ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 1, No. 9, pp. 1191-1196, September 2011

doi:10.4304/tpls.1.9.1191-1196

Probability, Scale of Delicacy and Proximity

Jianqing Wu
School of Philosophy and Sociology, Shandong University, Ji'nan, China;
School of Foreign Languages, Qingdao University of Science and Technology, Qingdao, China
Email: wjq58@hotmail.com

***Abstract*—This paper discusses the semantic constraints of English News probability, scale of delicacy and relationship between proximity and semantic proximity and its measurement method as well as the semantic constraints of the three factors on lexical information. Article introduces "semantic proximity" concept from the mandarin information processing and computer applications. By combing the algorithm of semantic proximity and semantic distance, the summary is obtained: the proposal algorithm of scale of delicacy in second language testing; semantic distance value and the measurement methods of the "semantic proximity".**

***Index Terms*—probability, scale of delicacy, proximity, semantic constraints, news English**

I. Probability

A. About Probability

Halliday has introduced the thought of probability from information theory. People tend to have a certain probability in the choice of words, probability is one of the inherent characteristics of the language, which is most evident regular feature when people make the word choice, for example, the word " Dian ti " (Elevator in Chinese) meaning "lift" or "elevator "embodies a specific word, phrase, or phrases in different languages using different register (Halliday, 1994: F48).

People in the language system in the context of the choice can be restricted by many factors, such as the selection process will inevitably show different language system which determines that the probability can be described. This has already been confirmed by discourse analysis.

The frequency of discourse is actually a probability of the realization in the syntax (HU, 2000, p.17). This can be used for the statistics of probability, the higher the using frequency is, the higher the probability rate is, the higher the probability is.

The chosen semantic news genre of English News Headlines is rather special, different from the conventional semantics. The word choice is to follow the rules of news vocabulary "probability ". Among these rules there are two most important characteristics "shortness" and "concision ", the common purpose of these two features is to save space.

Journalists write "dapper" news stories, Associated Press (AP) specialized in "shortness" as guidelines in the article, states: "Those who can not write articles with simple and powerful words, can not write for the Associated Press" (Wu, 2005, p. 116).

The word with the two "Probability Rules can be called "small words." This article mainly mentions the high probability of "small words". The following are two examples of "small words" in the headline.

Example 1: Mid-East peace Deal Hope

Example 2: All livestock banned across Europe

Example 1 deal to replace agreement, Example 2 ban in place of prohibit, restrain, are all using of high probability rules "small words". Similarly, with the foe instead of enemy or opponent; envoy instead of ambassador; eye instead of witness; air instead of broadcast; cut instead of reduction and so on.

B. The Digital Description of Probability

Probability is one of the core theories of Halliday, he thinks the semantic choice and its reflecting form can not be absolutely rigid description, because the natural language despite its logical side, after all, is "conventional" and influenced the complex Context.

In the description of semantics, he distinguishes four degrees of probability: Either will do, probably, almost certainly, and certainly, marked respectively, 1 / 2, 1 / 2 +, 1 - and 1, marked as 1/2-, 0 + and 0 on the contrary the situation. Later, when discussing the relationship between categories and description of the relationship between them he said: This is "not so much like an 'either-or' relationship, but rather a gradual increase of the 'more or less' relationship "(Halliday, 1955:78; Halliday, 1961:259, cited in Hu, 2000, p. 46).

News Stylistic likes using small words. The probability of using the small word in the Context of general Information is close to 0, but in the news language the probability is close to 1.

