International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 21, No. 2, 2025

Semantic Similarity Modeling and Preference Analysis of Chinese Disyllabic Hypothetical Conjunctions

Chenxiao Ma*

Institute of Language Sciences, Shanghai International Studies University, Shanghai 201620, China
*0233100657@shisu.edu.cn

Abstract: The pragmatic force of Mandarin conditionals is substantially influenced by the choice of hypothetical conjunctions. Speakers select markers such as *ruguo*, *ruoshi*, *yaoshi*, *wanyi*, and *yidan* according to communicative intentions and contextual constraints, thereby modulating the tonal strength of conditionals. This study develops a semantic vector matrix model based on the Word2Vec algorithm to quantify the semantic similarity among these conjunctions within conditional constructions. By combining computational modeling with pragmatic interpretation, we examine how each conjunction differs in contextual adaptability and strategic function. The findings show that although all conjunctions can introduce conditional antecedents, their semantic distributions and pragmatic effects are systematically differentiated, making them functionally non-interchangeable. The study further reveals that the selection of a conjunction reflects an optimization process integrating structural, tonal, and stylistic parameters. This work contributes to the computational analysis of Mandarin connectives and provides a quantitative framework for exploring semantic and pragmatic variation in conditional discourse.

Keywords: Hypothetical Conjunctions, Conditionals, Semantic Similarity Modeling, Vector Matrix, Pragmatic Optimization.

1. Introduction

Conditional sentences in Mandarin represent a key interface between logic, semantics, and pragmatics. As connective markers introducing the antecedent clause, hypothetical conjunctions establish inferential relations between antecedents and consequents, thereby shaping the epistemic and tonal force of conditional statements [13]. The subtle semantic diversity among conjunctions such as *ruguo* ("如果", if), *jiaru* ("假如", if), *yaoshi* ("要是", if), *ruoshi* ("若是", if), *tangruo* ("倘若", if), *wanyi* ("万一", in case), *yidan* ("一旦", once), *jiaruo* ("假若", if), and *jiashi* ("假使", if) has long attracted scholarly attention in Chinese linguistics.

Despite extensive descriptive work, previous studies have rarely quantified the semantic relationships among these conjunctions or modeled their pragmatic variation computationally. With advances in natural language processing, particularly in distributional semantics and vector-based representation, it has become possible to capture subtle semantic distinctions numerically and to link them to communicative strategies in natural discourse.

This study therefore aims to develop a semantic vector matrix model for nine Mandarin disyllabic hypothetical conjunctions, using the Word2Vec algorithm to compute their semantic similarity and contextual adaptability. The model enables a data-driven examination of how different conjunctions encode varying degrees of hypothetical force and pragmatic tone.

The research addresses three central questions: (1) From the perspective of linguistic economy, why do multiple conjunctions coexist to express overlapping hypothetical meanings in Mandarin? (2) How can a vector matrix model be constructed to quantitatively compute the semantic similarity among these conjunctions in conditional contexts? (3) What are the semantic and pragmatic distinctions that account for their differing strengths of hypothetical force and their distributional constraints?

To address these questions, this study first outlines the construction of the semantic vector matrix and the computational procedures for similarity measurement. It then analyzes the semantic distances, contextual preferences, and tonal strength associated with each conjunction, before proposing a set of pragmatic optimization strategies for conjunction selection. The goal is to provide a unified computational-pragmatic framework for understanding the functional differentiation of Mandarin hypothetical connectives.

2. Literature Review

Conditionals in Mandarin typically consist of two clauses: an antecedent, expressing the condition, and a consequent, expressing the result or inference. The antecedent provides a hypothetical premise upon which the consequent depends. For instance:

(1) 如果因循守旧、抱残守缺，就不可能在时代的潮流中有所发展和建树。

Ruguo yinxun shoujiu, baocan shouque, jiu bukenneng zai shidai de chaoliu zhong you suo fazhan he jianshu.

If one clings to outdated conventions and resists change, one cannot achieve progress or accomplishments in the tide of the times. (CCL)

Here, the antecedent establishes a hypothetical space distinct from reality, within which the consequent holds. As noted by Stalnaker (1975), conditional interpretation always operates within a context set that admits possible worlds differing from the actual one. The truth of the consequent depends on the assumed validity of the antecedent, reflecting the probabilistic or counterfactual nature of conditional reasoning [9, 10, 20].

Typological studies have long emphasized degrees of hypotheticality. Comrie (1986) proposed a continuum from factual to potential to counterfactual conditionals, corresponding to decreasing likelihood of realization [6, 11]. Both Chinese and cross-linguistic research typically classify

conditionals into factual, hypothetical, and counterfactual types (Wang, 2010; Ciardelli & Roelofsen, 2018) [3, 22].

Cognitive and pragmatic approaches further refine this typology. Sweetser (1990) identified three domains of conditional meaning-content, epistemic, and speech-act-each reflecting a distinct mapping between world knowledge and discourse intention [21]. Pragmatic studies (e.g., Birner, 2013; Xiang, 2012) have shown that conditionals can perform various functions such as advice, persuasion, reasoning, or mitigation, depending on discourse context [2, 23].

Within this framework, hypothetical conjunctions such as *ruguo*, *jiaru*, *yaoshi*, *wanyi*, and *yidan* function as explicit markers that anchor conditional relationships and signal the level of commitment or subjectivity of the speaker [24]. Their use enhances the contrast between potential worlds and increases the discourse's informational flexibility. Luo (2007) and Zhang (2009, 2014a) trace the historical evolution and semantic differentiation of these conjunctions [15, 24, 25], noting, for example, that *wanyi* tends to convey a narrower scope of hypotheticality and a cautious or mitigating tone that softens the illocutionary force.

