

Tracing Thematic Shifts in Literature Translation: A Topic Modeling Analysis of Yu Hua's *To Live*

Zhengyang Wang^{1, a}

¹Translation and Interpreting, University of Stirling, United Kingdom

^awangzyalex@outlook.com

Abstract

This study uses Latent Dirichlet Allocation (LDA) topic modeling to compare thematic structures in Yu Hua's novel *To Live* and its English translation, examining how themes are preserved or altered. After digitizing and preprocessing both texts, separate LDA models yielded six topics for the Chinese original and five for the English version. Key findings show high fidelity for core themes like war and family. For example, war-related topics strongly aligned across versions. However, secondary or culturally specific themes, such as gambling, were diluted or merged in the English translation, while familial themes were consolidated. The research concludes that translation reframes narratives by altering lexical and thematic organization. Topic modeling effectively highlights these shifts, underscoring the translator's role as a narrative architect and offering a valuable method for digital literary and corpus-based translation studies.

Keywords

Topic Modeling; Thematic Analysis; *To Live*; Literature Translation.

1. Introduction

Over the last few years, literature studies in translation have increasingly made use of empirical and computational methods to understand how structure and theme get carried across different languages more effectively [1]. Qualitative methods like close reading and textual comparison have been useful in illuminating changes at the level of tone, register, and cultural detail. But as texts get digitized and data-driven applications become increasingly available, there is an increasing demand for making use of quantification to analyze literature in translation at the level of structure as well as theme [2].

This article centers on the English translation of Yu Hua's novel *To Live* (《活着》). The novel charts the life of Xu Fugui, once a landowner, as he endures a sequence of personal and historical misfortunes, from the Chinese Civil War right up to the Great Leap Forward and the Cultural Revolution. In its austere narrative mode and affective writing, *To Live* is a searing meditation on survival, loss, and human endurance [3]. Michael Berry's English translation in 2003 made the novel known worldwide for the first time, but it has been noted for its accessibility as well as for its emotional accuracy [4][5]. Questions, however, remain as to how the novel's central themes are reconfigured or reweighted in translation.

For this purpose, this study uses topic modeling in the form of Latent Dirichlet Allocation (LDA) as an algorithmic method for uncovering thematic patterns in the target as well as the source texts [6]. Using the pre-translational Chinese novel as one corpus and the post-translational English version as another, this approach makes it possible for systematic comparison of how the themes are maintained, lost, or restructured across linguistic as well as cultural borders [7]. The central research questions are as follows:

What are the main thematic clusters present in the Chinese original and the English translation of *To Live*, as identified through topic modeling?

To what extent are these themes preserved or altered in translation?

What types of thematic shifts—such as merging, omission, or re-weighting—can be observed, and what might they indicate about the translational approach?

Unlike analysis based on an explicit theoretical foundation, this research features an approach-oriented focus in which the findings are allowed to be uncovered within the data rather than derived from an explicit interpretive framework. Focusing on lexical salience and topical cohesion, the objective is to determine if the emotional/narrative center of the novel is preserved in the translation, as well as how particular culture-specific sub-themes are managed. This work is an addition to the expanding area of digital literary studies as well as corpus-based analysis of translation studies. It illustrates how unsupervised machine learning may be utilized in literary texts in order to reveal patterns that may not be easily observed through conventional reading. It thereby proposes an extensible as well as reproducible protocol for subsequent comparable studies in literature translation, especially in the case of long narratives as well as intricate thematic structures.

2. Methodology

2.1. Corpus Preparation

The source text consists of Yu Hua's *To Live* in its original Chinese version (approx. 55,000 characters), and the target text is Michael Berry's English translation (approx. 60,000 words), excluding paratextual material such as forewords or afterwords. Both texts were digitized and cleaned to facilitate topic modeling.

