# POETRY STYLISTIC ANALYSIS TECHNIQUE BASED ON TERM CONNECTIONS

LIANG-YAN LI, ZHONG-SHI HE, YONG YI

College of Computer Science, Chongqing University, Chongqing, 400030, China
E-MAIL: LF813@swnu.edu.cn , zshe@cqu.edu.cn

**Abstract:**

**Literary Language Processing deserves its due attention in the current research atmosphere of Natural Language Processing (NLP). Since poetry fully reveals Literary Language features such as vividness, sensibility and individuality, it is the appropriate start-point in NLP. Stylistic Analysis thus contributes as an important task in Literary Language Processing with lots of challenges. This paper looks into the research object, Poetic Language, strongly recommends and carefully proves Poetry Stylistic Analysis Technique based on Term Connection with the support of NLP technique as the background. Furthermore, the corresponding algorithm is proposed and questionnaires are applied to evaluate Poetry Stylistics in surveys. Both theories and experiments confirm us that commonness exceeds individuality concerning Poetry Stylistic Analysis, therefore Poetry Stylistic Analysis Technique based on Term Connection is valid in evaluating Poetry Stylistics.**

**Keywords:**

**Literary Language Processing; Stylistic Feature; Analysis Technique; Term Connection**

## 1. Introduction

Literary Language is an important part of Natural Language, however, Non-literary Language has been contributing as the main focus in Natural Language Processing (NLP) so far and language literary features such as vividness, sensibility and individuality are often neglected. Thus a foreseeable crisis reveals here: any technique developed under current conditions is hard to handle Literary Language, which violates the theoretical definition of NLP. As is known to all that Literary and Non-literary Languages are not entirely different. Analyzed in syntax layer, literary works often include massive Non-literary Language, and certain non-literary works must also contain Literary Language for betterment in effects. We can better explain this as follows: human beings are Natural Language Application main users, and with both requirements of material gain and aesthetic tastes in natural language, we demand Literary Language as a result. Quite naturally, Literary Language Processing is unavoidable in NLP.

Up to now, Literary Language has not been clearly defined in NLP areas. Generally Speaking, literature is "an art that vividly reveals social living conflicts with language and characters as tools, and is further divided into forms such as drama, poetry, novel, prose, etc [1]. Among them, poetry is the most literary stylistic form because of "vividness" as its most typical feature. *Motets of Poetry*, the earliest Chinese poetry collection, employs rhetorical devices as Bi, Xing to portray love. Take "关关雎鸠，在河之洲" as an example. From the aspect of literary Sensibility, poetry expresses the strongest passion. No wonder traditional Chinese literature essays agree "poems derive from emotions", "poems root in passion, describe in language, prosper in rhymes and strengthen in meaning". Chinese traditional poetry beats into and is fully worthy of its title, world treasure. With the wonderful unity of content and form, it is still quite influential and powerful nowadays. From the aspect of Art, the so-called revolutionary modern poems enjoy fewer and fewer readership, yet traditional poetry analysis[2] and creation[3][4] become more and more popular. From the aspect of research, traditional Chinese poetry corpus collection, studies and creation with the aid of computers are gaining more concern and interests, and thus computer turns into a tool of vital importance in traditional poetry modernization. In the above social atmosphere, Poetry Language Processing not only proves its own feasibility, but also meets the urgent requirements of poetry art and NLP.

Poetry Language Processing is no doubt a great challenge for computers, because computers are good at logic thinking while poetry derives from human visual thinking. Concerning NLP, the very challenge expresses in specific hard tasks, among which are Poetry Evaluation and Stylistic Analysis. Style is a collective representation of Literary Language individualism, and even for human being comprehensive stylistic analysis of literary works can only be carried out by experts. So far, researching conditions of NLP are as follows: Grammar and word

0-7803-8403-2/04/$20.00 ©2004 IEEE

processing are comparatively mature, semantics and sentences under discussion, while pragmatics and discourse analyses far out of the reach. Since stylistics belongs to discourse semantics, difficulties are foreseeable to handle styles with our current techniques. Even though, as early as 60s in 20th century, foreign experts have tried to identify real authors[5] of certain literary works according to their term frequency statistics and sentence length. But this approach essentially is based on word processing instead of discourse semantics, and thus a comprehensive especially aesthetical analysis of styles is beyond the concern.

A new technique, NLP Technique based on Term Connections[6], features in series of Literary Language oriented skills, mainly representing as skills of Knowledge Representation and Acquisition, Language Analyses and Evaluation, as well as Language Creation and Perfection. This paper targets at the outstanding representation poems in traditional Chinese Poetry Art and specifically introduces the related Stylistic Analysis Technique.