II. The Scale of Delicacy

The concept of choosing a word or word meanings is the problem of the "scale" of the "delicacy", referred to as "scale of delicacy". Halliday's "System Grammar" marks the advent of the birth of a language system theory, and introduces "scale" and "delicacy". Christian M. I. M. Matthiessen and Halliday in their book "Systemic Functional Grammar: A First Step into the Theory" states that in order to expand the Words of the spatial dimensions , you can go through delicacy: from the more general to more specific (Matthiessen, Halliday author, Huang Guowen, Hongyang translation, 2009). His new book, "Complementarities in Language" further elaborated on this idea (Li Li, 2010). "Scale of delicacy" is used to describe the detailed description of the semantic delicacy, it is [0,1] continuum, from 0 to 1, becoming the high-order delicacy. One end is the basic level of the concept base category, its scale of delicacy is close to 0, and the other end is the level that can not further distinguish between the syntax, is set to 1, i.e., from the selection of the most common system into the most specific system. "Scale of delicacy" is fit for "scale" of all units, and can be measured. Different languages will be different on the order of the bands. A Sense of the order from general to specific, the more points the more detailed. For example, the general Anglo-American people adopt the concept of probability and the scale of delicacy when describing "weapons" which can be defined as: "guns, pistols, revolvers, machine guns, rifles, knife, knives, daggers, hunting rifles, swords, bombs , grenades, bomb and bayonet "(Wang Su, 1992: 268). The number of scales of delicacy depends on the number of order of words in different meanings of the concept of group-level. Likewise in several scales, the distribution of meanings is related to the delicacy of its corresponding "scale".

Four levels means four scales, based on the various meanings of corresponding scales of four levels of delicacy, it is divided into four scales, low, low, high, high-delicacy. In the scale of delicacy [0,1] interval values, the scale of delicacy of the "non-end scale" of values changes from low to high with the increase in the number of scale. If there are four "scales", then the highest scale is 1, the lowest is 0, the four-scale semantic distance between each other are equal, the other two "non-end scale" value is 0.666, 0.333. If there are five scales, then five are equal portions, ie, 1, 0.75, 0.5. 0.25 and 0, and so on. The higher the scale of delicacy is, the higher the value becomes, it is increasingly close to 1, whereas the lower, close to 0. Items of the same scale of delicacy have the same value. Conventional semantic scale of delicacy is low, special-purpose stylistic has higher semantic scale of delicacy. This paper discusses the Semantic News English and semantic relationships among the three as well as proximity measurement.

III. The Proximity and Its Measurement Methods

The calculation of semantic proximity has a wide range of applications in information retrieval, information extraction, text classification, word sense disambiguation, machine translation based on examples and in many areas alike. The calculation methods of semantic proximity is mainly focused on foreign information technology (Sheth & Kashyap, 1993; Chugur, et al, 2002; Kandola, et al, 2002; Voulgaris, et al, 2004; Ziegler, et al , 2006) and domestic computer applications and information processing (Liu & Song, 2001; Liu Qun, Li Su-Jian, 2002; Li Su-Jian, 2002; Li Bin, Cai DF, 2003; Li Bin, Liu Ting et al, 2003; Guan Yi, Wang Xiaolong, 2003; Yu Chao, Dongfeng Cai, 2006; the summer of 2007; Li Feng, Li Fang, 2007; Wangjia Qin, LI Ren-fa, etc., 2007; He Tingting, Wen Bin, etc., 2008; Wu Kui, ZHOU Xian-secondary, 2010). Calculation of the semantic proximity is to compare proximity between the estimated or calculated words and the selected standard words. Based on the characteristics of English news, this paper combs different calculation methods and screens proximity algorithms in second language tests. In this paper, there is no distinction between the concept of similarity and proximity, the similarity between two words is a word with respect to proximity to another word.

A. The Semantic Proximity

Semantic proximity is a stronger notion of subjectivity. A single definition of semantic proximity can only be obtained in specific applications. The relationship between the words is very complex, and thus the proximity or difference between the values is difficult to carry out with a simple measure. There will be very large differences from different angles. From the perspective of News English, the news genre has very different meanings in the other angles in the conventional meanings of words in the body of proximity. Specifically, the greater the likelihood of syntactic and semantic structure of a meaning in different contexts can be replaced without changing the original text is, the higher the proximity between the two is, otherwise, lower. Proximity is a number, usually in the range [0,1]. The semantic proximity of a word with its own is 1. If the two words in a specific genre, i.e., news genre, can not be replaced, their proximity is 0.

