Analysis of Linguistic Description Features of New Discoveries in Chinese Scientific and Technical Literature within the Field* Postprint

Authors: Mao Chenyu, Le Xiaoqiu

Date: 2017-10-11T00:00:00+00:00

Abstract

Objective

This study aims to analyze the linguistic description features of new discoveries in Chinese scientific literature within the field.

Method

We employed semantic annotation of linguistic description features of new discoveries and investigated their patterns through syntactic analysis, frequency distribution statistics, and co-occurrence analysis.

Result

We summarized sentence patterns for new discovery language in Chinese scientific literature within the field and identified characteristic collocations of new discovery language.

Limitation

The results have domain-specific limitations and require further comparative studies.

Conclusion

Utilizing semantic annotation, frequency statistics, and co-occurrence analysis can effectively discover the linguistic description characteristics of new discoveries in Chinese scientific literature.

Full Text

Preamble

ChinaXiv Collaborative Journal, Issue 270, 2016, Issue 5

Linguistic Feature Analysis of New Findings Descriptions in Chinese Scientific Literature within Specific Fields

Mao Chenyu[1],[2], Le Xiaoqiu[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract

Objective This study analyzes the linguistic features used to describe new findings in Chinese scientific literature within specific fields. **[Methods]** We employed semantic annotation to identify linguistic features of new findings, then investigated their patterns through syntactic analysis, frequency distribution statistics, and co-occurrence analysis. **[Results]** We summarized the sentence patterns for expressing new findings in Chinese scientific literature and identified characteristic collocations used in such descriptions. **[Limitations]** The results are limited by disciplinary boundaries and require further comparative research. **[Conclusions]** Semantic annotation, frequency statistics, and co-occurrence analysis can effectively reveal the descriptive features of new findings in Chinese scientific literature.

Keywords: New finding; Linguistic feature; Semantic annotation
Classification Number: TP393

Scientific literature aims to provide new knowledge to other researchers working on the same problems [1]. A paper that qualifies as a scientific research achievement must contain new discoveries, new hypotheses, or new theories [2]. Scientific discoveries and theoretical innovations are important manifestations of scientific and technological innovation [3]; therefore, authors adopt specific descriptive methods to declare their originality. From the perspective of natural language understanding, analyzing the linguistic features of new findings to reveal their patterns in academic literature is of significant practical importance for improving the recall rate of rule-based information extraction.

To grasp the linguistic characteristics of new findings in Chinese scientific literature, this paper focuses on the language used to describe new discoveries within specific fields. Through methods such as semantic annotation, word frequency statistics, and co-occurrence analysis, we examine how new findings are described, explore their linguistic patterns, and analyze features such as characteristic words and syntactic collocations, providing a foundation for constructing patterns for identifying new findings.

2 Research Status of Linguistic Features for New Findings in Scientific Literature

Scientific discovery refers both to the process of making discoveries and to the results of those discoveries. This study focuses on the latter as manifested in scientific literature. Qian Shiti [4] and Li Xingmin et al. [5] argue that discovering new facts from nature or proposing new concepts, principles, hypotheses, laws, and theoretical systems in scientific research all constitute scientific discoveries. Qiu Renzong [6] notes that "scientific discoveries must involve finding something previously unknown, with the scientific community as the reference frame, and must be, in principle, testable. Moreover, the results of scientific discoveries can be integrated into the scientific knowledge system as a new chapter or supplement." Tan Shusheng [7] points out that scientific discovery and theoretical innovation involve discovering new scientific facts and establishing new scientific theories (both positive and negative), which represent revelations and understandings of previously unknown substances and their properties, laws of motion, and new phenomena in nature, primarily expressed in the form of academic papers or monographs.

The above definitions primarily address "scientific discovery" in general, while the specific content of new findings as manifested in scientific literature lacks clear definition. Building on previous theories, this paper defines "new findings" in scientific literature as discoveries and innovations regarding natural phenomena, objects, principles, characteristics, and laws within a specific research field, achieved through research or experience, as well as the revelation of new facts [8].

