A Comparative Study of Word-Formation Patterns in Chinese and Japanese Internet Slang - A Computational Linguistics Perspective

Pengfei Gao*
Institute of Foreign Languages, Tianjin University of Science and Technology, Tianjin 300457, China
*Corresponding email: fire0818@mail.tust.edu.cn

Abstract

This study constructs a systematic and quantifiable comparative framework to investigate the word-formation patterns of Internet slang in China and Japan by integrating quantitative text analysis with pre-trained language models (PLMs). Utilizing computational linguistics tools, we conducted fundamental metric analyses - including word frequency distribution, co-occurrence network modeling, and correspondence analysis - on large-scale corpora harvested from prominent social media platforms. Furthermore, the research introduces pre-trained models to transform core lexical items into high-dimensional semantic vectors, enabling advanced computations such as semantic similarity matrix construction, hierarchical clustering, and dimensionality reduction visualization. These methods allow for a deep exploration of the underlying semantic generation and association mechanisms of online neologisms. By comparing the divergences between the two, this study reveals the similarities and differences in linguistic innovation paths within Chinese and Japanese cyber-contexts, providing an empirical analytical framework for applying computational linguistics to the comparative study of dynamic sociolinguistic phenomena.

Keywords

Internet slang, Sino-Japanese comparison, Word formation, Computational linguistics, Pre-trained language models, Semantic vectors

Introduction

Research background and problem statement

In the digital epoch, social media platforms have evolved into the most active frontiers for linguistic innovation. Platforms such as Weibo, Douyin, X (formerly Twitter), and TikTok have not only reshaped interpersonal interaction but have also catalyzed the birth of highly resilient linguistic variants - Internet slang - that are inherently tethered to cyberspace. These lexical items often erupt, propagate, mutate, and eventually fade at an astonishing velocity. Their generative mechanisms are flexible and diverse, characterized by the recontextualization of archaic terms, cross-domain blending, or idiosyncratic abbreviations, all of which vividly reflect collective social mentalities, cultural flashpoints, and youth subculture trends during specific periods. Due to the widespread dissemination of Internet slang words, it is necessary to detect them by the word form features and a variety of semantic-changed registered words in social media texts into accounts [1].

Although China and Japan share the East Asian cultural sphere, they possess radically different linguistic typologies. Chinese is a quintessential isolating language, lacking morphological inflection and relying heavily on word order and functional particles with logographic characters at its core. Conversely, Japanese is an agglutinative language with a sophisticated system of conjugations and particles, employing a hybrid writing system of Kanji and Kana. Investigating how these distinct linguistic systems generate creative responses to the immediacy, entertainment-oriented, and communal demands of online communication - and whether their word-formation mechanisms exhibit systematic typological differences - is of profound academic value. Such an inquiry not only deepens our understanding of linguistic adaptability and creativity but also provides a unique perspective on the digital interactive modes of these two socio-cultural entities.

However, traditional comparative linguistic research, particularly regarding the dynamic phenomenon of Internet slang, often encounters methodological limitations. Existing scholarships tend to focus on etymological anecdotes, semantic interpretations, or cultural analyses of individual terms, primarily utilizing qualitative descriptions and illustrative examples. While these studies offer valuable case-level insights, they struggle to capture the overall distribution and structural characteristics of word-formation patterns at a macroscopic level, and they lack objective metrics for quantifying linguistic differences. Meanwhile, the rapid advancement of computational linguistics, especially the emergence of deep-learning-based pre-trained language models, has initiated a paradigm shift in linguistic research. These models can autonomously induce complex linguistic regularities from massive datasets and map the semantics of words and phrases into high-dimensional mathematical spaces, transforming them into computable and comparable vectors. This provides unprecedented technical possibilities for the refined, empirical comparative analysis of large-scale linguistic phenomena. Regrettably, there remains a lack of research that systematically applies this data-driven computational paradigm to compare the non-standard morphological mechanisms of Chinese and Japanese Internet slang.

Research objectives and significance

Building upon the aforementioned background, the core objective of this study is to construct an analytical framework that fuses classical linguistic theory with modern computational methods. This framework is designed to conduct a systematic, quantifiable, and in-depth comparative exploration of Chinese and Japanese Internet slang word-formation patterns. Specifically, the study aims to achieve three levels of objectives. At the descriptive level, it seeks to move beyond subjective enumeration by objectively presenting the distribution frequency and preferential characteristics of major word-formation types, such as clipping, compounding, derivation, semantic shift, and borrowing. At the explanatory level, the study leverages the robust semantic representation capabilities of pre-trained models to penetrate surface morphological forms. It analyzes the internal associations, clustering structures, and generative logic of these terms within semantic vector spaces to reveal the underlying semantic operations driving formal differences. At the methodological level, it aims to demonstrate a viable technical path that organically combines macro-textual metric analysis tools (such as KH Coder) with micro-semantic mining models (such as Sentence-BERT). This establishes a reproducible and scalable empirical paradigm for future cross-linguistic dynamic vocabulary research.