## 2. Purpose of Poetry Stylistic Analysis based on Term Connections

The purpose of Poetry Stylistic Analysis mainly discussed in this paper is to distinguish Bold-and-Unrestrained and Graceful-and-Restrained poetry styles, both of which serve as a distinctive classification of poetry styles and bears emphasis in Poetry Stylistic Analysis with difficulties.

In Literature researches, Stylistic Analyses account for an immense number of books. In dynasties of Wei and Jin, the term "style" has come into being, and "Ti" or "Pin" took its place for common use in later generations. In general, "style is features and patterns of a work, which is directly-related to both the contents of the writing and its aesthetic representation, and furthermore style is the reflection of the author's personality such as spirit and mood, talent and culture, taste and interest, etc. "[7]

From the aspect of Poetry Processing, classifications of style have more manipulable meaning. *Stylistic Features* by Liu Xie in his famous collection, *Wen Xin Diao Long*, divides style into 8 categories: elegance, profound, concise, complicated, detailed, splendour, novel and extravagant. *Twenty-four Poems Appreciation* by Sikong Tu employs more subtle division of style. At modern and contemporary times, Chen Wangdao's 4-group-and-8-type classification is more representative: "Group1---concise and complicated, according to content and form proportion; Group2 ---vigorous and tender, according to the strength or mildness in atmosphere; Group3---watery and flowery, according to rhetoric language applications; and Group4---cautious and sparse, according to interline correspondence." [8] Actually, these four groups construct four dimensions, and since it is a more quantitative definition and enjoys higher manipulability, this kind of style classification can be easily carried out in Computer-aided Poetry Stylistic Analysis. For example, Group1 can be evaluated by word number and sentence length, while Group4 is easy to measure with standard grammar rule calculation. Also, aesthetic meanings reveal fully in Group2 and 3, and these two groups concern much about experiences of language, which in itself explains the existing importance and difficulties. Vigorous and tender styles of poetry research in literary fields have their respective "bold-and-unconstrained" and "graceful-and-restrained" as common terms. For instance, we often label a poet's school as Bold-and-unconstrained or Graceful-and-restrained. Thus this paper employs these commonly-used terms, and therefore sets our research goal at Group2--- "bold-and-unconstrained" and "graceful-and-restrained" styles.

## 3. Knowledge Representation Models of both styles

Human beings or computers, a certain set of knowledge is required to distinguish "bold-and-unconstrained" and "graceful-and-restrained" styles, and it's impossible to fulfill the goal only by statistics of word frequency and sentence length.

The basic principle of knowledge representation is to construct a formalized system of pieces of knowledge and their interrelation. Stylistics falls into the category of discourse semantics and it is generalized through the corresponding semantic comprehension of lower layers (paragraph, sentence, word) of the discourse. Similarly, paragraph depends on sentences which depends on words. From the viewpoint of semantics, word is the minimum language unit carrying integrated semantic meanings. So, knowledge representation based on words is the basic idea.

What word semantic meanings contribute as basis of "bold-and-unconstrained" and "graceful-and-restrained" styles? The answer comes to word connotation semantic meanings in reference to knowledge representation based on term connections. Bold-and-unconstrained poems mainly consist of powerful words, while graceful-and-restrained ones include gentle expressions.

Take the word "Rose" for example, and refer to Fig.1 for its semantic structure. Word format semantic meaning includes pronunciation and spelling, which are material representations of language. Word reference semantic meaning is the related object in real world, sensible representation of language. Word experience semantic meaning is experiences gained from objective forms, experiences derive from contents of objects.