Through manual reading and analysis of a substantial number of scientific articles, we found that the abstract, introduction, and conclusion sections centrally reflect the new findings of a paper. We selected unstructured abstracts from scientific literature as our experimental corpus. Based on our definition of new findings, we conducted linguistic analysis of sentence groups describing new discoveries, phenomena, characteristics, and laws in these abstracts, manually annotating characteristic words (e.g., "discovered"), phrases (e.g., "revealed the law of⋯"), sentence patterns (e.g., contrastive sentences), and structures (e.g., parallel structures: "The study found: 1⋯; 2⋯; 3⋯").

Example 1: "This study, focusing on woody plants in Tiantong, Zhejiang, #studied#FT the relationship between twig size (cross-sectional area) and quantity (density) and #found#FT: 1) twig density was #significantly#FT #negatively correlated#FT with branch cross-sectional area⋯, and there was #no#FT #significant#FT #difference#FT in twig density between the two life forms; ⋯" —From Xu Yue et al., "The 'Size-Number' Trade-off of Twigs in Woody Plants in Tiantong, Zhejiang"

Example 2: "⋯systematically #studied#FT the adaptation mechanisms of Masson pine families to different types of low-P stress and the variation #laws#FT of P efficiency. #The results show#FT that P efficiency indicators

such as seedling height, ground diameter, and biomass of the tested Masson pine families all #showed#FT #significant#FT family variation, ···" —From Yang Qing et al., "Genetic Variation in Root Architecture and Phosphorus Efficiency of Masson Pine Families Under Heterogeneous Low-Phosphorus Stress"

3.2 Extraction of Collocation Features for New Finding Language

After completing corpus annotation, we analyzed characteristic collocations in new finding descriptions. Regarding collocation, researchers such as Choueka et al. [26], Benson et al. [27], and Church et al. [28] define it as recurrent, interrelated, and somewhat arbitrary word combinations. Collocation can also be distinguished as narrow or broad [29-30]: narrow collocation refers specifically to fixed collocations requiring restrictive co-occurrence relationships between words, while broad collocation refers to co-occurring words that appear in context with syntactic and lexical associations. This study defines new finding language feature collocations as combinations of annotated feature words that appear in the context of new finding sentences and have certain grammatical relationships.

During annotation, we observed that most new finding language feature collocations appear within specific contexts. Therefore, when extracting feature collocations, we traversed sentences in the abstracts and extracted combinations of annotated feature words from clauses, such as "studied the law of···" and "showed significant···indicators."

After extraction, we systematically categorized all feature collocations. For instance, collocations indicating new finding categories (such as "explored the characteristics of···" and "revealed the law of···") were classified as "Type," while words indicating specific new finding results (such as "the study found" and "the results show") were classified as "Result." The specific categorization is shown in Table 1 .

Table 1 Feature Collocation Categorization Correspondence Table (Partial)

Category	Examples
Type	explored the characteristics of··· /analyzed the laws of···/studied the influence of···/···
Result	the results show/indicate/research shows/···/the study found/analysis found/···

Some collocations, although conforming to the above categorization rules, were

listed separately for analysis due to their high frequency within a category. For example, "significant difference" was not grouped under "significant N" but was instead counted separately.

3.3 Statistical Analysis of Feature Collocation Frequency Distribution

Statistical analysis of feature collocation frequency distribution can effectively identify which collocations appear more frequently in the corpus and which words are more widely distributed in abstracts describing new findings.

The IDF (Inverse Document Frequency) of a feature collocation can be calculated by dividing the total number of documents by the number of documents containing the collocation, then taking the logarithm of the quotient:

$$\mathrm{IDF} = \log \frac{D}{\mathrm{Num} + 1}$$

where D is the number of abstracts in the corpus, and Num is the number of abstracts containing the collocation. We use Num+1 to avoid division by zero when a collocation does not appear in the corpus.