However, most existing research has relied on qualitative or descriptive analysis, with limited quantitative validation. Few studies have employed large corpora or computational models to assess semantic distance or to explain why certain conjunctions are pragmatically preferred in specific contexts. The lack of empirical modeling leaves unresolved how degrees of hypothetical force can be numerically characterized and how subtle tonal variations emerge from distributional differences.

Recent developments in natural language processing (NLP) have provided new tools for addressing such questions. Distributional semantic models (Mikolov et al., 2013; Pennington et al., 2014; Devlin et al., 2018) capture meaning through vector representations derived from co-occurrence patterns, allowing quantitative comparison of semantic similarity [6, 16, 18]. While widely used in lexical semantics, this approach has seldom been applied to functional categories such as conjunctions or discourse markers, which encode higher-order relational meaning.

By adopting a vector matrix-based modeling framework, this study extends computational semantics into the analysis of Mandarin conditionals. It integrates corpus-based evidence and quantitative modeling to explain the pragmatic distinctions and substitution constraints among hypothetical conjunctions. This approach bridges the gap between computational and pragmatic analyses, offering a replicable and data-driven account of conjunction choice in Mandarin.

The following section introduces the methodological design of the semantic similarity model, detailing corpus selection, preprocessing, and vector computation procedures.

3. Core Technical Framework

Building on the theoretical rationale set out above, this study implements a reproducible, empirically driven pipeline that situates distributional semantic modeling within a rigorous corpus-linguistic and pragmatic analysis of Mandarin hypothetical conjunctions. The overall objective is not merely to produce word embeddings, but to derive robust, interpretable measures of semantic proximity that inform linguistic claims about hypothetical force, register, and pragmatic optimization. The pipeline comprises (i) corpus curation and normalization, (ii) embedding training with controlled hyperparameter search, (iii) similarity computation and statistical validation, and (iv) visualization and interpretive analysis. Below we describe the rationale and operational choices for each component, together with steps taken to ensure reproducibility and linguistic interpretability.

3.1. Implementation Environment and Reproducibility

All experiments were implemented in Python ($\geqslant$3.8) to leverage its mature scientific ecosystem and to facilitate reproducibility. Key libraries and versions used in the study include: NumPy (for numerical arrays and matrix operations), Pandas (corpus manipulation), Gensim (Word2Vec implementation), NLTK / Jieba (tokenization and basic preprocessing), and Matplotlib / Seaborn (visualization). We document software versions and random seeds in the repository accompanying the paper; full scripts for corpus preprocessing, model training, and evaluation are provided to enable independent replication.

Selecting Python as the implementation platform supports (a) transparent data transformations, (b) straightforward integration with downstream statistical analyses, and (c) easy extensibility to contextualized models (BERT-family) or deep-learning frameworks (PyTorch/TensorFlow) if future work requires it. More importantly, deterministic aspects of the pipeline (vocabulary construction, corpus splits, tokenization rules) are fixed a priori and recorded as configuration files to avoid hidden sources of variability.

3.2. Representation Design: Tokenization, Granularity, and NumPy-backed Matrices

A crucial design decision concerns representation granularity-character-level vs. word-level tokenization-in Chinese. We adopted word-level segmentation using a validated Chinese tokenizer (and manual correction for ambiguous cases in a balanced development subset) because hypothetical conjunctions are multi-character lexical items whose pragmatic functions are better captured at word granularity. To mitigate segmentation noise, we post-processed outputs with part-of-speech heuristics and filtered non-conditional uses of target tokens.

NumPy underpins all vector and matrix manipulations. After embedding training, each target conjunction is represented as a D-dimensional NumPy array (D = 150 in the final model). Collecting these vectors yields a semantic matrix V $\in$ R^{N$\times$D} (N = number of target conjunctions). Pairwise similarity is computed via the cosine similarity metric implemented as efficient vectorized NumPy operations, enabling large-scale bootstrapping and permutation tests without costly Python-level loops.

3.3. Embedding model: Word2Vec choices and hyperparameter strategy

We use Word2Vec as our workhorse for distributed semantic representation, because (i) its assumptions align with the distributional hypothesis central to our linguistic questions, (ii) it is computationally efficient for the corpus scales involved, and (iii) its low-dimensional dense vectors are readily interpretable and amenable to classical linear algebraic analyses.

Model design choices and their justifications:

Architecture: We adopt Skip-gram with Negative Sampling (SGNS). Skip-gram better captures infrequent contextual

associations-important for pragmatic markers that may have skewed frequency distributions. Negative sampling stabilizes learning and reduces computational load.

Dimensionality (D): After exploratory grid search ($D \in \{100, 150, 200\}$), we selected 150 as a trade-off between representational richness and overfitting risk for this corpus size.

Context window: A wide context window = 15 was chosen to capture discourse-level cues (register, collocational frames) that inform pragmatic function beyond immediate local syntax.

Min_count: Set to 1 for the target conjunctions but higher for background vocabulary (empirically tuned) to avoid noisy rare tokens.

Negative sampling & subsampling: We used negative sampling (ns_exponent=0.75) and subsampling for frequent tokens (sample=1e-5) to reduce dominance of function words and topical high-frequency terms.

Training regimen: Models were trained for 30 epochs to ensure stable embeddings; early stopping and loss monitoring were applied in pilot runs.

To address randomness inherent in embedding training (initial weights, sampling order), we trained the full pipeline ten independent runs with different seeds and report mean similarity scores and their standard errors. This ensemble approach reduces the influence of any single stochastic instantiation and allows statistical comparison across conjunction pairs.

3.4. Evaluation, Validation, and Statistical Testing

Beyond reporting raw cosine scores, we subjected similarity results to multiple validation procedures to ensure linguistic and statistical robustness:

Intrinsic diagnostics: we inspected nearest-neighbour lists for each target conjunction and computed cluster cohesion metrics (e.g., silhouette score) on the embedded vectors to assess whether conjunctions form linguistically plausible clusters (e.g., formal vs. colloquial, risk-oriented vs. deterministic).