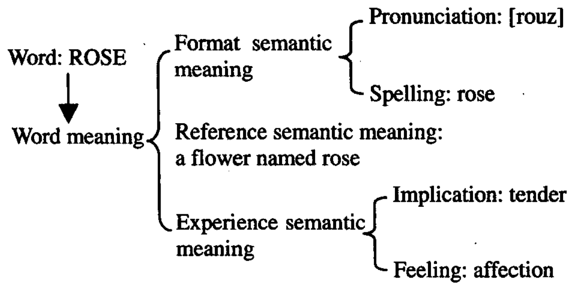

Figure1. Word Semantic Structure

Forms of objects include factors of quantity, body, color, taste, weight, strength, rhythm, meter, speed, quality, etc, which present implication experiences to readers. Some persons such as Formalism Art Critic Claffer Bell (British) argue even to extremes that Art is "a meaning format" [9]. Generally speaking, objects featured with formats of small quantity, little body, soft color, soft sound, thin taste, light weight, slow speed, etc. are easy to arouse our gentle implications, tiled as "elegance", implying "refined, considerate, and soft beauties"; while objects featured with formats of large quantity, big body, bright color, coarse sound, strong taste, heavy weight, fast speed, etc. are more likely to pass powerful implications to readers, named "splendor" in aesthetics, implying "vast, rough and powerful beauties". The above common rules are concluded from longtime art experiences, which combine personal aesthetic experience qualities and object format quantities together, providing great theoretical and practical senses as a result. The macro "bold-and-unconstrained" and "graceful-and-restrained" styles depend on the referred object format quantities. Compare:

北国风光，千里冰封，万里雪飘。 (1)《沁园春·雪》 by Mao Zedong

小城风光，一里冰封，两里雪飘。 (2) A try by author of this paper

Sentence 1 is the famous lines of our bold-and-unconstrained poet, Mao Zedong; Sentence 2 reaches graceful-and-restrained style by reducing object factor levels in volume and quantity of certain words.

Gentle or powerful implications on quantity scales of object format have two challenges. One is to set the reference. Quantity is somewhat relative, and the degrees of gentleness and power have different implications to different people. And therefore subjectivity and unsteadiness lie in personal feelings. The other one is to scale implications. Objects enjoy different levels of gentleness and power even though they belong to the same stylistic category. This very level is advisable to calculate in scales for precision.

Solution to the above two challenges is to apply Expert Trichotomy, i.e. set up an expert group, divide poetry collections into three subsets of gentle, middle and powerful, and carry out necessary manipulations recursively with appropriate levels of recursion performed according to precision requirements. This approach preferably reveals human common aesthetic rules. There are altogether seven subsets in Knowledge Representation Technique based on Term Connections, and poems in the collection are labeled with -3、-2、-1、0、+1、+2、+3 seven different scales. Besides, word implications are relative to its reference semantic meanings, and thus various reference semantic meanings of polysemies point to their respective connotation semantic meanings.

In specific language environments, word reference semantic meanings vary according to its context, and this result in the corresponding variations of word experience semantic meanings. Negative adverbs such as "not", "very", "a little", etc. may shift semantic meanings of the headword (word to be defined). For example: "not---big", "very---majestic" and "a little---push". This influential power on headwords of certain vocabulary is called semantic value, which is ascertained by experts with Expert Trichotomy, varying in 0.25、0.5、0.75、1、1.25、1.5、1.75 seven different scales. Here 1 stands for no influence, and figures below that means decrease level in experience semantic meanings while figures above means increase.

## 4. Algorithm for Poetry Style Analysis

Stylistic Analysis Technique is built upon Knowledge Representation Technique and Language Analysis Technique based on Term Connections. And therefore the input data of "bold-and-unconstrained" and "graceful-and-restrained" style analysis is outputs of Knowledge Representation Technique and Language Analysis.

From the aspect of Knowledge Representation, the set up of poetry labeled corpus, word corpus and term connection corpus is required, among which the latter two serve as data basis of poetry language analysis and language appreciation. According to Knowledge Representation Technique based on Term Connections, human-computer interaction approaches are employed to label target poems in the corpus in order to get the required labeled collection. Then necessary information is retrieved to form term connection corpus, which holds up language rules about term connections, i.e. the specifically applied meaning of connections appeared in the labeled collection. After that, related information is again retrieved from term connection corpus to further construct word corpus, in which word language knowledge is collected, i.e. the all possible would-be word meaning in term connections. Hence, different from traditional knowledge representation skills, the core of term connection based technique is term connection corpus rather than word corpus.

Term connection is a sequential language unit of two words with definite meanings, and can be illustrated in a triplet ($w_a$, $R$, $w_b$). Here, $w_a$, $w_b$ stand for words, $R$ is the relationship between them, and $w_a$ appears before $w_b$ in sentences. When $w_a$ defines features of $w_b$, their relationship is subordinate-center ← with subordinate $w_a$ and center $w_b$. When $w_b$ defines $w_a$, a center-subordinate relationship is formed → with center $w_a$ and subordinate $w_b$. Take the verse sentence “更上一层楼” for example, different kinds of term connections are described as follws:

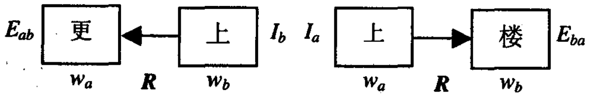

Figure2. Term Connection Examples

In term connection ($w_a$, ←, $w_b$), $I_b$ represents connocation meaning of $w_b$, and $E_{ab}$ indicates semantic value of $w_a$ towards $w_b$. In term connection ($w_a$, →, $w_b$), $I_a$ stands for connotation meaning of $w_a$, and $E_{ba}$ indicates semantic value of $w_b$ towards $w_a$. Both connotation meaning and semantic value are acquired through corpus labeling.