The value of this research is manifested across multiple dimensions. Theoretically, it seeks to test and deepen existing hypotheses regarding how typological differences between Chinese and Japanese constrain and shape linguistic innovation from a computational empirical perspective. It provides empirical evidence from the emerging register of online society for theoretical linguistics. Methodologically, it represents a meaningful interdisciplinary attempt to facilitate a deep dialogue between the problem consciousness of sociolinguistics and comparative linguistics and the advanced tools of computational linguistics, potentially broadening the technical boundaries of linguistic inquiry. In terms of application, the findings offer insights into second language acquisition, cross-cultural communication, social sentiment analysis, and natural language processing. For instance, a clear understanding of the divergent word-coining logic between China and Japan can help language learners more deeply comprehend the contemporary culture and cognitive habits of the target country. Meanwhile, the identified morphological and semantic patterns can assist search engines and machine translation systems in better processing Internet neologisms and out-of-vocabulary terms.

Literature review

Existing research on Chinese and Japanese Internet slang

Domestic research on Internet language in China began at the turn of the century, with early work focusing on the collection, organization, and preliminary socio-cultural interpretation of emerging cyber-vocabulary, emphasizing its characteristics as a social dialect and its impact on modern Chinese. Subsequent research has gradually delved into the transmission mechanisms and discursive functions of specific terms, as well as the

youth subcultures behind them. The Japanese academic community has an even longer history of focusing on Shingo/Ryukogo (new and vogue words), which in the modern era is closely linked to studies on Wakamono-kotoba (youth speak) and Internet slang. It can be seen that the word formation used in China's interesting compound word is "word + word", and it is also applicable in Japan. To this point, kango (words deriving from Chinese) is used more frequently in Japan's compound words. It has been rare to see interesting utterances in Japan with the form of "wago (words deriving from Japanese) + wago" during the last five years, and the form of "others + kango" is more popular [2]. Scholars have continuously tracked annual vogue words through yearbooks, while sociolinguists have analyzed social stratification and intergenerational differences from the perspective of Isōgo (phase words). These studies have accumulated an exceptionally rich repository of case materials and cultural interpretations, forming an essential intellectual foundation for this study. However, a significant deficiency lies in the fact that the vast majority of research remains confined within a single linguistic system or performs only superficial cultural analogies. Although a few studies have attempted simple comparisons between Chinese and Japanese vogue words, most remain at the level of phenomenological listing, lacking a unified theoretical framework based on linguistic ontology - such as morphology - for systematic structural comparison. Furthermore, there is a notable absence of empirical research utilizing large-scale corpora and quantitative methods to verify differences in word-formation patterns.

Morphological theories

Morphology, a core branch of linguistics, investigates the rules and patterns by which new words are generated. Chinese morphological research traditionally emphasizes compounding (including coordinate, subordinate, verb-object, subject-predicate, and complementary structures), derivation (affixation), conversion (zero-derivation), and abbreviation (clipping and contraction). Since Chinese lacks morphological inflection, its word-formation logic relies heavily on semantic association and syntactic relationships. Japanese word formation presents a different landscape; its advanced agglutinative nature makes derivation - through the addition of prefixes and suffixes with strong grammatical or semantic functions - exceptionally active. Furthermore, compounding, hybridity (such as Wasei-Kango and Wasei-Eigo), clipping, and the large-scale borrowing and assimilation of loanwords are all highly productive morphological means. Japanese language in general, clipping of Sino-Japanese compounds and noun of foreign origin in common. Japanese internet slang also clips adjectives, adverbial phrase, and even idiomatic expression [3]. These classical theories provide the indispensable taxonomic lens and analytical categories for decoding the "word-coining cipher" of Internet slang. However, as online language frequently breaks conventional rules to produce "non-standard" or "marginal" formations, the application of traditional theory must remain flexible and open, potentially requiring the creation of new sub-categories to accommodate vivid linguistic facts.