Bootstrapping and confidence intervals: cosine similarities were bootstrapped across sentence-level samples to produce 95% confidence intervals for each pairwise estimate, allowing us to test whether observed differences are statistically reliable.

Permutation testing: to test whether similarity patterns exceed chance, we conducted label-permutation tests where conjunction labels were shuffled across vectors; the observed similarity distribution was compared against the null distribution to obtain p-values.

Extrinsic corroboration: where applicable, we compared clustering and similarity-derived groupings with corpus-derived distributional statistics (register metadata, genre frequencies), verifying that computational clusters align with pragmatic metadata (e.g., *wanyi* more frequent in advisory/personal genres).

These validation steps transform vector-space observations into defensible linguistic claims, and they provide a principled basis for subsequent pragmatic interpretation.

3.5. Interpretability and Visualization

We use heatmaps of target vectors (dimension vs. token) to reveal concentrated vs. diffuse activation patterns across dimensions, which can suggest narrow vs. broad semantic scopes. All plots are produced via Matplotlib (and Seaborn for aesthetics) with consistent color-mapping and annotated with Chinese token labels to support cross-linguistic readers. Visual diagnostics are accompanied by quantitative cluster metrics to avoid overinterpretation of projection artifacts.

We acknowledge several practical constraints. First, Word2Vec captures distributional regularities but is not intrinsically context-sensitive; it may conflate different pragmatic senses that share distributional contexts. To mitigate this, we (a) restricted context windows to preserve discourse signals, (b) manually filtered non-conditional occurrences, and (c) plan to compare results with contextualized embeddings (e.g., BERT) in follow-up work. Second, segmentation and tokenization errors are a non-negligible source of noise; these were addressed through targeted manual correction on a development subset.

To promote transparency and reuse, we provide: (i) corpus sampling scripts (with anonymized/permissioned data pointers where applicable), (ii) all model configuration files, (iii) random seeds, and (iv) plotting and analysis notebooks. Together these artifacts ensure that reviewers and future researchers can reproduce and extend the present experiments.

In sum, the core technical framework integrates principled representation design, controlled embedding training, rigorous statistical validation, and interpretable visualization. The next section describes the concrete computational workflow-corpus extraction, preprocessing rules, training schedule, and the similarity matrices-followed by the empirical results and their pragmatic interpretation.

4. Semantic Similarity Computation and Experimental Procedures

Following the computational framework established above, this section details the procedures for modeling and evaluating semantic similarity among Mandarin disyllabic hypothetical conjunctions. The computation integrates linguistic interpretability with quantitative reproducibility, combining corpus-driven semantics, distributed representation learning, and statistical control. This dual perspective ensures that the resulting patterns are not only mathematically valid but also linguistically meaningful within the semantics-pragmatics interface.

4.1. Semantic Fields and Compositional Semantic Space

Understanding the semantic relations among lexical items is one of the central challenges of computational linguistics. The notion of semantic field (Firth, 1957) and its modern formalization as compositional semantic space provide the theoretical basis for this study[8]. According to the distributional hypothesis-"you shall know a word by the company it keeps"-lexical meaning can be inferred from contextual co-occurrence patterns. With the advent of large-scale corpora and neural representation learning, this hypothesis has evolved into quantitative distributional semantics, enabling the projection of words into a continuous, high-dimensional space where geometric distances correspond to semantic similarity (Pennington et al., 2014; Devlin et al., 2018) [5, 18].

In this study, each hypothetical conjunction-*ruguo*, *jiaru*, *yaoshi*, *ruoshi*, *tangruo*, *wanyi*, *yidan*, *jiaruo*, and *jiashi*-is embedded as a point in a shared semantic vector space derived from contextual usage across millions of sentences.

Conjunctions that frequently appear in similar syntactic or pragmatic environments (e.g., *ruguo* and *jiaru*) are expected to occupy proximate positions, whereas those with distinct discourse profiles (e.g., *wanyi* vs. *yidan*) are positioned farther apart.

This continuous semantic topology allows us to quantify previously qualitative judgments about register, tone, and hypothetical strength in conditional constructions.

Cosine similarity serves as the primary metric of semantic proximity. A higher cosine value (approaching 1) reflects stronger contextual overlap, while lower values indicate divergence in pragmatic or functional properties. Such modeling provides an empirically grounded complement to typological and intuition-based analyses of Chinese conditionals.

4.2. Vector Matrix Construction

Each conjunction is encoded as a 150-dimensional embedding vector, trained through the Skip-gram Word2Vec model. The complete set of vectors forms a semantic matrix $V \in R^{9\times150}$, where rows represent conjunctions and columns represent latent semantic features.

Pairwise cosine similarities are computed as:

$$Sim(i,j) = \frac{v_i \cdot v_j}{|v_i| \cdot |v_j|}$$

yielding a symmetric similarity matrix $S \in R^{9\times9}$, with diagonal values normalized to 1.

While cosine similarity captures the geometric alignment of conjunctions, eigen-decomposition of

S reveals principal semantic axes, corresponding to latent factors such as formality, temporal immediacy, and risk orientation. These eigenvectors summarize the dominant dimensions of variation across the conjunction system, offering interpretable insight into how Mandarin encodes degrees of hypotheticality.

For example, although *wanyi* and *yidan* may exhibit moderate co-occurrence similarity due to shared syntactic positions, their projection along the "probability" axis diverges sharply-*wanyi* signaling low-probability, anxiety-laden conditions, and *yidan* indicating deterministic or temporal causality.

Thus, vector distance provides a computational measure of pragmatic polarity.