B. Two Important Indicators in the Measurement of the Semantic Proximity

1. Semantic distance

In general, the semantic distance is the real number between [0, ∞). The distance between a word with itself is 0. Semantic distance and semantic proximity has a close relationship. The greater the semantic distance between the estimated words and the standard words provided by testees is, the lower its proximity is; the other hand, the smaller the distance between the two terms, the greater its proximity. A simple correspondence can be established between the two. The correspondence need to meet the following conditions (Liu Qun, Li Su-Jian, 2002):

a. The distance of the two terms is 0, its proximity is 1;

b. The distance of the two terms is infinite, its proximity is 0;

c. the greater the distance between the two terms, the smaller its proximity is (monotonically decreasing).

The semantic distance algorithm of He Ronggui and Lan Yuru(1999) are incorrect (see formula (1)), in order to reduce uncertainty and get the mean square value, the algorithm derived from the semantics of SLA is as follows (see formula (2)):

$$Dis(W_1, W_2) = \sqrt{\sum_M dis(w_1, w_2)^2} \quad .....(1)$$

$$Dis(W_1, W_2) = \sqrt{\frac{\sum_M dis(w_1, w_2)^2}{m}} \quad .....(2)$$

In the formula above, Dis (W1, W2)shows the semantic space vocabulary words W1 and W2 and the linear distance of two points, dis (w1, w2) is the algebraic difference between the coordinates of words W1 and W2 words in the same factors of dimension M. In this experiment, 57 testees identify the scale semantic graph below at the circle the number in their view that "cool" should be in the semantics of the scale (as shown in Figure .2), this concept of different figures represent the direction and strength of the Top value that the testees evaluate. Different views on the concept of testees by the "semantic distance" between the meanings of words show the degree of difference.

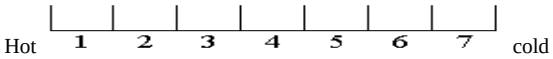
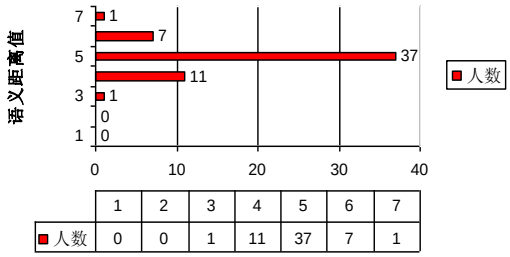


Figure .2 scale of the semantic identification
(Note: the size of the ruler scale is set according to level of delicacy of the experimental data needed, here is set to 7)



	1	2	3	4	5	6	7
人数	0	0	1	11	37	7	1

Figure .3 the data distribution of the distance scale of "cool" in the "hot - cold" semantic meaning rule
(N = 57, the bars refers to the number of testees on each choice)

Statistics shown in Figure .3, according to the formula (2), the final value of 4.929 is obtained, that is, in the rule of the total scale of 7, the semantic distance of "cool" and "hot" is 4.929, and 2.071 with "cold". More people think that "cool" more close to "cold", its proximity with "cold" is much larger than and the proximity with "hot". In many cases, the direct calculation of the proximity of words is rather difficult. We can often calculate the semantic distance first, and then convert to the proximity of words. And the distance and the proximity of words can be two-way conversion.

2. Word correlation

The second indicator is the correlation between words, that the extent of the two words related to each other. It can be measured according to the possibility of two words presenting in the same context. Correlation is a term [0,1] between the real number.

Words correlation and proximity are two different concepts. Such as "factory" and "machine", their proximity is very low, and correlation is very high. The proximity of words reflects the polymer characteristics between the words, and word correlation between the words reflect a combination of features. At the same time, Words correlation and word proximity is closely linked to each other. If two words are very similar, then the correlation of two words associated with other terms will be very high, and vice versa. However, this situation may not reflect the low proximity of true level of testees, so the word correlation in the test should also be cause for concern.

C. The Calculation Method of the Word Proximity

1. Comparison of two common calculating ways

There are two common ways of calculation of semantic distance, one is based on a knowledge of the world (Ontology) to calculate the semantic distance (Agirre & Rigau, 1995; Wang Bin, 1999), another method of calculating word proximity is based on statistics of large-scale corpus. For example, computing the proximity of words by using word correlation. A set of feature words is selected beforehand, and then use this set of words in the actual corpus in the context of the term to measure the frequency, and calculate the characteristics of this group of words and the relevance of each word. This approach assumes that all semantic similar words should also be similar in their context.