For the Chinese corpus, word segmentation was performed using the jieba Python library, as Chinese text lacks inherent word boundaries. Stopwords, punctuation, and proper names (e.g., character names) were removed to minimize noise. For the English corpus, tokenization was performed using nltk, with all words converted to lowercase. Stopwords (e.g., "the," "and") and punctuation were removed, and stemming was applied using nltk to reduce inflectional forms. The preprocessing pipeline was implemented in Python 3.13, using standard NLP libraries including:

jieba for Chinese word segmentation

nltk for English tokenization and cleaning

scikit-learn for LDA model training

pyLDAvis for topic visualization

wordcloud for generating word cloud diagrams

Each novel (Chinese and English) was treated as a single document for modeling purposes, given the focus on macro-level thematic structure rather than paragraph-level granularity.

2.2. Topic Modeling Procedure (LDA)

Latent Dirichlet Allocation (LDA) is utilized to model topics in each corpus independently. All modeling was conducted using Python's scikit-learn package. After experimenting with different numbers of topics and evaluating performance via perplexity and topic coherence measures, topics number K=6 produced coherent and interpretable results for the Chinese texts, and K=5 for the English texts.

The LDA models were trained under the following conditions:

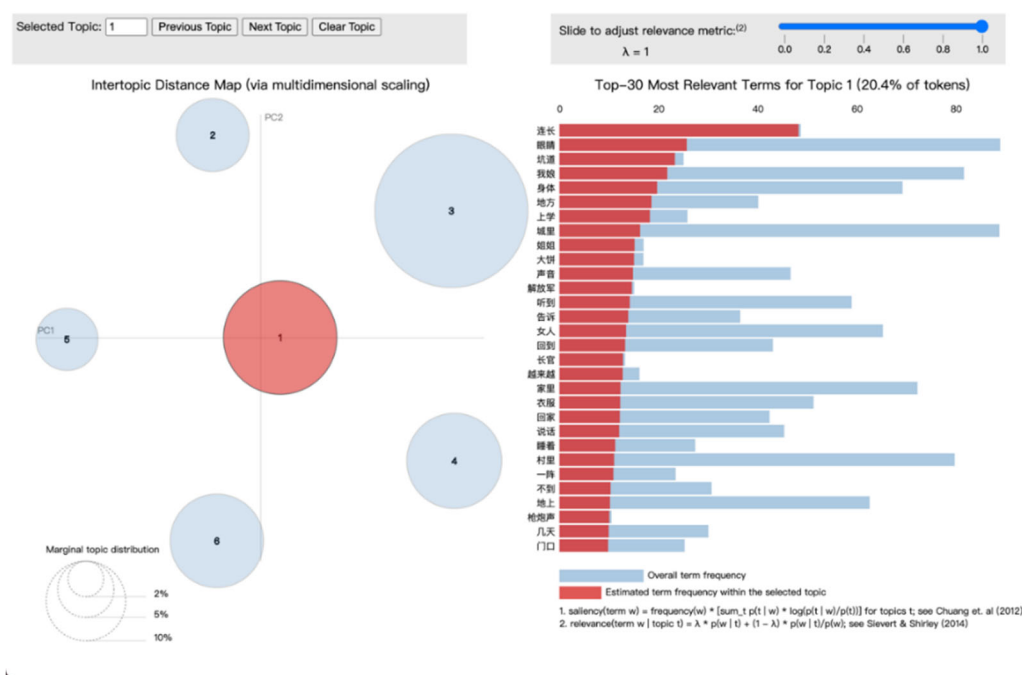
Number of Topics (K): 6/5

Learning Method: 'batch'

Maximum Iterations: 20

Hyperparameters: The Dirichlet priors were set as `doc_topic_prior=0.1` (alpha) and `topic_word_prior=0.01` (beta) to allow the model to adjust topic sparsity.

Other Settings: The model was run for 1000 iterations, with appropriate settings for randomness and parallel processing. Some other python packages are used to visualize the results, e.g. pyLDavis showing the model in a html file (Graph 1 is a screenshot of the file), and wordcloud showing the topics in word cloud diagram (see Graph for some Chinese topics and Graph for English).



Graph 1: topic model visualization with pyLDavis

After obtaining, the topics are labeled by analyzing the top words and cross-referencing with the novel's content. For example, if a Chinese topic's top words included “连长” (company commander), “医生” (doctor), “坑道” (tunnel), this topic would be recognized this as relating to the wartime and military hospital scenes, and label like “War and Army.” English topics were done the same, e.g., an English topic with “commander, soldier, liberation, bullet, army, tunnel” clearly corresponded to war.