4.3. Corpus Preparation and Preprocessing

Corpus data were obtained from the Center for Chinese Linguistics (CCL) at Peking University, encompassing diverse registers and genres of Modern Chinese. From an initial pool of 89,884 conditional sentences, we retained 77,830 valid instances after rigorous filtering. The nine most frequent disyllabic conjunctions listed above were selected for modeling.

Given the structural asymmetry of Mandarin conditionals-where the antecedent precedes and conditions the consequent-the corpus preprocessing was tailored to preserve full contextual frames. Each token occurrence was extracted with a $\pm$ 50-character window to ensure that both local collocational and broader discourse cues were retained.

Data cleaning involved three layers:

(1) Regular-expression filtering using EmEditor to remove HTML artifacts and metadata;

(2) Python-based tokenization and normalization, ensuring consistent segmentation and character encoding;

(3) Manual verification of ambiguous tokens (e.g., *yaoshi* used as a noun meaning "key").

Table 1. Corpus statistics of hypothetical conjunctions used in training

Conjunction	*Ruguo*	*Jiaru*	*Ruoshi*	*Tangruo*	*Wanyi*	*Yaoshi*	*Yidan*	*Jiaruo*	*Jiashi*
Total downloads	10000	10000	10000	10000	20000	10000	10000	5396	4488
Number used for training	10000	9991	10000	10000	7955	10000	10000	5396	4488

This preprocessing pipeline ensures linguistic validity and computational consistency, allowing the embeddings to reflect genuine contextual distributions rather than genre or tokenization artifacts.

4.4. Model Construction and Similarity Computation

4.4.1. Model Design and Vector Dimensionality

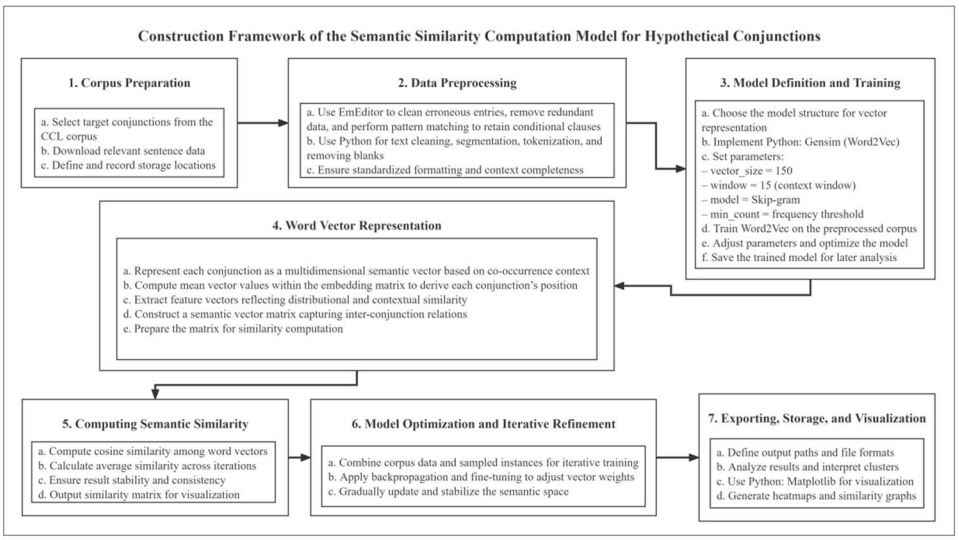

Figure 1. Construction Framework of the Semantic Similarity Computation Model for Hypothetical Conjunctions

The Word2Vec Skip-gram model was implemented using Gensim in Python. A parameter optimization grid search determined that a 150-dimensional embedding with a window size of 15 and 30 training epochs achieved the best balance between semantic granularity and generalization. Subsampling of frequent words (1e-5) and negative sampling (5 negatives per update) were applied to reduce frequency bias.

Each conjunction's embedding position results from iterative stochastic optimization, where surrounding context words act as probabilistic constraints. Consequently, the resulting semantic space encodes latent pragmatic dimensions, including stylistic preference (e.g., *jiashi* and *jiaruo* in formal registers) and emotional stance (e.g., *wanyi* conveying caution or anxiety).

4.4.2. Iterative Computation and Averaging

To counter stochastic variation inherent in neural embeddings, we trained and evaluated the model ten independent times, each with distinct random seeds. The mean cosine similarity and standard deviation were computed for every conjunction pair, producing confidence intervals for subsequent interpretation.

This ensemble approach enhances robustness and supports statistical reproducibility, allowing stable cross-run similarity patterns to be interpreted as linguistically meaningful rather than artifacts of initialization.

4.4.3. Visualization of Semantic Distributions

To facilitate interpretability, high-dimensional embeddings were visualized through heatmaps and dimensionality reduction projections. Each heatmap illustrates the activation intensity across vector dimensions for a single conjunction, revealing whether its meaning is broadly distributed or narrowly concentrated.

For instance, *jiaru* and *ruguo* show extensive overlap in high-intensity dimensions, indicating shared semantic roles as neutral hypothetical conjunctions. By contrast, *wanyi* exhibits concentrated activation in fewer dimensions, suggesting a specialized pragmatic domain emphasizing uncertainty and risk.