Relevant applications in computational linguistics

In recent years, the application of computational linguistic methods has seen explosive growth. At the macro-level of text analysis, text mining and scientific mapping tools represented by KH Coder and VOS viewer have been widely applied in academic literature analysis, news sentiment research, and social media analysis. These tools efficiently process large-scale texts to reveal latent thematic structures and conceptual evolutions through word frequency statistics, co-occurrence network analysis, and correspondence analysis. At the micro-level of semantic analysis, static word embedding models such as Word2Vec and GloVe, alongside context-aware pre-trained language models like BERT and the GPT series, have revolutionized lexical semantic research. The "semantic vectors" they generate enable the precise measurement of semantic similarity, semantic analogical reasoning, and the detection of historical semantic shifts. However, a significant interdisciplinary gap exists: the integration of macro-structural metric analysis with micro-semantic deep learning models specifically focused on the comparative word formation of Chinese and Japanese Internet slang. Existing computational research tends to focus on single languages or general domains, leaving this dynamic sociolinguistic phenomenon largely unexplored through a hybrid computational framework. This study targets

this intersection, aiming to contribute both methodologically and empirically.

Theoretical framework and methodology

Theoretical framework and core concepts

The execution of this study is built upon two key theoretical pillars. The first pillar is linguistic word-formation theory. TikTokers tend to ignore the rule of word formation. It can be concluded that the aim of different types of word formation processes assisted people to understand the function in order to avoid mistakes in their writing [4]. By synthesizing classical categories of Chinese and Japanese morphology and tailoring them to the specificities of Internet language, we have constructed an operational analytical framework comprising five primary categories for the classification of selected slang terms: Clipping/Abbreviation (e.g., Chinese yyds, Japanese tsura), Compounding/Derivation (e.g., Chinese Tianhuaban, Japanese Bocchizatsu), Semantic Shift/Extension (e.g., Chinese Pofang, Japanese Kusa), Borrowing/Hybridization (e.g., Chinese Shuan Q, Japanese Guguru), and Functional Conversion (particularly prominent in Japanese, such as nominalization via the suffix -mi). This framework serves as the foundation for our formal comparison.

The second, more innovative pillar is the "Distributional Hypothesis" from computational linguistics and its technical implementation. The distributional hypothesis posits that the meaning of a word can be defined by the patterns of its distribution across different contexts. The dynamic nature of language, particularly in the realm of internet slang and memes, poses significant challenges for the adaptability of large language models (LLMs) [5]. Pre-trained language models, such as Sentence-BERT, represent a superior engineering realization of this hypothesis. By undergoing self-supervised learning on massive corpora, these models automatically encode every word or phrase as a dense vector in a high-dimensional space (e.g., 768 dimensions), known as a "semantic vector". This vector is designed to capture the deep semantic information of the term. Crucially, within this vector space, semantically similar words are geometrically close to one another. The semantic similarity between two terms can be precisely quantified by calculating the cosine value of their vectors (ranging from -1 to 1). This provides us with a "semantic ruler" that transcends surface orthographic differences to directly measure and compare the conceptual proximity of Chinese and Japanese slang, which can then be projected onto a two-dimensional plane via dimensionality reduction for visual observation of internal clustering patterns.

Overall research design

To achieve the research objectives, we adopt a progressive "mixed methods" design. The entire research process is divided into two primary, interconnected stages. The first stage is "Macro-Metric and Associative Analysis", where the core task is to utilize KH Coder for fundamental, exploratory data analysis of our constructed corpora. We systematically generate high-frequency word lists, compare part-of-speech distributions, and construct "co-occurrence networks". Co-occurrence analysis is based on the principle that words appearing frequently within the same textual window are likely to be semantically related. By visualizing these networks and calculating graph-theory metrics such as density, centrality, and modularity, we can grasp the macro-structural morphology of the slang semantic fields in both countries. This allows us to determine whether they cluster around a few core concepts or form a multi-centric, dispersed structure.

The second stage is "Micro-Semantic and Vector Space Analysis", which transitions to a Python programming environment utilizing the sentence-transformers library to invoke pre-trained multilingual Sentence-BERT models. We select dozens of representative terms from each language that best reflect various word-formation types for in-depth analysis. The model generates corresponding semantic vectors for each term, after which we conduct a series of vector-based computational experiments. These include calculating a "semantic similarity matrix" for all word pairs to compare structural differences between the Chinese and Japanese datasets and performing unsupervised K-means clustering to observe if the model automatically groups terms with similar cross-lingual concepts (such as negative emotions or social phenomena). Finally, the t-SNE algorithm is used to reduce high-dimensional vectors to two or three dimensions. This results in a "semantic space map" that visually demonstrates the distribution and relative positioning of Chinese and Japanese slang in a unified

space, revealing whether they are distinct or overlapping. The results of these two stages are integrated to provide a multi-dimensional comparative understanding of word-formation patterns.