3. Results and Discussion

This section presents the results of the topic modeling analysis and discusses key findings regarding thematic preservation and transformation in the English translation of *To Live*. Rather than offering a one-to-one mapping of all extracted topics, the discussion focuses on a few representative cases to illustrate patterns of alignment, merging, and lexical dilution across the Chinese and English versions. These patterns reveal not only what themes are emphasized or backgrounded, but also how translation choices may influence thematic accessibility for the target audience.

3.1. Comparison of Topics in the Original and Translation

The LDA models produced six topics from the Chinese original and five from the English translation. Each topic was represented by its top 15 high-probability keywords, and thematic

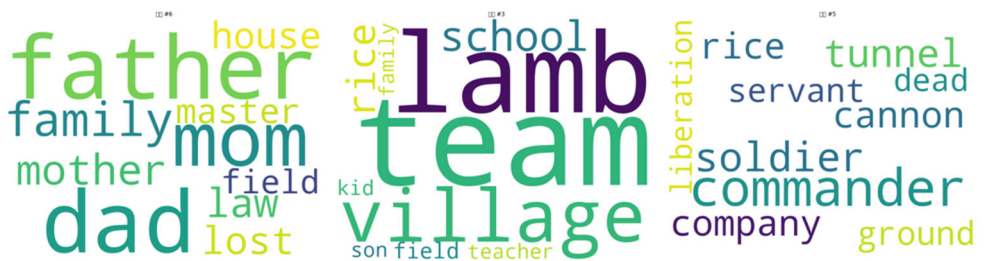
labels were assigned through interpretive reading of the associated content. **Table 1** summarizes the extracted topics.

Table 1. LDA Topics from Chinese and English Texts (Simplified)			
Corpus	Topic ID	Label	Top Keywords
Chinese	C1	War and Military	连长, 医生, 坑道, 解放军, 大饼
	C2	Gambling and Old Life	骰子, 青楼, 铜钱, 我爹, 少爷
	C3	Family and Home	儿子, 家里, 眼泪, 村里, 干活
	C4	Collectivization and Hardship	食堂, 人民公社, 力气, 钢铁, 汽油桶
	C5	Generational Change and Tradition	老人, 私塾, 光耀, 长衫, 野菜
	C6	Marriage and Women	新娘, 女人, 衣服, 肩膀, 我娘
English	E1	Family and Home	father, mother, house, family, land, ox
	E2	Community and Daily Life	team, village, rice, town, kid, sick
	E3	Death and Suffering	dead, life, door, body, feel, tear
	E4	War and Military	commander, bullet, tunnel, army, flatbread
	E5	Descriptive Odds and Ends	lucky, dressed, ash, expert, blankly

While the core themes like war, family, and suffering are consistently present, differences emerge in how secondary or culturally embedded topics are treated. The following subsections explore three representative cases.



Graph 2: Chinese topic C1 C2 C3



Graph 3: English topic E1 E2 E4

3.2. Analysis of Thematic Transfer and Shifts

Case 1: Strong Cross-Lingual Preservation - War and Survival

The war and military theme is the most clearly aligned between both versions. In Chinese Topic C1, terms such as “连长” (company commander), “医生” (doctor), “坑道” (tunnel), and “解放军”

(People's Liberation Army) dominate. These reflect the scenes where Fugui is conscripted, injured, and interacts with soldiers and medics during combat.

English Topic E4 reflects this with similarly high-weight terms: "commander," "soldier," "bullet," "tunnel," "liberation," and "flatbread." The appearance of culturally specific items such as "flatbread" suggests a conscious effort to retain unique details in translation. The word "liberation" may be a simplified rendering of "解放军," preserving ideological implications while keeping terms familiar to English readers.

This preservation may be attributed to the universality of wartime narrative. The translator likely recognized the high narrative weight of these scenes and rendered them with minimal adaptation, trusting that the underlying emotions like fear, hunger, trauma would resonate regardless of cultural background. War vocabulary also tends to be relatively direct, reducing the risk of semantic loss in translation.