TF (Term Frequency) represents the total number of times a collocation appears in the corpus. The IDF value measures the discriminative power of a collocation across the entire corpus. If a collocation appears many times in a particular document but not frequently across the whole corpus, it indicates high discriminative power for that document' s topic. Conversely, if a collocation appears frequently in both individual documents and the entire corpus, its discriminative power is low. Since this study aims to identify common collocations that characterize new finding language features, we prefer collocations with smaller IDF values.

3.4 Co-occurrence Analysis of Feature Collocations

Co-occurrence analysis can reveal which feature collocations frequently appear together and which collocation combinations have sequential links in new finding corpora. The calculation method represents each abstract in the corpus as a combination of feature collocations and statistically calculates the co-occurrence frequency of pairwise feature collocation combinations.

4.1 Experimental Data Sources and Preprocessing

Considering literature sources, core journals with high impact factors and inclusion in major databases tend to contain more and higher-quality new findings. From a disciplinary perspective, comparative analysis across multiple fields helps derive generalizable patterns. This study selected three fields for comparative analysis: botany, physics, and chemistry. For botany, we chose the journal

Acta Phytoecologica Sinica; for chemistry, *Acta Chimica Sinica*, *Acta Polymerica Sinica*, and *Chinese Journal of Organic Chemistry*; and for physics, *Acta Physica Sinica*. Data from each journal were customized and exported from CNKI [31] in EndNote format and stored in a database. The main fields of the data table are shown in Figure 1 [Figure 1: see original paper].

Figure 1 Initial Abstract Field Display

In the experiments, we selected the first 120 abstracts from the botany field and manually annotated their new finding features. For comparison, we randomly selected 300 abstracts each from the chemistry and physics fields for manual annotation, obtaining 116 and 114 abstracts containing new finding features, respectively.

Figure 2 New Finding Feature Collocation Normalization Table (Excerpt)

4.2 Experimental Results

The experimental results primarily include the main expression patterns, high-frequency feature collocation analysis, and co-occurrence analysis of new finding language in Chinese scientific literature.

(1) Analysis of Expression Patterns for New Finding Language

After completing annotation of the new finding corpus, we counted the categories of new findings in each field. They can be broadly divided into categories such as "influence," "characteristics," "laws," and "relationships," though differences exist between fields. The specific categories are shown in Table 2 .

Table 2 Statistics of Main Categories of New Findings by Field

Category	Botany	Chemistry	Physics
Influence	23.3%	25.9%	27.1%
Characteristics/Properties	27.1%	15.5%	13.1%
Laws	13.1%	13.3%	3.5%
Effects	-	-	-

Analysis of Table 2 reveals: The distribution of new finding categories varies across fields; In botany, physics, and chemistry, the main categories of new findings all involve "influence," "characteristics/properties," and "laws" ; Compared with botany, physics and chemistry have more similar categories, both including descriptions of mechanisms, behaviors, and various properties (e.g., activity, stability, conductivity), and their new finding categories are more dispersed.

After extracting features from the annotated abstracts, we obtained new finding feature collocation tables for each field. Following the normalization method

described in Section 3.2, we standardized the feature collocations, with the standardized data shown in Figure 2.

Through syntactic analysis of the main categories of new finding language, we gained a macroscopic understanding of sentence patterns, which facilitates the later design of linguistic regular expressions.

The common descriptive sentence patterns, phrases, and words for the main categories are shown in Table 3 , where "cue sentences" indicate the category of new findings, and "result sentences" contain specific new finding results. Result sentences typically include cue words such as "the results show/found."