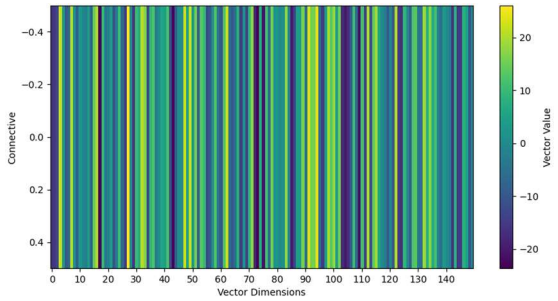

Figure 2. Semantic Vector Matrix of *jiaru*

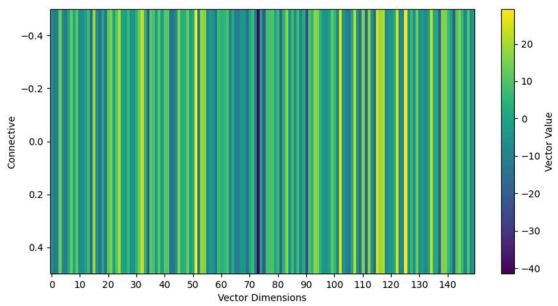

Figure 3. Semantic Vector Matrix of *jiaruo*

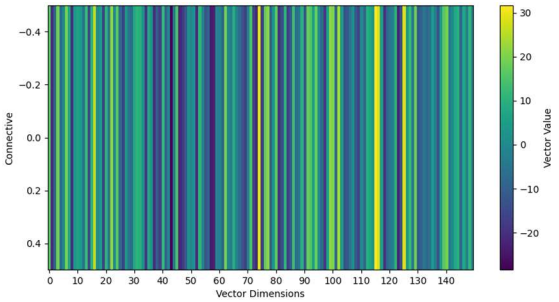

Figure 4. Semantic Vector Matrix of *ruoshi*

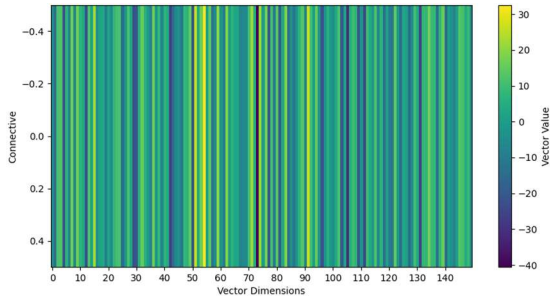

Figure 5. Semantic Vector Matrix of *tangruo*

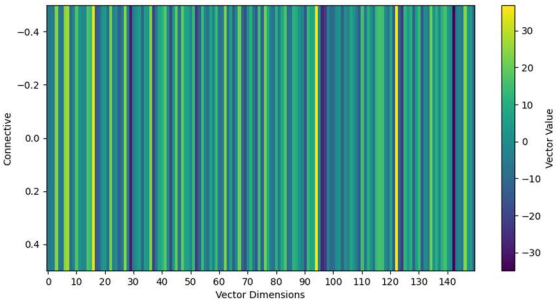

Figure 6. Semantic Vector Matrix of *yidan*

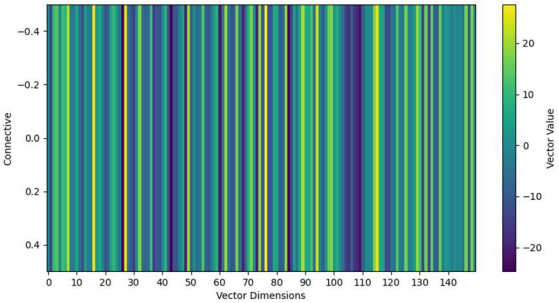

Figure 7. Semantic Vector Matrix of *jiashi*

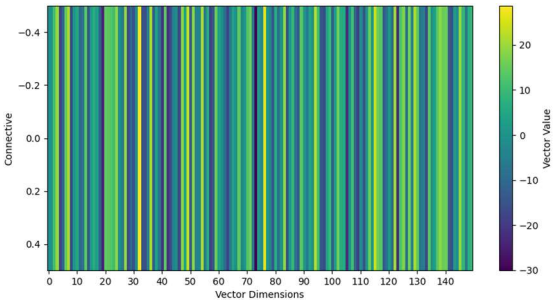

Figure 8. Semantic Vector Matrix of *ruguo*

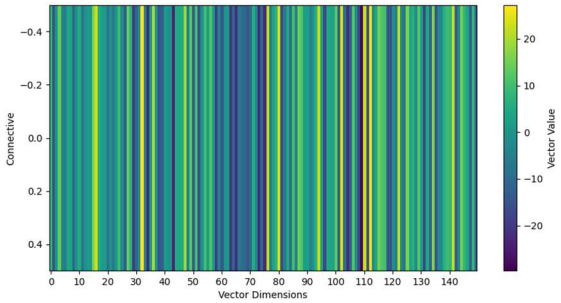

Figure 9. Semantic Vector Matrix of *wanyi*

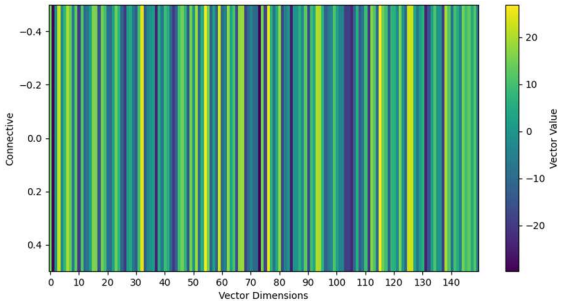

Figure 10. Semantic Vector Matrix of *yaoshi*

These visualizations provide both global and local perspectives on semantic clustering. For instance, *ruguo*, *yaoshi*, and *jiaru* cluster within a neutral-formal semantic zone, whereas *wanyi* and *yidan* form a peripheral risk-oriented cluster.

Such visualization translates abstract vector relations into linguistically interpretable structures, bridging computational representation and semantic theory.

Through the integration of corpus-based computation, distributed semantic modeling, and quantitative visualization, this section establishes the empirical and operational foundation for data-driven comparison among Mandarin hypothetical conjunctions. The next section presents the experimental findings, discusses their contextual implications, and analyzes the pragmatic preference patterns revealed by the similarity scores.