Case 2: Thematic Merging - Family, Labor, and Emotional Loss

The theme of family and domestic life presents a more complex case. In the Chinese corpus, this theme is distributed across at least three topics (C3, C6, and partially C2), reflecting distinct dimensions: parent-child relationships, gender roles, marriage customs, and emotional responses to loss. Frequent terms include "儿子" (son), "眼泪" (tears), "新娘" (bride), and "家里" (home).

In contrast, the English model merges these into Topic E1, combining terms like "father," "mother," "house," "land," and "ox." Interestingly, this cluster also includes economic indicators such as "money" and "field," suggesting that the translated narrative tightly binds familial experience with economic decline. The symbolic use of the ox, Fugui's final companion, reinforces the intersection of family, material deprivation, and personal identity.

Here, lexical cohesion is achieved through generalization. While the Chinese version makes frequent use of relational markers ("我娘," "儿子," "丈人"), English naturally favors pronouns or varied expression ("he," "his mother," "the boy"), resulting in lower term recurrence. This affects the LDA model's clustering behavior, leading to broader, more inclusive topics.

Thematically, the consolidation simplifies the emotional arc. English readers may experience the story as one continuous narrative of loss and struggle, whereas Chinese readers may perceive distinct emotional episodes—losing a son, marrying off a daughter, caring for aging parents. This shift does not dilute meaning but alters its narrative rhythm.

It is also worth noting that emotional terms like "tear," "lost," and "feel" appear more prominently in the English E3 topic ("Death and Suffering") than in the Chinese equivalents. This may reflect a stylistic emphasis on affect in English storytelling, possibly amplified during translation to make the narrative emotionally accessible to a non-Chinese audience.

Case 3: Thematic Suppression - Gambling and Decline

The clearest example of thematic suppression or dilution is the portrayal of Fugui's youthful hedonism and financial ruin. In Chinese Topic C2, a dense cluster of culturally rich terms appears: "骰子" (dice), "青楼" (brothel), "铜钱" (copper coin), "少爷" (young master), "沈先生" (Mr. Shen), and "赌博" (gambling). These establish the social and moral context of Fugui's early fall.

In the English corpus, however, there is no topic dominated by gambling-related words. While E5 contains vague terms such as "lucky," "raggedy," and "cordial," these lack sufficient cohesion to constitute a thematic cluster. Term frequency analysis confirms this: "gamble" and "dice" appear fewer than five times in the entire English text, compared to dozens in the Chinese version. Cosine similarity between C2 and any English topic is below 0.3.

This absence does not imply outright omission. Rather, the translator likely treated these elements as narrative setup, necessary but not central. The language may have become more

narrative and less terminologically marked. For example, a sentence like “我爹被我赌输了家业” might be rendered as “I lost everything, even my father’s trust,” eliminating repeatable tokens like “gambling.”

Possible motivations for this reduction include:

Narrative compression: The gambling phase is confined to early chapters and does not recur later in the novel.

Cultural sensitivity: Terms like “brothel” and “young master” may lack clear Western analogs or risk exoticizing the text.

Stylistic flattening: English literary norms often favor streamlined moral arcs over episodic moral lessons.

Thematic loss here is subtle but significant. Chinese readers encounter a clear arc of moral downfall and karmic retribution. English readers may perceive *Fugui* as unlucky rather than self-destructive, a shift that alters the novel’s moral framing.

3.3. Emergent Patterns and Translational Implications

The three examples above illustrate broader tendencies observed throughout the modeling results:

Table 2. Broader tendencies

Pattern	Evidence Example	Implication
Preservation	War theme: C1–E4 match (cosine similarity: 0.67)	Universality of suffering aids translation
Merging	Family split in Chinese (C3, C6) merged into E1	Lexical economy and narrative flow
Dilution	Gambling present in C2 but not in English	Low repetition and cultural filtering

These shifts are not simply losses or gains. They reflect how translation, even when faithful in content, re-structures meaning at a thematic and lexical level. Topic modeling reveals patterns that traditional close reading might overlook, not just what is translated, but how it is lexically distributed, prioritized, and narratively organized.

4. Conclusion

This study investigated the thematic transmission between the Chinese original and the English translation of Yu Hua’s *To Live* using Latent Dirichlet Allocation (LDA). The results offer insights into how translation reshapes literary meaning not just at the sentence level, but at the scale of narrative structure.