Table 3 Main Sentence Patterns for "New Finding" Categories

Category	Main Patterns/Phrases (Cue Sentences)	Main Patterns/Phrases/Words (Result Sentences)
Influence	studied the influence of···on··· observed the influence of···on ···	the results show/indicate/found··· has significant (considerable) influence on·····is influenced by···
Characteristics/Properties	studied the (optical/electrical ···) propertiesexplored the stability of···	significantly in- creased/decreased/raisedmaximum/optimal ···
Laws	studied the law of···revealed the law of···	has obvious boundariesshows significant positive/negative correlation
Behaviors	studied the behavior of··· observed the behavior of···	showed significant···highest/lowest maximum value
Mechanisms	studied the reasons for changeexplored the factors of ···	behavior is more significantas···, ···

(2) Frequency Statistical Analysis of New Finding Feature Collocations

After normalizing and categorizing the annotated feature collocations from the new finding corpus, we calculated their IDF values and sorted them in descending order, selecting the Top 15 for each field, as shown in Tables 4 –6 .

Table 4 IDF Calculation for Feature Collocations in Botany

Collocation	IDF
shows···trend	···
significant···difference	···
shows···correlation	···

Table 5 IDF Calculation for Feature Collocations in Chemistry

Collocation	IDF
shows···trend	···
beneficial to	···

Table 6 IDF Calculation for Feature Collocations in Physics

Collocation	IDF
shows···trend	···

Integrating the IDF statistics across the three fields reveals: Both Type and Result collocations are universally present across all three fields, indicating that scientific literature consistently uses a combination of category cue words (Type) and specific result indicators (Result) to describe new findings; "Discovery" phrases (e.g., "the study found" , "analysis found" , "comparison found") show domain-specific characteristics, appearing more frequently in chemistry and physics than in botany; Degree modifiers such as "significant" , "obvious" , and "most···"have high frequencies across all three fields, highlighting "noticeable, remarkable" facts commonly used in control experiment results—an important aspect of new findings; When describing laws, the "as/with···" collocation pattern is universally adopted across all three fields.

(3) Co-occurrence Statistical Analysis of New Finding Feature Collocations

Based on the normalization tables used for IDF calculation, we computed pairwise co-occurrence relationships of feature collocations across different fields. By analyzing high-frequency co-occurrence relationships, we examined the linguistic patterns commonly used to describe new findings and explored similarities and differences across fields. The results are shown in Tables 7 –9 (Top 10).

Table 7 Co-occurrence Statistics for New Finding Feature Collocations in Botany

Collocation Pair	Occurrence
Type-Result	High

Table 8 Co-occurrence Statistics for New Finding Feature Collocations in Chemistry

Collocation Pair	Occurrence
Type-Result	High

Table 9 Co-occurrence Statistics for New Finding Feature Collocations in Physics

Collocation Pair	Occurrence
Type-Result	High

Analysis of the co-occurrence tables across the three fields shows: The co-occurrence frequency between new finding category (Type) and result indicator (Result) is high, suggesting all three fields tend to adopt this Type-Result structure for describing new findings; "Discovery" collocation patterns appear frequently in physics and chemistry to introduce new findings; When describing specific new findings, authors tend to use attention-grabbing words such as "significant" and "most…" .

4.3 Experimental Validation

Based on the new finding language expression patterns identified in Section 4.2, we compiled tables of new finding type expressions, result cue words, and characteristic words, and designed experiments to validate the accuracy of these features.

The experimental approach involves: for a test abstract, first segmenting it into sentences, then performing regular expression matching in the order of new finding category rules, result cue words, and new finding content feature rules. If the abstract contains both new finding category collocations and content feature collocations, it is classified as new finding-related (result cue words are optional). We then calculated Precision and Recall using the following formulas:

$$\text{Precision} = \frac{\text{Number of relevant documents in results}}{\text{Total number of returned results}}$$

$$\text{Recall} = \frac{\text{Number of relevant documents in results}}{\text{Total number of relevant documents}}$$

The test corpus used journal data from the same sources as Section 4.1, with 100 new abstracts randomly selected from each field and manually judged for new finding content, as shown in Table 10 .