Both methods have their own characteristics. The former is mostly used for computer applications and Chinese information processing, simple, effective, more intuitive and easy to understand, but the results were greatly affected by the subjective sense of impact, not always accurately reflect the objective facts. In view of the proximity of the semantic proximity between the research field of linguistics and information processing, this paper chooses the corpus-based method to improve the algorithm. Corpus is more objective and comprehensive reflects the proximity and difference of words in the syntax, semantics, pragmatics and other aspects. However, this approach relies on using the corpus to calculate a large amount of complex calculation, besides, it is affected by sparse data and the data noise interferences, sometimes by obvious mistakes. Liu Qun, Li Su-Jian (2002) provides five calculation methods of semantic proximity, among which the formula (1) and formula (2) are suitable for the testing in the News English, since these two algorithms have overcome the disadvantages of corpus. This paper below respectively calls them the formula (3) and (4).

2. The calculation method of proximity of word meanings

For two words W1 and W2, we note their proximity as Sim (W1, W2), Proximity is the real number between [0,1], its semantic distance is a path length, it is a positive integer, that is the Dis (W1, W2). Then we can define as the conversion relationship of a simple monotonic decline:

$$Sim(W_1, W_2) = \frac{\alpha}{Dis(W_1, W_2) + \alpha} \text{.....(3)}$$

Where α is an adjustable parameter. The meaning of α is semantic distance value when the proximity id 0.5, in the calculation of proximity here, α is set to 1.6. This conversion is not the only relationship where only one of the potentials is given. In order to more scientifically express the estimated true value and reduce uncertainty, this paper adopts the value under conditions of maximum uncertainty as the standard, using the following formula (4), selecting the maximum word proximity. As for semantic proximity of the two words W1 and W2, if W1 has n meanings: S11, S12, ... , S1n, W2 has m meanings: S21, S22, ... , S2m, we require, the semantic proximity of W1 and W2 is the maximum value of statistical calculated proximity per person:

$$Sim(W_1, W_2) = \max_{i=1...n,\, j=1...m} Sim(S_{1i}, S_{2j}) \text{.....(4)}$$

Thus, the proximity problem between the two terms boils down to the proximity between the two meanings. As for the two words we compared here, W1 is the correct answer sample provided by the tester, i.e., the benchmark meanings, W2 is the different semantic options provided by the testees, that is, meanings to be estimated. Based on formula (2), (3), using the formula (4), we can calculate the semantic proximity of semantic options in different language domain of news English.

The subsequent calculation of the semantic proximity is still prevalent in the field of computer applications, Wang Jiaqin, LI Ren-fa (2007) calculated by combining the two coefficients x, y between the concept of semantic similarity. He Tingting, Wen-bin et al (2008:90) proposed a lexical meanings of the similarity of the polarity between the formula and the formula for calculating the degree of emotional words. Wu Kui, ZHOU Xian-Zhong (2010) proposed the semantic similarity algorithm based on the concept of Bayesian estimation. However, these formulae do not apply to News English study, so they are not adopted

IV. Conclusion

The semantics of News English is restricted by three factors, they are, probability, scale of delicacy and proximity. The three are objective as parameters, but subjective as constraints, the learners are proposed to improve cognitive abilities according to the degree of changeability of these constraints.

The relationship among the three constraints is summarized below: the calculation of proximity is constrained by probability and scale of delicacy, semantic proximity becomes different when probability and scale of delicacy differ. In the conventional context, semantic proximity is proportional with probability, and inversely proportional with scale of delicacy. Probability is inversely proportional to the semantic scale of delicacy, that is, the higher the frequency effect is, the lower the semantic delicacy. For News Writing Style has preference of rare semantics of low word frequency in the polysemy, so the probability of the meanings in news genre is relatively low in the conventional context, news lexical semantics is of low probability, high scale of delicacy. But within the news context, the relationship among the three constraints are positive. Learners pay unconscious attention to the cognitive psychology with various styles, including the semantic proximity in the context conversion.