5. Experimental Results and Pragmatic Preference Analysis

Building upon the computational modeling framework established in Section 4, this section presents the empirical results of semantic similarity computation and interprets them from a semantic-pragmatic and strategic-communicative perspective. The analysis integrates corpus statistics, distributional semantics, and game-theoretic reasoning to reveal how conjunction choice in Mandarin conditionals reflects both probabilistic meaning alignment and pragmatic optimization.

5.1. Quantitative Results and Semantic Interpretation

Table 2. Semantic similarity matrix of Mandarin disyllabic hypothetical conjunctions

	Ruguo	*Jiaru*	*Yaoshi*	*Ruoshi*	*Tangruo*	*Wanyi*	*Yidan*	*Jiaruo*	*Jiashi*
Ruguo	0.000	0.404	0.382	0.465	0.487	0.526	0.539	0.551	0.556
Jiaru	0.404	0.000	0.396	0.472	0.503	0.505	0.507	0.564	0.557
Yaoshi	0.382	0.396	0.000	0.527	0.524	0.553	0.561	0.569	0.563
Ruoshi	0.465	0.472	0.527	0.000	0.383	0.495	0.489	0.550	0.543
Tangruo	0.487	0.503	0.524	0.383	0.000	0.449	0.477	0.526	0.516
Wanyi	0.526	0.505	0.553	0.495	0.449	0.000	0.527	0.546	0.530
Yidan	0.539	0.507	0.561	0.489	0.477	0.527	0.000	0.480	0.488
Jiaruo	0.551	0.564	0.569	0.550	0.526	0.546	0.480	0.000	0.242
Jiashi	0.556	0.557	0.563	0.543	0.516	0.530	0.488	0.242	0.000

The data reveal several statistically significant patterns. The highest similarity appears between *yaoshi* and *jiaruo* (0.569), indicating a close overlap in contextual usage and pragmatic orientation. In contrast, *jiaruo* and *jiashi* (0.242) exhibit the lowest similarity, suggesting distinct discourse functions despite morphological resemblance.

Overall, *ruguo* maintains moderate similarity with *jiaru* and *ruoshi*, confirming its role as a neutral, context-general conditional marker frequently used in formal registers. By contrast, *wanyi*, *yidan*, and *yaoshi* form a looser semantic cluster associated with spoken and event-oriented contexts emphasizing contingency or risk.

Interestingly, although *wanyi* and *yidan* both show moderate similarity to *yaoshi*, they diverge sharply in pragmatic polarity: *yidan* expresses temporal immediacy and determinacy, while *wanyi* encodes precaution, anxiety, and low-probability speculation [14].

These findings highlight that semantic proximity does not equate to pragmatic equivalence. Conjunctions positioned closely in the vector space may diverge in tone, emotional valence, and discourse strategy — dimensions that the embedding space captures only implicitly through distributional cues.

In short, while all nine conjunctions serve as hypothetical markers, their pragmatic force and stylistic positioning vary along continuous axes of formality, probability, and speaker stance, reinforcing the non-interchangeability observed in natural discourse.

5.2. Contextual Applicability and Strength of Hypothetical Force

To bridge quantitative similarity and pragmatic interpretation, this section evaluates each conjunction's contextual applicability and hypothetical strength, drawing upon both corpus evidence and prior qualitative research (Zhang 2014a; Han 2012) [12, 25].

Previous studies have suggested that *wanyi* and *yidan* occupy the narrowest functional range, typically introducing potential rather than counterfactual conditions [1]. Our corpus-based findings support this claim: both markers cluster near the lower end of the hypotheticality scale, emphasizing temporal immediacy (*yidan*) or low-probability

precaution (*wanyi*).

Consider the following examples:

(2) 万一某天家道中落，子孙后代凭着这些财宝还能复兴祖业。

Wanyi moutian jiadao zhongluo, zisun houdai pingzhe zhexie caibao hai neng fuxing zuye.

If one day the family fortune declines, future generations could rely on these treasures to restore their heritage.

(The Wisdom of Tolerance, Hsing Yun & Liu Changle)

(3) 假使（*万一）当初他们不坚决大干起来，哪里会有这样好的局面呢！

*Jiashi (*wanyi) dangchu tamen bu jianjue dagan qilai, nali hui you zheyang hao de jumian ne!*

Had they not resolutely taken action back then, how could such a good situation have arisen!

(People's Daily, 1958)

(4) 要是（*万一）老舍都已经被打倒在地踏上了一只脚了，我就没什么可委屈的。

*Yaoshi (*wanyi) Laoshe dou yijing bei dadao zai di tashang le yi zhi jiao le, wo jiu mei shenme ke weiqu de.*

If even Lao She had already been crushed underfoot, I would have no reason to feel wronged.

(Deng Youmei, Remembering Mr. Lao She)

As examples (3)-(4) illustrate, *wanyi* cannot substitute for *jiashi* or *yaoshi* without pragmatic incoherence, since its semantics presuppose potentiality rather than counterfactuality [7].

Integrating both distributional and contextual evidence, we observe a gradient of hypothetical strength:

Table 3. Hypothetical strength and pragmatic orientation of conjunctions

Strength	Conjunctions	Pragmatic Orientation
High	*ruguo*, *jiaru*, *yidan*	Broadly applicable; factual + counterfactual
Medium	*yaoshi*, *jiaruo*, *tangtuo*	Polite or mitigated hypotheticality
Low	*ruoshi*, *jiashi*	Concessive, indirect, or formal hypotheticals
Weakest	*wanyi*	Potential or event-triggered conditions

Although *wanyi* and *yidan* exhibit moderate cosine similarity, their pragmatic roles differ systematically: *yidan* signals temporal determinism and event causality, while *wanyi* conveys emotional uncertainty and anticipatory caution[26]. This divergence demonstrates that semantic similarity operates under pragmatic modulation, where speaker intention and discourse type reshape the interpretation of conditional meaning.