Table 10 Distribution of New Finding Content in Test Corpus

Field	With New Finding Content	Without New Finding Content	Total
Botany	…	…	100
Chemistry	…	…	100

Field	With New Finding Content	Without New Finding Content	Total
Physics	···	···	100

The experimental results are shown in Table 11 :

Table 11 New Finding Feature Identification Performance by Field

Field	Precision	Recall
Botany	81.48%	78.57%
Chemistry	70.00%	67.74%
Physics	62.29%	76.00%

These results demonstrate that the feature set summarized in this study achieves relatively high precision and recall in describing new finding content. Botany shows the best performance because new finding categories in this field are more concentrated, whereas chemistry and physics are relatively more dispersed (see Table 2). In summary, the new finding language feature set compiled in this study demonstrates satisfactory accuracy.

5 Conclusion

This study examined the language of new findings in Chinese scientific literature within specific fields. Through semantic annotation, word frequency statistics, and co-occurrence analysis, we conducted a preliminary exploration of descriptive patterns for new finding language, analyzing features such as sentence patterns and characteristic collocations. We also performed comparative research on sentence patterns and feature collocations across different fields, analyzing similarities and differences in their expressions, thereby achieving quantitative research on linguistic expressions of new findings in Chinese scientific literature across different fields.

The limitation of this study is that the analysis of new finding language descriptions is currently restricted to natural science literature, and the extensive semantic annotation work was conducted manually, which is time-consuming and labor-intensive. Future work should employ machine learning methods to learn expression patterns for large-scale computation.

In subsequent research, we will build upon these results to further explore the descriptive characteristics of new finding language, establish a descriptive model for new finding language, and make the results more practically meaningful.

References

[1] Berkenkotter C, Huckin T N. Genre Knowledge in Disciplinary Communication: Cognition/Culture/Power [M]. Lawrence Erlbaum Associates, Inc, 1995.

[2] Wen Youkui, Wu Guangyin. Dynamic Mining of Fragmented Scientific Research Innovation Points[J]. Digital Library Forum, 2014(7): 25-32.

[3] Zhu Daming. Four Essential Factors in Appraising Innovation of Papers of Sci-tech Periodicals[J]. Science and Technology Management Research, 2011(9): 199-201.

[4] Qian Shiti. Thinking About the Structure of Scientific Discovery [J]. Science Technology and Dialectics, 1989(3): 37-40.

[5] Li Xingmin, Song Desheng, Wang Shenli. The Highest Musical Charm in Ideological Field: Case Studies of Scientific Discovery [M]. Changsha: Hunan Science and Technology Press, 1998.

[6] Qiu Renzong. Road to Success: Scientific Discovery Mode [M]. Beijing: People' s Press, 1987.

[7] Tan Shusheng. Evaluation Standard of Scientific Discovery and Theoretical Innovations[J]. Invention & Innovation, 2006(1): 38-39.

[8] Zhou Luyang. Index System for Identifying Innovation Factors in Academic Papers [J]. Acta Editologica, 2006, 18(1): 68-70.

[9] Swales J M. Genre Analysis: English in Academic and Research Settings [M]. Cambridge University Press, 1990.

[10] Swales J M. Research Genres: Exploration and Applications [M]. Cambridge University Press, 2004.

[11] Hyland K. Disciplinary Discourses: Social Interactions in Academic Writing [M]. University of Michigan Press, 2004.

[12] Hyland K. Metadiscourse [M]. John Wiley & Sons, Inc., 2005.

[13] Hunston S. Professional Conflict: Disagreement in Academic Discourse [A]. //Text and Technology: In Honour of John Sinclair [M]. John Benjamins Publishing Company, 1993.

[14] Dahl T. Contributing to the Academic Conversation: A Study of New Knowledge Claims in Economics and Linguistics [J]. Journal of Pragmatics, 2008, 40(7): 1184-1201.

[15] Dahl T. The Linguistic Representation of Rhetorical Function: A Study of How Economists Present Their Knowledge Claims [J]. Written Communication, 2009, 26(4): 370-391.