Analysis and core findings

Data collection and preprocessing

To ensure the timeliness and representativeness of the research, we selected the 2022 to 2024 as the temporal window. The core data sources are the most authoritative annual vogue word lists from both countries: China's "Top Ten Vogue Words of the Year" by Yiuwen Jiezi and Baidu's annual lists, and Japan's "Ucan New Word and Vogue Word Awards" Top 30 lists. While these lists provide a core lexicon, they lack the rich context necessary for deep co-occurrence and semantic analysis. Consequently, we utilized Python scraping frameworks (such as Scrapy) to harvest popular original posts containing these terms from Weibo and X within their peak periods, ultimately building a contextual corpus of hundreds of thousands of words for each language. During preprocessing, Chinese data was segmented using the jieba tool with part-of-speech tagging, while Japanese data was processed using the MeCab parser with the UniDic dictionary to ensure better handling of out-of-vocabulary terms like names and neologisms. The corpora underwent standardized cleaning - including the removal of URLs, usernames, and emojis - to result in clean, structured text data for analysis.

Quantitative text analysis results

Upon importing the preprocessed data into KH Coder, frequency analysis revealed that Chinese slang discussions featured a higher proportion of verbs and adjectival phrases, such as Tangping (lying flat) and Neijuan (involution), suggesting that Chinese slang excels at providing highly generalized names for actions, states, and emotions. Conversely, Japanese corpora showed a prominence of nominal phrases and Sa-hen verb structures (e.g., Oshikatsu), reflecting a tendency toward substantivization and the flexible derivation of verbs from nouns. Co-occurrence network analysis provided even more enlightening results. The Chinese network often displayed "core-periphery" characteristics. For example, terms like Neijuan acted as hubs for a dense cluster of related terms like "competition", "pressure", and "anxiety". In contrast, the Japanese co-occurrence network appeared more "flat" and "divergent". While core nodes existed, their associated terms were more diverse, spanning direct emotional words, specific activity domains like gaming or idols, and generalized expressions. From the examples above, most of the Japanese internet slang is back clipping, like aka, ime, otsu, jidora, china, ripu and resu. Some of them are complex clipping, like furiso, anisuto, and sukusho [6]. Correspondence analysis further confirmed that Chinese slang positions in the conceptual space shifted significantly year-over-year, reflecting rapid changes in social focus, whereas Japanese slang exhibited stronger continuity in certain core concepts related to interpersonal distance and self-expression.

Deep semantic analysis results

Semantic vector analysis provided another dimension of understanding. Statistical analysis of the internal semantic similarity matrices indicated that Chinese slang terms possessed a slightly higher average internal similarity and a more concentrated distribution than their Japanese counterparts [7]. This implies that, within the model's "cognition", Chinese slang tends to orbit a relatively unified spectrum of social issues, whereas Japanese slang covers a broader and more dispersed range of social life, entertainment, and self-identity. When we combined the 60 terms for cross-lingual K-means clustering, the model did not simply split them by language [8]. Instead, several "cross-lingual" clusters emerged. For instance, a cluster expressing "negativity and fatigue" contained both Chinese emo and Babi Q le alongside Japanese pien and mō dame [9]. This empirically proves that despite formal differences, online youth in both countries share strong cross-cultural resonance in certain emotional expressions. The need for quickness, effectiveness, and clarity in online interactions has led to the frequent use of various word formation processes, such as compounding, derivation, clipping, borrowing and some others [10]. Finally, t-SNE visualization produced a "semantic map" where Chinese and Japanese points were intermingled in a "mixed-living" pattern. While high-level semantic overlap was significant, Chinese terms were more densely packed in regions expressing "macro-social phenomena and criticism", while Japanese terms were more prevalent in

areas reflecting “nuanced personal feelings and Otaku culture”, subtly reflecting divergent discourse focuses.