Such findings confirm that computational similarity models approximate-but do not fully capture-the pragmatic gradient of hypothetical force. The linguistic reality involves a non-linear mapping between semantic density and contextual salience, requiring pragmatic reasoning to complete the interpretation [28].

This understanding sets the stage for exploring how speakers strategically select among conjunctions in actual discourse.

5.3. Pragmatic Optimization as Communicative Strategy

From a game-theoretic standpoint (von Neumann & Morgenstern 1944; Clark 1996)[4, 17], the speaker's choice of hypothetical conjunction represents a bounded-rational optimization within a communicative game. Each conjunction yields distinct payoffs in terms of informational precision, tone, and interpersonal alignment. The speaker thus seeks a Nash-equilibrium balance—maximizing clarity while minimizing face threat or interpretive ambiguity.

Example (5) demonstrates a neutral equilibrium scenario:

(5) 如果明天暴雨不停，组委会还要研究采取进一步的措施。

Ruguo mingtian baoyu buting, zuwei hui hai yao yanjiu caiqu jinyibu de cuoshi.

If heavy rain continues tomorrow, the organizing committee will consider additional measures.

(People's Daily, 1986)

Here, *ruguo* may be replaced by *jiaru* or *yaoshi* without affecting the conditional logic, as the situation is informationally balanced and semantically neutral [27].

However, when epistemic commitment strengthens, substitution disrupts pragmatic equilibrium:

(6) 如果（*万一）因循守旧、抱残守缺，就不可能在时代的潮流中有所发展和建树。

*Ruguo (*wanyi) yinxun shoujiu, baocan shouque, jiu bukenneng zai shidai de chaoliu zhong you suo fazhan he jianshu.*

If one clings to outdated conventions and resists change, one cannot achieve progress or accomplishments in the tide of the times. (CCL)

Here, *ruguo* encodes an assertive epistemic stance and a deterministic causal link between antecedent and consequent. Substituting *wanyi* would weaken the propositional force, shifting from normative assertion to subjective speculation.

Hence, pragmatic optimization operates along three axes:

Epistemic strength - the degree of commitment between antecedent and consequent;

Social alignment - the balance between assertiveness and politeness;

Contextual risk - the degree of uncertainty tolerated in communication.

Formally, we may model speaker payoff U_S as a function of these factors:

$$U_S = \alpha E + \beta A - \gamma R$$

Where E = epistemic precision, A = alignment (cooperativity), and R = communicative risk.

Speakers choose the conjunction that maximizes U_S under contextual constraints. Thus:

ruguo and *jiaru* dominate in formal exposition (high E, low R);

yaoshi and *wanyi* optimize in spoken discourse (medium E, high A);

jiashi and *jiaruo* serve concessive or mitigated contexts (low E, high A).

This strategic differentiation confirms that Mandarin conjunction choice is not random but equilibrium-driven, reflecting an implicit optimization between semantic transparency and pragmatic face management.

6. Conclusion and Future Directions

This study has proposed and empirically validated a computational-pragmatic framework for analyzing Mandarin disyllabic hypothetical conjunctions. By integrating corpus-based modeling, vector-space semantics, and game-theoretic

pragmatics, we have shown that the choice of conjunctions such as *ruguo*, *yaoshi*, and *wanyi* reflects a structured interplay between semantic similarity and strategic optimization.

The semantic vector matrix constructed via Word2Vec effectively captured graded similarities among conjunctions, revealing continuous variation along dimensions of formality, probability, and epistemic strength. These computational measures, when aligned with pragmatic interpretation, uncovered the non-interchangeability of conjunctions that appear superficially synonymous. In particular, while *wanyi* and *yidan* occupy adjacent positions in semantic space, their pragmatic functions diverge sharply-one expressing subjective caution, the other temporal inevitability.

By linking semantic embedding geometry to communicative payoff dynamics, this research provides quantitative evidence for the long-observed intuition that conditional conjunctions encode both propositional relations and interpersonal stance. The framework thus bridges the methodological gap between distributional semantics and pragmatic reasoning, demonstrating that lexical similarity can be understood not only as a measure of contextual co-occurrence but also as a predictor of strategic substitutability in discourse.

6.1. Theoretical and Methodological Contributions

From a linguistic standpoint, the findings contribute to three interrelated domains: (1) Semantic typology of conditionals. The graded similarity results support a scalar model of hypotheticality, extending Comrie' s (1986) typology by embedding Chinese conditionals into a continuous vector space. (2) Computational pragmatics. The study demonstrates that neural embedding models can be repurposed for pragmatic interpretation when coupled with communicative utility functions, allowing quantitative modeling of tone modulation and politeness strategies. (3) Game-theoretic linguistics. The optimization analysis provides a formalized account of conjunction choice as a strategic equilibrium. Speakers select forms that maximize epistemic clarity and social alignment while minimizing communicative risk, reflecting bounded rationality in language use.

Methodologically, this study advances quantitative linguistic analysis in several ways: (1) It operationalizes semantic fields through measurable vector geometry rather than categorical typology; (2) It introduces robust averaging and cross-validation techniques to stabilize semantic similarity scores; (3) It integrates visualization-based interpretation to facilitate linguistic insight beyond raw numerical output; (4) And most crucially, it formalizes speaker decision-making as a payoff-based optimization process within the same computational framework that models meaning distribution.

This convergence of corpus linguistics, quantitative semantics, and pragmatic reasoning illustrates a scalable methodology for bridging symbolic interpretation with statistical modeling in natural language research.

6.2. Limitations and Future Directions

Despite its contributions, the present study has several limitations that open new avenues for future research.