From the viewpoint of Language Analysis based on Term Connections, sentence semantic information is concluded from poetry analysis grounded on word and term connection corpuses. This mainly lies in: confirm word sequences; set up Tree-structure of words; and ascertain word format, reference, connotation meanings together with semantic value towards its center. The basic technical idea of Language Analysis based on Term Connections is analogism and its detailed principles and algorithm are illustrated in another paper [6]. Take “更上一层楼” for example, the corresponding word sequence, Tree-structure and word semantic meanings are shown in Fig.3.

On studies of both word connotation meanings and “bold-and-unconstrained” or “graceful-and-restrained” styles, with supportive data from Knowledge Representation and Language Analysis based on Term Connections, this paper proposes the following definitions about poetry stylistic calculation:

**Definition 1 Word Context Semantic Value *ES*:** Word Semantic Value (in case of no subordinates) or multiplication of Word Semantic Value and Context Semantic Value of its subordinates (if subordinates' number>0).

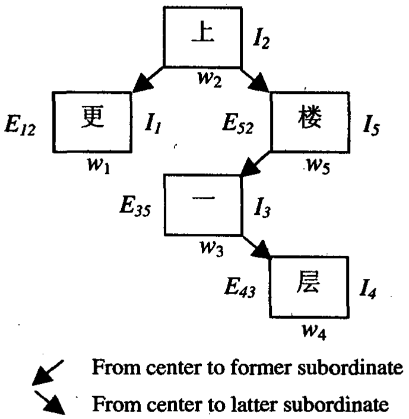

Figure3. Semantic Information from Language Analvsis

This definition is recursive, because subordinate semantic value might influence that of the center, for example: in phrase “not quite well”, semantic value of “quite” towards “well” is influenced by semantic value of “not” towards “quite”. Define $E_{ab}$ Semantic Value of $w_a$ towards $w_b$ and $w_a$ has subordinate subset {$w_i$, $i = 1$, ……, $n$}; and define $ES_{ab}$ Context Semantic Meaning of $w_i$ towards $w_a$, then $ES_{ab}$ is Context Semantic Value of $w_a$ towards $w_b$. So:

$$ES_{ab} = \begin{cases} E_{ab} \times \prod_{i=1}^{n} ES_{ia}, & n > 0 \\ E_{ab}, & n = 0 \end{cases} \tag{1}$$

**Definition 2 Word Context Connotation Meaning *IS*:** Word Context Connotation Meaning (in case of no subordinates) or multiplication of word Context Connotation Meaning and Context Connotation Meanings of its subordinates (if subordinates' number>0). Define $I_a$ Word Connotation Meaning of $w_a$, and $w_a$ has subordinate subset {$w_i$, $i = 1$, ……, $n$}; and define $ES_{ia}$ Context Semantic Value of $w_i$ towards $w_a$, then $IS_a$ is Context Connotation Meaning of $w_a$. So:

$$IS_a = \begin{cases} I_a \times \prod_{i=1}^{n} ES_{ia}, & n > 0 \\ I_a, & n = 0 \end{cases} \tag{2}$$

**Definition 3 Discourse Connotation Meaning *DI*:** the average Context Connotation Meaning of all words in the discourse. If a poem $D$ has word sequence set {$w_i$, $i = 1,\ldots, n$}, and Context Connotation Meaning of $w_i$ is $I_i$, then Connotation Meaning of poem $D$ is $DI$. So:

$$DI = \frac{1}{n}\sum_{i=1}^{n} IS_i \tag{3}$$

**Definition 4 Poetic “bold-and-unconstrained” or “graceful-and-restrained” style *DIZ*:** standard marks of Discourse Connotation Meaning. Poetry Stylistic Analysis turns into Discourse Connotation Meaning calculation via this definition. Discourse Connotation Meaning measuring in countable standard marks helps a lot in the orderliness and comparison of styles. Define *DI* Connotation Meaning of Poem *D* and the collection in which it is contained *DS*. If Average Connotation Meaning of *DS* is $\bar{I}$, and Connotation Standard Subtraction of *DS* is $\sigma$, then Connotation Meaning of Poem *D* is *DIZ*. So:

$$DIZ = \frac{DI - \bar{I}}{\sigma} \qquad (4)$$

The algorithm of "bold-and-unconstrained" or "graceful-and-restrained" poetic style is described as follows:

**Input** Poem *D*: character sequence set $\{c_i, i=1,\ldots,n\}$

**Output** Poetic "bold-and-unconstrained" or "graceful-and-restrained" style

**Step 1** Pretreatments of Poetry Analysis, including calculations of term grouping, semantic meaning tagging, optimal tree searching, etc.

**Step 2** Calculate Word Context Semantic Value $\{ES_i, i = 1,\ldots,n\}$

**Step 3** Calculate Word Context Connotation Meaning$\{IS_i, i = 1, \ldots, n\}$

**Step 4** Calculate Poetic Discourse Connotation Meaning $DI$

**Step 5** Calculate Poetic "bold-and-unconstrained" or "graceful-and-restrained" style *DIZ*.

## 5. Experiments on Poetry Stylistic Analysis Technique

In Definition 4, it is actually a hypothesis to equal Discourse Context Meaning to the manipulability of Poetic "bold-and-unconstrained" or "graceful-and-restrained" style. Though this hypothesis is concluded from longtime human Arts experiences and is based on Aesthetic principles, the Computer-aided Poetry Stylistic Analysis Technique discussed in this paper will prove its validity.

The basic principle of our validation is as follows: select sample poems, comment on their poetic styles both by expert evaluation and with computer analyses. If conclusions from the different experts enjoy obvious correlation, expert evaluation is coherent; and if computer analyses results are obviously correlated with expert evaluation, the algorithm is valid.

Experiments on Poetry Stylistic Analysis can be carried out in the following five procedures:

**Procedure 1** Organize an expert team: 38 Chinese major seniors with necessary background knowledge of Traditional Chinese Poetry Arts.

**Procedure 2** Select sample poems: 55 pieces of poems of Dynasty Tang, 348 lines in total.

**Procedure 3** Expert Evaluation: Print and circulate *Poetic "Bold-and-unconstrained" or "Graceful-and-restrained" Style Questionnaires*. Every member in the expert team is required to fill in their comments individually, marking each poem with a figure from -3、-2、-1、0、+1、+2、+3. Here minus numbers stand for various degrees of "Graceful-and-restrained" Style, zero means no obvious stylistic features and positive numbers show various degrees of "Bold-and-unconstrained" style. The consideration why comments are made on each poetic line is due to the simplicity of capacity and manipulability of marking. Since a poem can be too long or there are possibly shifts in style between lines, it's rather difficult to judge in a whole and comments on poems instead of lines might reduce reliability.

**Procedure 4** Computer Analysis: "Bold-and-unconstrained" or "Graceful-and-restrained" style of each poetic line is calculated with Algorithm mentioned in the above Part 3. Since Expert Evaluation is performed on each individual line, computer analyses also aim at poetic lines in accordance. At present, all the calculations are integrated as modules in Poetic Language Processing Experiment System, on which Computer-aided Poetic Style Analysis is performed.

**Procedure 5** Correlation Studies: carry out correlation studies on conclusions from expert team and outputs of the computer.

Stylistic evaluation on 55 pieces of poems with 348 lines totally in Dynasty Tang receive both data from experts and the computer, amounting to 39 arrays, 348 lines and 13572 original data. Correlation analyses are carried out in data couple of the target 39 factor elements, resulting to 741 correlated coefficient values in Fig.4. Results from Commentator A in expert team and Computer Calculation are compared in Fig.5. The Correlated Coefficient Critical Table tells us $r_{0.01}(346) = 0.137$, $r_{0.05}(346) = 0.105$.

**Analysis 1** All commentators conclude high correlated results and the correlated coefficient values display in normal school. Among 741 correlated coefficient values, 4 values are unobvious, 4 values are obvious in level 0.05, and 733 values obvious in level 0.01. Few commentators have low or high correlation and most of them present middle level correlation results.

**Analysis 2** Commentator A differs a lot from others and thus becomes the core factor influencing the whole correlation. A is unobviously correlated with the other 3 commentators, has obvious correlation with 2 in level 0.05 and 33 in level 0.01 (Fig.5).