This study has some referential significance in setting the difficulty index, the test index and weight. The proximity reflects the final effect of testee's information output. Domestic and foreign research in the field of Chinese and information technology has been done a lot on the measurement method of semantic proximity, but it has only about a decade of history, China has a late start of academic research in the field of foreign language, and the related research has just started. This paper, by studying the semantic constraints of English news, has discussed and analyzed the relationship among the three constraints, and has found the proposed measurement methods of probability, scale of delicacy and proximity in news language. However, the measurement of semantic proximity should also focus on the

reduction of uncertainty to a greater degree and introduce subjective factors appropriately to make the algorithm perfect; in addition, the inner relation between the proximity and learner's individual Cognition is also worth further exploration.

ACKNOWLEDGEMENT

This paper is financed by Humanities and Social Sciences Fund of Qingdao of University of Science and Technology in 2010, the project number is 10XB15.

REFERENCES

[1] Agirre E. and Rigau G (1995). A proposal for word sense disambiguation using conceptual distance. In International Conference "Recent Advances in Natural Language Processing" RANLP'95, Tzigov Chark, Bulgaria.
[2] Halliday, MAK (1994). An Introduction to Functional Grammar. London: Arnold.
[3] Kandola, J., Shawe, J.,-Taylor & Cristianini, N. (2002). Learning semantic similarity. Proceedings of NIPS '2002. Also at: books.nips.cc.
[4] Laufer, B. (1997). 'The lexical plight in second language reading'. In J. Coady and T. Huckin (eds): *Second Language Vocabulary Acquisition: A Rationale for Pedagogy*, Cambridge: Cambridge University Press.
[5] Liu, Wei-Yi. & Song, Ning. (2001). The fuzzy association degree in semantic data models. *Fuzzy Sets and Systems.* Volume 117, Issue 2, 16 January 2001, Pages 203-208
[6] Sheth, Amit and Kashyap, Vipul. (1993). So Far (Schematically yet So Near (Semantically). Proceedings of the IFIP WG, Citeseer. Also at: portal.acm.org.
[7] Voulgaris, S., Kermarrec, A.-M. & Massoulie, L. (2004). Exploiting semantic proximity in peer-to-peer content searching. Distributed Computing Systems, 2004. FTDCS 2004. Proceedings. 10th IEEE International Workshop.
[8] Ziegler, Cai-Nicolas. Simon, Kai. & Lausen, Georg. (2006). Automatic computation of semantic proximity using taxonomic knowledge. Conference on Information and Knowledge Management, Proceedings of the 15th ACM international conference on Information and knowledge management. Arlington , Virginia, USA. Pages: 465 - 474. Also at: portal.acm.org.
[9] Chugur, Irina., Gonzalo, Julio. & Verdejo, Felisa. (2002). Polysemy and sense proximity in the Senseval-2 test suite. Annual Meeting of the ACL. Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions. Volume 8, Pages: 32 - 39. Also at: portal.acm.org.
[10] Guan Yi, Wang Xiaolong, (2003). statistical computing semantic similarity between Chinese words. Language computing and content-based text processing - National Seventh Joint Conference on Computational Linguistics Proceedings.
[11] He Ronggui, Lan Yuru, (1999). the computerized testing system to identify fuzzy semantics. Eighth International Seminar on Computer-assisted instruction.
[12] He Tingting, Wen Bin, Song Yue, Wang Qian, Luo Le, (2008). word recognition and affective perspective taking tendency of . http://www.ir-china.org.cn/coae2008/ Chinese orientation among the first evaluation Proceedings.
[13] HU Zhuang-lin, (2000). On Functionalism. Beijing: Foreign Language Teaching and Research Press.
[14] HU Zhuang-lin, Zhu Yongsheng, Zhang Delu, Li Zhanzi, (2008). Introduction to Systemic Functional Linguistics (second edition) . Beijing: Peking University Press.
[15] Li Bin, Cai Dongfeng. (2003). Semantic disambiguation methods based on semantic distance and context semantic proximity. Advances in Computation of Oriental Languages - Proceedings of the 20th International Conference on Computer Processing of Oriental Languages.
[16] Li Bin, Liu Ting, Chi, Li Sheng. (2003). On similarity Computing based on semantic-dependent Chinese Sentence. *Application Research of Computers*, No. 12, page 17.
[17] Li Feng, Li Fang. (2007). the Chinese word semantic similarity calculation - based on the "Knowledge Network" (2000) . *Chinese Information Processing*, 3, page 24.
[18] Li Li. (2010). Book review, why the same thing on language interpretation - "Language System and the Association and the complementary". *Foreign Language Teaching and Research*, No. 3, page 57.
[19] Li Su-Jian. (2002). On the semantic relevance of the statements based on calculating. *Computer Engineering and Applications.* No. 7, page 39.
[20] Liu Qun, Li Su-Jian. (2002). The Semantics similarity calculation based on the "Knowledge Network". The 3rd Chinese Lexical Semantics Workshop Proceedings.
[21] Madison, Halliday co-author, Huang Guowen, Wang Hongyang translation. (2009). Systemic Functional Grammar: Introductory of Theory. Beijing: Higher Education Press.
[22] Qian Xu Jing. (2005). The process of meaning guessing and knowledge used of speculation - case studies of word learning. "*World Chinese Language Teaching*", No. 1, page 73.
[23] Wang Bin. (1999). On the automatic alignment of Chinese-English bilingual corpus. PhD thesis, Institute of Computing Technology, Chinese Academy of Sciences.
[24] Wang Jiaqin, LI Ren-fa, Li Chung-sheng, Tang Jianbo. (2007). Ontology-based semantic similarity method of the concept. *Computer Engineering*, No. 11, page 68.
[25] Wang Su. (1992). Cognitive Psychology. Beijing: Peking University Press.
[26] Wang Yulong, Wu Jianqing. (2005). Journalistic English. Beijing: National Defense Industry Press.
[27] Wu Jianqing. (2005). On the language comparison in western general news reports. *Popular Science*, No. 8, page 113.
[28] Wu Kui, ZHOU Xian-zhong, Wang Jian-Yu, Zhao Jiabao. (2010). Semantic similarity estimation algorithm based on the Bayesian concept. *Journal of Chinese Information Processing*, No. 2, page 90
[29] Xia Tian. (2007). Chinese Semantic Similarity. *Computer Engineering*, No. 6, page 77
[30] Yu Chao, Cai Dongfeng, Zhang Guiping. (2006). The analysis of related technologies in the calculation of lexical semantic similarity. Third Student Conference of Computational Linguistics Proceedings.