First, the Word2Vec-based embedding captures only co-occurrence-level semantics and lacks explicit modeling of syntactic structure or discourse prosody. Future work could employ contextualized embeddings such as BERT or RoBERTa to refine the semantic space, allowing the representation of sentence-level and pragmatic nuances.

Second, the game-theoretic model presented here remains conceptual; subsequent research could formalize it using Bayesian decision frameworks, explicitly quantifying communicative risk and epistemic uncertainty. Such models could simulate how speakers update beliefs about hearer interpretation in real time—a natural extension of Shannon information entropy[19] and Bayesian belief updating principles.

Third, cross-linguistic comparison offers fertile ground for generalization. Applying this integrated framework to conditional systems in English, Japanese, or Korean could reveal universal vs. language-specific mechanisms in tone modulation and hypothetical reasoning.

Finally, the proposed methodology holds promise for applied NLP tasks such as sentiment-conditioned text generation or style-adaptive machine translation, where pragmatic appropriateness is as crucial as semantic accuracy.

6.3. Concluding Remarks

In summary, this research advances a unified account of how meaning, probability, and strategy interact in the linguistic encoding of hypotheticality. The computational quantification of semantic similarity not only deepens our understanding of lexical structure but also provides an empirical foundation for modeling pragmatic reasoning as a form of communicative optimization.

Through the fusion of quantitative linguistics, information theory, and game theory, this study contributes to an emerging paradigm of computational pragmatics — one in which meaning is treated as both an informational signal and a strategic choice.

Future extensions of this framework will continue to refine the dynamic interface between semantic representation and pragmatic inference, moving toward a more comprehensive, data-driven theory of language as an adaptive, cooperative system.

References

[1] Bai, Guangliang. (2018). An error analysis and acquisition order investigation of the "ruguo"-type hypothetical conjunctions based on a corpus study [Master's thesis, Fujian Normal University]. [in Chinese]

[2] Birner, B. J. (2013). Introduction to Pragmatics. New Jersey: Wiley-Blackwell.

[3] Ciardelli, I., & Roelofsen, F. (2018). An inquisitive perspective on modals and quantifiers. Annual Review of Linguistics, 4, 129–149.

[4] Clark, R.L., (2012). Meaningful games: Exploring language with game theory. Cambridge: MIT Press.

[5] Comrie, B. (1986). Conditionals: A typology. In E. C. Traugott, A. ter Meulen, J. S. Reilly, & C. A. Ferguson (Eds.), On Conditionals (pp. 77–99). Cambridge: Cambridge University Press.

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[7] Fang, Xianghong. (2004). A semantic study of modern Chinese connectives and related sentence structures based on

intensional logic [Doctoral dissertation, Shanghai Normal University]. [in Chinese]

[8] Firth, J. R. (1957). Papers in Linguistics 1934–1951. Oxford: Oxford University Press.

[9] Goodman, N. (1947). The problem of counterfactual conditionals. The Journal of Philosophy, 44(4), 113–138.

[10] Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. Topics in Cognitive Science, 5, 173–184.

[11] Grice, H. P. (1989). Studies in the Way of Words. Cambridge, MA: Harvard University Press.

[12] Han, Qizhen. (2012). A cognitive study of concessive conditional sentences in modern Chinese [Doctoral dissertation, Huazhong University of Science and Technology]. [in Chinese]

[13] Harris, M. B. (1986). The historical development of si-clauses in Romance. In E. C. Traugott, A. Meulen, J. S. Reilly, & C. A. Ferguson (Eds.), On Conditionals (pp. 265–284). Cambridge: Cambridge University Press.

[14] Jiao, Min. (2022). An analysis of differences in the use of "ruguo"-type hypothetical conjunctions. Cultural Innovation Comparative Research, (04), 37–41. [in Chinese]

[15] Luo, Ronghua. (2007). The grammaticalization of "wanyi". Journal of Yichun College, (01), 74–78. [in Chinese]

[16] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems (pp. 3111–3119).

[17] Neumann, J.V., Morgenstern, O., (1944). Theory of Games and Economic Behavior. Princeton: Princeton University Press.

[18] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543).

[19] Shannon, C. (1948). The mathematical theory of communication. Bell System Technical Journal, 27, 379–423, 623–656.

[20] Stalnaker, R. (1975). Indicative conditionals. In R. Stalnaker, W. L. Harper, & G. Pearce (Eds.), Ifs: Conditionals, Belief, Decision, Chance, and Time (pp. 193–210).

[21] Sweetser, E. (1990). From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure. Cambridge: Cambridge University Press.

[22] Wang, Chunhui. (2010a). "Hypothetical hierarchy" and Chinese conditional sentences. Studies in the Chinese Language, (04), 59–69, 96. [in Chinese]

[23] Xiang, Chengdong. (2012). A pragmatic analysis of logical conditional sentences. Foreign Languages Journal, (02), 96–100. [in Chinese]

[24] Zhang, Xueping. (2009). A discourse analysis of "wanyi". Chinese Teaching in the World, (01). [in Chinese]

[25] Zhang, Xueping. (2014a). The semantic functions and pragmatic distribution of "ruguo"-type hypothetical conjunctions. Chinese Language Learning, (01). [in Chinese]

[26] Zhang, Xueping. (2014b). "Yidan" and "wanyi". Chinese Teaching in the World, (02). [in Chinese]

[27] Zhang, Xueping. (2019). The co-occurrence patterns and causes of "ruguo(shuo)" hypothetical expressions. Chinese Language Learning, (02), 24–31. [in Chinese]

[28] Zhou, Simin. (2019). The influence of counterfactual elements on the comprehension of counterfactual semantics in Chinese hypothetical conditionals [Master's thesis, Northeast Normal University]. [in Chinese]