**Analysis 3** Computer and Expert Team share high correlation in common, higher than that of Commentator A with other commentators. Computer Analysis is unobvious with only one commentator, is obvious with one in level 0.05 and 36 in level 0.01. Compared with Commentator A, computer analysis has smoother results with the maximum value below Commentator A's maximum value and the minimum value above Commentator A's minimum value (Fig.5). This reveals that computer analysis reflects common ideas of the expert team, and there is no chance of an extremely high or low correlation with few commentators.

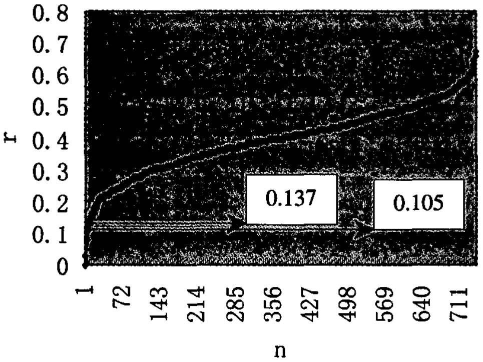

Figure4. Data Couple Correlation in Expert Team

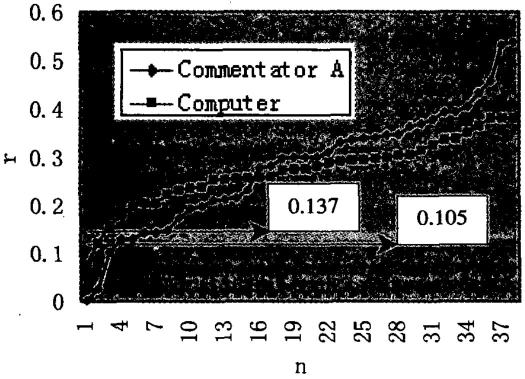

Figure5. Data Couple Correlated Coefficient Value Comparison between Commentator A and Computer

Conclusions are drawn from the above analyses both theoretically and experimentally:

**Conclusion 1** Poetic Stylistic Evaluation by human beings shows commonness rather than individuality.

**Conclusion 2** There are individualities in human comments on "Bold-and-unconstrained" or "Graceful-and-restrained" Style (Analysis 2).

**Conclusion 3** Poetry Stylistic Analysis Technique based on Term Connections is of validity and mainly reveals commonness of human commendations.

From theoretical basis of Poetry Stylistic Analysis Technique based on Term Connections, this technique can be applied in other discourse styles besides poetry. Therefore, it contributes a lot to Literature Theory Research and Creation Practice and further developments of NLP.

Poetry Stylistic Analysis Technique based on Term Connections has been applied in our National Natural Science Funded Project, Computer-aided Literary Art Creation Research ------ Poetry, Ci, Melody and Couplet (I) (grant number: 60173060) and is sure to strive for betterment through practice. "Bold-and-unconstrained" or "Graceful-and-restrained" Style is only one dimension of poetry stylistic features. Later on, efforts will be contributed in more dimensions such as "concise and complicated", "watery and flowery" or "cautious and sparse" styles, so that Poetry Stylistic Analysis Technique based on Term Connections can possess more powerful abilities in Poetry Stylistic Evaluation and Appreciation.

## Acknowledgements

This work is supported by the NSFC under the grant number 60173060.

## References

[1] Dictionary Compiling Office of the Institute of Linguistics, Chinese Academy of Social Sciences. A Modern Pocket Chinese Dictionary. Beijing: Commercial Press,1984,1:577

[2] Institute of Computational Linguistics, Beijing University. Computer-Aided Research System for Ancient Poetry. http://icl.pku.edu.cn/

[3] Henan Centrix Technology Co., LTD. Intelligent Assistant of poetry written.http://www.xianbo.com

[4] Hong-cheng Lin. Computer-Aided Poetry Written Machine. http://www.oligood.com/oldpeasant/web/

[5] Xuesen Qian, Zaifu Liu. Literature, Aesthetics and Modern Science. China Social Sciences Press, 1986:202-7.

[6] Liangyan Li, Zhongshi He, Yong Yi. Principles And Algorithms of Semantic Analysis. Proceedings of 2003 ICMLC. 2003: 1613-1618

[7] Yong-ji Zhao. Essentials of the Ancient Poetry. Tianjin Ancient Books Publishing House.1989: 689

[8] Wang-Dao Chen. Essencial Presentation of Rhetoric. Shanghai Education Publishing House, 2001,7: 264

[9] Clive Bell.Art.China Literature Union Publishing House, 1984:.