4.1. Major Findings

At a macro level, topic modeling reveals a high degree of thematic fidelity. Core themes such as war, family, survival, and death were strongly represented in both the Chinese and English corpora. For instance, war-related topics in both versions shared numerous high-frequency lexical items (e.g., “commander,” “soldier,” “tunnel”), indicating that experiences grounded in physical and emotional universality tend to be faithfully preserved during translation.

However, the analysis also uncovered important structural shifts in how secondary or culturally embedded themes were represented. Topics such as gambling, collectivization, and gender roles, which appeared as distinct and lexically cohesive themes in the Chinese original, were either diluted, merged, or lexically generalized in the English translation. This suggests a

reconfiguration of narrative emphasis: while major emotional arcs are retained, some socio-cultural nuances may be flattened in the process.

4.2. Interpretive Implication

These shifts do not necessarily reflect a loss of content, but rather a form of narrative reframing. Thematic merging, for instance, is consistent with English literary norms that prioritize cohesive emotional arcs over episodic or role-based narration. Additionally, the de-emphasis of culturally specific lexical items may reflect pragmatic translation strategies, such as simplification or normalization, designed to improve accessibility for global readers unfamiliar with the Chinese historical context.

Topic modeling provides a valuable lens for observing these transformations. It not only quantifies changes in lexical salience but also helps uncover implicit shifts in narrative framing and emotional pacing that would be difficult to detect through close reading alone.

4.3. Methodological Contributions

This study also demonstrates the potential of LDA-based topic modeling as a tool for comparative literary analysis. When used in conjunction with contextual interpretation, topic models allow researchers to identify:

Which themes are foregrounded or backgrounded across languages;

How lexical frequency contributes to thematic visibility;

Where translation may re-structure rather than merely transfer narrative meaning.

By treating entire novels as analyzable corpora, this approach offers a scalable, replicable, and language-agnostic methodology—especially useful for long-form literary works.

4.4. Future Directions

Several extensions of this research are possible. Future studies could:

Apply the same method to multiple translations of the same work (e.g., different English versions, or into other languages);

Integrate other computational techniques, such as sentiment analysis or named entity tracking, to enrich theme-emotion mappings;

Combine modeling with reader reception studies, to evaluate how thematic restructuring affects interpretation across cultural boundaries.

Such research would contribute to a broader understanding of how translation mediates not just language, but also cultural resonance and thematic hierarchy.

In conclusion, this study shows that while meaning is largely preserved across translation, it is also repackaged, sometimes subtly, sometimes significantly, through changes in lexical emphasis and thematic organization. Computational methods like topic modeling can help us make these processes visible. By highlighting what remains constant, what shifts, and what disappears, we can better understand the translator's role not just as a linguistic mediator, but as a narrative architect.

References

- [1] Krasnov F, et al. Clustering of Translation via Topic Modeling[C]. Big Data and AI Conference, Journal of Physics: Conference Series, IOP Publishing, 2019.
- [2] Li D, et al. Applying Topic Modeling to Literary Analysis: A Review[J]. Digital Studies in Language and Literature, 2024, 1(1-2): 113-141.
- [3] Chang Q, et al. Suffering and Redemption: The Two Major Thematic Discourses in Yu Hua's Novels of the 1990s[J]. Journal of Huazhong University of Science and Technology (Social Science Edition), 2001, (2): 96-101.(in Chinese)

- [4] Zu D, Xu Y. Evaluation of Michael Berry's Translation of *To Live* from Metafunctional Approach[J]. *Theory and Practice in Language Studies*, 2017, 7(4): 270-279.
- [5] Lu H. An Analysis of Translation Methods for Culture-Loaded Words in the English Translation of *To Live*[J]. *Modern Linguistics*, 2023, 11(12): 5698-5702.(in Chinese)
- [6] Blei D M, et al. Latent Dirichlet Allocation[J]. *Journal of Machine Learning Research*, 2003, 3: 993-1022.
- [7] Lefevere A. *Translation, Rewriting, and the Manipulation of Literary Fame*[M]. London/New York: Routledge, 1992.