can be made concerning proper authorship, writing style, and other literary qualities.

REFERENCES

[1] C. H. C. Leung and C. Zeng. The Use of Multi-Step Markov Chains in the Characterization of English Literary Works. In Procedings of the 11th International Conference on Data Analytics, Valencia, Spain, pp. 43-48, 2022.

[2] C. Zeng and C. Leung, The Use of Stochastic Models in the Analysis of Vast English Literary Data Corpora, 2020 6th International Conference on Big Data and Information Analytics (BigDIA), Shenzhen, China, pp. 282-288, 2020.

[3] T. Plotz and G. A Fink, Markov Models for Handwriting Recognition. London: Springer London, 2011.

[4] G. T. Wright, *Shakespeare's Metrical Art.* Berkeley: University of California Press, 1988.

[5] W. Shakespeare and P. Alexander, *William Shakespeare; the complete works*. London: Collins, 1964.

[6] G. Taylor, Shakespeare and Others: The Authorship of "Henry the Sixth, Part One". *Medieval & Renaissance Drama in England*, Vol. 7 pp. 145-205. Rosemont Publishing & Printing Corp DBA Associated University Presses, 1995.

[7] C. Patrick, *Shakespeare's literary authorship*. Cambridge: Cambridge University Press, 2008.

[8] P. Edmondson and S. Wells, *Shakespeare Beyond Doubt*. Cambridge: Cambridge University Press, 2013.

[9] G. Taylor and G. Egan, *The New Oxford Shakespeare: Authorship Companion*. Oxford: Oxford University Press, 2017.

[10] E. Martina, The Use of Dialects and Foreign Languages in Shakespeare's King Henry V—Characteristics of the Fool Explored. *English Studies*, vol. 100, pp. 767-784. Colchester, Informa UK Limited, June 2019.

[11] D. Berend and A. Kontorovich, A Finite Sample Analysis of the Naive Bayes Classifier. *Journal of Machine Learning Research*, 16(1), pp. 1519-1545, 2015.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006.

[13] W. Feller, *Introduction to Probability Theory and Its Applications*, Volume I, 3rd. Ed. Wiley, 2008.

[14] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[15] D. D. Lewis and W. A. Gale, A Sequential Algorithm for Training Text Classifiers. In Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3-12, 1994.

[16] N. L. J. Kuang and C. H. C. Leung, Analysis of Evolutionary Behavior in Self-Learning Media Search Engines, in Proceedings of the IEEE International Conference on Big Data, Los Angeles, USA, pp. 643-650, 2019.

[17] N. L. J. Kuang and C. H. C. Leung, Performance Dynamics and Termination Errors in Reinforcement Learning - A Unifying Perspective. In Proceedings of the IEEE International Conference on Artificial Intelligence and Knowledge Engineering, pp. 129-133, 2018.

[18] E. Parzen. *Stochastic Processes*. Dover, 2018.

[19] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, Cheap and Fast - but is It Good? Evaluating Non-expert Annotations for Natural Language Tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 254-263, 2008.

[20] N. L. J. Kuang and C. H. C. Leung, Leveraging Reinforcement Learning Techniques for Effective Policy Adoption and Validation. in Misra S. et al. (eds) in Computational Science and Its Applications – ICCSA 2019, 311-322, Lecture Notes in Computer Science, Vol. 11620. Springer, 2019.

[21] N. L. J. Kuang and C. H. C. Leung, Performance Effectiveness of Multimedia Information Search Using the Epsilon-Greedy Algorithm, in Proceedings of the IEEE International Conference on Machine Learning and Applications, Florida, USA, pp. 929-936, 2019.

[22] W. Shakespeare, *The Sonnets.* London: Macmillan Collector's Library, 2016.

[23] C. Alexander and M. Smith *The Oxford Companion to the Brontës*. Oxford University Press, 2006.

[24] L. Pykett, *Emily Brontë*. Rowman & Littlefield Publishers, 1989.

[25] S. M. Gilbert and S. Gubar, *The madwoman in the attic: the woman writer and the nineteenth-century literary imagination.* Yale University Pres, 1979.

[26] C. Brontë, *Jane Eyre*, Oxford University Press, 2019.

[27] L. Judith and P. Christopher, From the Red Room to Rochester's Haircut: Mind Control in 'Jane Eyre'. *ESC: English Studies in Canada, 32*(4), pp. 169-188, 2008.

[28] A. Brontë, C. Brontë, and E. Brontë, *Poems by Currer, Ellis, and Acton Bell.* Project Gutenberg, 1997.

[29] WC. Barnwell, The Rapist in "Leda and the Swan". *South Atlantic Bulletin*, 42.1, pp. 62-68, 1977.

[30] J. Rivkin and R. Michael, *Literary Theory.* Blackwell Anthologies. 3rd ed. New York:Wiley, 2017.

[31] W. B. Yeats, Leda and the Swan. *Modern English Literature*, p.8, 1935

[32] M. Yang (楊牧), Translated poem of 'Leda and the Swan' 譯作《麗達與天鵝》. *Yi Shi (譯事)*. Cosmos Books Ltd. (天地圖書有限公司), p. 26, 2007.

[33] G. Yu (余光中), Translated poem of 'Leda and the Swan' 譯作《麗妲與天鵝》. *Songs of Innocence: An Anthology of Guangzhong Yu's Translated Poems* (天真的歌 : 余光中經典翻譯詩集). Jiangsu Phoenix Literature and art publishing, Ltd. (江蘇鳳凰文藝出版社), p. 76, 2019.

[34] R. A. Howard, Dynamic Probabilistic Systems, Volume I, Dover, 2017.

[35] T. Reedy and D. Kathman, *How We Know That Shakespeare Wrote Shakespeare: The Historical Facts*. Kathman & Ross, Shakespeare Authorship Page, 2005.

[36] M. Koppel, J. Schler, and S. Argamon, Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 54(4), pp. 9-26, 2009.

[37] M.B. Malyutov, Authorship attribution of texts: A review. *General Theory of Information Transfer and Combinatorics*, pp.362-380, 2006.

[38] S. Segarra, M. Eisen, and A. Ribeiro, Authorship attribution through function word adjacency networks. *IEEE Transactions on Signal Processing*, *63*(20), pp. 5464-5478, 2015.