Conclusion

By integrating quantitative metrics with deep semantic modeling, this study proposes a preliminary explanatory framework for the divergences in Chinese and Japanese Internet slang word-formation patterns. We contend that these differences are deeply rooted in the typological traits of both languages, which are amplified under the pressures of online communication. As a language highly dependent on semantic synthesis and parataxis, Chinese displays a “semantic-driven” innovation path. Whether through the metaphorical summarization of social phenomena or the precise semantic expansion of existing terms, Chinese reflects the robust ideographic capability of characters and the cognitive habit of constructing new categories through conceptual synthesis. The findings of this research underscore the transient nature of Chinese internet buzzwords, with most experiencing a sharp decline in popularity after two years. This trend aligns with the concept of ephemeral discourse, highlighting the rapid emergence, circulation, and obsolescence of linguistic forms in online communication. In contrast, Japanese Internet slang demonstrates the potential of “morphological-driven” innovation. As an agglutinative language, its prolific affixation system and flexible word-class conversion provide ready-made molds for creating short, intimate, and phonetically specific terms. The more divergent and widely distributed nature of the Japanese network stems from the versatility of these morphological tools, allowing neologisms to permeate parallel dimensions of life rather than focusing on a few grand narratives. Such so-called “anti-language” is important for constructing a private communication system incomprehensible and intelligible for outsiders from the dominant, ultimately warding off being absorbed or assimilated. We analyze the dynamics of frequency shift and se-mantic change in slang words and compare them to those of nonslang words. Our analysis shows that slang words change more slowly in semantic meaning but exhibit more rapid frequency fluctuations and are more likely to undergo significant frequency declines.

This study successfully implements a hybrid computational framework to systematically compare the word-formation patterns of Chinese and Japanese Internet slang. A lingo can be one of the central aspects in constructing an identity of a community. The morphological creativity observed in this online environment can be explained by the concept of language energy introduced by Ezra Pound language sustains vitality and energy by interconnecting words and generating novel expressions. The research confirms that computational methods can effectively transcend the limitations of traditional qualitative inquiry, revealing systematic structural differences in morphological tendencies and semantic distributions. These findings provide a vivid annotation to linguistic typology from a digital humanities perspective. Methodologically, the study demonstrates the efficacy of combining macro-analytical tools like KH Coder with deep semantic models like Sentence-BERT. While limitations exist regarding the temporal span and “black-box” nature of vector interpretation, this research opens a new window for understanding linguistic contact and innovation in the digital age and establishes a foundation for future cross-lingual studies.

Funding

This paper is supported by the “Tianjin University Student Innovation and Entrepreneurship Training Program” and the “Tianjin University of Science and Technology Student Innovation and Entrepreneurship Training Program Project ‘Ink into gold, preserving art and revitalizing - the innovator of intangible cultural heritage New Year paintings and digital interactive products’ (Project Number: 202510057013)”.

Acknowledgements

The author would like to show sincere thanks to those techniques who have contributed to this research.

Conflicts of Interest

The author declares no conflict of interest.

References

[1] Carney, T. R. (2023) Drafting definitions with polisemy and semantic change in mind. *Obiter*, 44(3), 561-574.

[2] Tuptim, N. (2023) Japanese industrial technical terms: Word formation, word type and word pedagogical applications. *Kasetsart Journal of Social Sciences*, 44(3), 797-806.

[3] Meilantari, N. L. G., Aritonang, B. (2024) Clipping and blending words in Japanese internet Slang: philosophical perspective. *Mahadaya: Jurnal Bahasa, Sastra, Dan Budaya*, 4(1), 11-18.

[4] Al Hikmah, I., Machmoed, H., Sahib, H. (2024) An analysis of word formation processes found in TikTok application. *ELS Journal on Interdisciplinary Studies in Humanities*, 7(1), 160-169.

[5] Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., Seals, C. (2025) Hate speech detection using large language models: a comprehensive review. *IEEE Access*, 13, 20871-20892.

[6] Meilantari, N. L. G., Aritonang, B. (2024) Clipping and blending words in Japanese internet Slang: philosophical perspective. *Mahadaya: Jurnal Bahasa, Sastra, Dan Budaya*, 4(1), 11-18.

[7] Hamasalih, H. A., Ghafoory, D. S. (2025) Word formation processes used in Kurdish community Facebook platform. *Zanco Journal of Human Sciences*, 29(SpA), 275-297.

[8] Yuniarto, H. (2025) Evolving Neologisms and Ephemeral Discourse: WeChat index analysis of 2021's Chinese Internet buzzwords. *PAROLE: Journal of Linguistics and Education*, 15(1), 90-99.

[9] Yang, Y., Foucault Welles, B. (2026) Subversive humor and platformed Asianness: How Asians use neologisms to define themselves and others online. *The Communication Review*, 1-36.

[10] Fitria, T. N. (2022) Analysis of word formation process in online shop's terminologies. *Kajian Linguistik Dan Sastra*, 7(2), 67-80.