[16] Publishing with Us [EB/OL]. [2015-11-13]. http://www.sciencemag.org/site/help/authors/publishing.xhtm

[17] Wen Youkui, Wen Hao, Xu Duanyi, et al. Knowledge Element Mining in Knowledge Management [J]. Journal of the China Society for Scientific and Technical Information, 2005, 24(6): 663-670.

[18] Wen Youkui, Wen Hao. Sentence Group Distribution of Keywords and Innovation Idea Words [J]. Journal of the China Society for Scientific and Technical Information, 2007, 26(1): 50-55.

[19] Teufel S, Moens M. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status [J]. Computational Linguistics, 2002, 28(4): 409-445.

[20] Sandor Á. Modeling Metadiscourse Conveying the Author' s Rhetorical Strategy in Biomedical Research Abstracts [J]. Revue Française de Linguistique Appliquée, 2007, 12(2): 97-108.

[21] Sandor Á, De Waard A. Identifying Claimed Knowledge Updates in Biomedical Research Articles [C]. In: Proceedings of the Workshop on Detecting Structure in Scholarly Discourse. Association for Computational Linguistics, 2012: 7-15.

[22] Ibekwe-SanJuan F, Chen C, Pinho R. Identifying Strategic Information from Scientific Articles Through Sentence Classification [C]. In: Proceedings of the 6th International Language Resources and Evaluation (LREC' 08), Marrakech, Morocco. 2008.

[23] Ibekwe-Sanjuan F, Silvia F, Eric S, et al. Annotation of Scientific Summaries for Information Retrieval [C]. In: Proceedings of ECIR' 08 Workshop on: Exploiting Semantic Annotations for Information Retrieval. 2008.

[24] Leng Fuhai, Bai Rujiang, Zhu Qingsong. A Hybrid Semantic Information Extraction Method for Scientific Research Papers[J]. Library and Information Service, 2013, 57(11): 112-119.

[25] Lai Yuangen. Research on Linking Method Between Periodical Thesis and Patent Literature[J]. Document, Information & Knowledge, 2011(1): 63-68.

[26] Choueka Y, Klein T, Neuwitz E. Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus [J]. Journal of the Association for Literary and Linguistic Computing, 1983, 4(1): 34-38.

[27] Benson M, Benson E, Ilson R F. The BBI Combinatory Dictionary of English: A Guide to Word Combinations[M]. John Benjamins Publishing Company, 1986.

[28] Church K W, Hanks P. Word Association Norms, Mutual Information, and Lexicography [C]. In: Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics. 1989: 76-83.

[29] Chen Yaju. Automatic Extraction of Chinese Collocation [D]. Shanghai: East China Normal University, 2006.

[30] Shen Xiuying. A Study of Chinese Collocation [D]. Shanghai: Fudan University, 2007.

[31] CNKI [EB/OL]. [2015-11-05]. http://www.cnki.net/.

Author Contributions

Mao Chenyu: Designed and implemented the technical solution and roadmap, collected and cleaned data, analyzed experiments, wrote the manuscript, and revised the final version.
Le Xiaoqiu: Proposed the research direction and main ideas, designed the research plan and technical roadmap, and revised parts of the article.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

The supporting data is self-archived by the authors and can be accessed via E-mail: maochenyu@mail.las.ac.cn.

[1] Mao Chenyu, Le Xiaoqiu. nf_{idf}.xlsx. TF-IDF Statistics for New Finding Feature Collocations in Botany, Chemistry, and Physics (Frequency > 1).
[2] Mao Chenyu, Le Xiaoqiu. nf_{occurrence}.xlsx. Co-occurrence Relationships of New Finding Feature Collocations in Botany, Chemistry, and Physics (Frequency > 1).

Received: 2015-11-26
Revised: 2016-03-07

Note: Figure translations are in progress. See original paper for figures.