Jianqing Wu was born in Shandong, China in 1971. He earned his master's degree in Shandong University, China in 2003 and now a doctoral candidate in Shandong University.

He is meantime an associate professor of Linguistics and Journalism in Qingdao University of Science and Technology, Shandong, China. He has been working in QUST for over 16 years. In 2009, he was financed by Chinese government to do his research in MTSU, U.S.A. In the recent decade, he has published 30 articles and 15 books. Some of them are below: Evaluation in Media Discourse Analysis of a Newspaper Corpus. *Journal of Quantitative Linguistics* (SSCI, Netherland), 2010. 3: 256-260; Mini Discourse Training—Review on Discourse Analysis 2nd edition. *Contemporary Foreign Language Research* (CN), 2010. 4: 58-60; The Analysis of Cultural Gaps In Translation and Solutions. *English Language Teaching* (Canada). 2008: 2. "*ENGLISH DEBATE*", China Science and Culture Press, Feb. 2003, author; "*JOURNALISTIC ENGLISH*", Defense Industry Press, Aug.2005, author; "*Practical College English Grammar*", China Commerce Press, Aug.2002, co-author. His Current and previous research interests are discourse analysis, pragmatics and sociolinguistics.

Prof. Wu is a member in professional societies like the Newspaper In Education Research Society (NIE) -China, IQLA, ELT-China, AILA-China as well as Association of China's Sociolinguistics (ACS).