# Developing a Computer-facilitated Tool for Acquiring Near-synonyms in Chinese and English

Shixiao Ouyang, Helena Hong Gao*, and Soo Ngee Koh
Nanyang Technological University, Singapore
{helenagao}@ntu.edu.sg

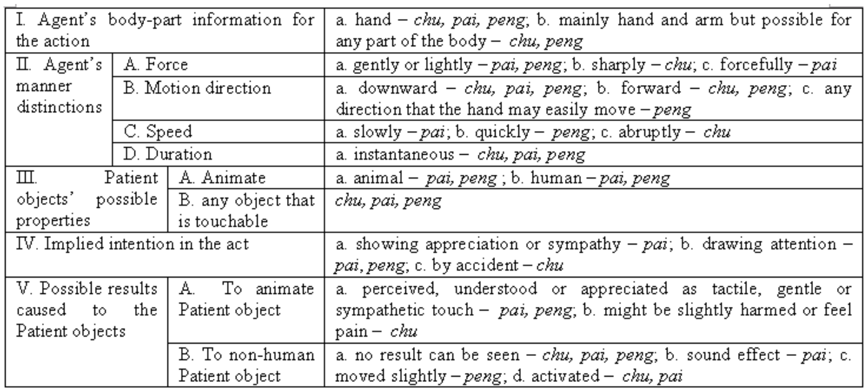
## 1 Introduction

This paper is a multi-disciplinary study on cognitive linguistics, computational linguistics and language acquisition. It focuses on application issues of meaning, semantic structures and pragmatics to near-synonyms in Chinese and English languages. The near-synonyms of physical action verbs (PA Verbs) can be distinctive from each other in the way in which their actions are depicted, but in terms of language acquisition, it is not an easy task to identify the nuances among near-synonyms. Normal dictionaries do not explain the differentiae that are crucial for the choice among near-synonyms, and research on how to tackle the nuances among near-synonyms of PA Verbs has hardly been done. We aim to develop a computer-facilitated language-learning tool for L2 learners to master different sub-classes of PA Verbs in both Chinese and English.

We believe that all the fractions of the arm movement and actions are perceptibly the base for linguistic expressions of human physical actions. By emphasizing the specifications of the manners of action as the crucial semantic components embedded in the verb roots, Gao [1] provided a demonstration of her decomposition method for the semantic properties of near-synonyms of PA Verbs (see Table 1). Differences between the members of each subclass are marked by different notions found in PA Verbs, such as Bodypart Contact, Instrument, Force, Motion Direction, Speed, Effect, and Intention. These are considered to be the most important ones in projecting the lexical semantic prominence in the classifications of a word's meaning components among its near-synonyms.

We assume that to thoroughly understand the nuances between the near-synonyms of PA Verbs, the learners need to understand the perspectives from which the action is depicted linguistically. Gao's [1] work on semantic decomposition of near-synonyms of PA Verbs provides a theoretical guideline and solid ground for designing an e-learning tool for L2 learners.

*Proceedings of the 8th International Conference on Computational Semantics*, pages 316–319,
Tilburg, January 2009.



| | | |
|---|---|---|
| I. Agent's body-part information for the action | | a. hand – *chu, pai, peng*; b. mainly hand and arm but possible for any part of the body – *chu, peng* |
| II. Agent's manner distinctions | A. Force | a. gently or lightly –*pai, peng*; b. sharply – *chu*; c. forcefully –*pai* |
| | B. Motion direction | a. downward – *chu, pai, peng*; b. forward – *chu, peng*; c. any direction that the hand may easily move –*peng* |
| | C. Speed | a. slowly –*pai*; b. quickly – *peng*; c. abruptly – *chu* |
| | D. Duration | a. instantaneous – *chu, pai, peng* |
| III. Patient objects' possible properties | A. Animate | a. animal – *pai, peng* ; b. human –*pai, peng* |
| | B. any object that is touchable | *chu, pai, peng* |
| IV. Implied intention in the act | | a. showing appreciation or sympathy – *pai*; b. drawing attention – *pai, peng*; c. by accident –*chu* |
| V. Possible results caused to the Patient objects | A. To animate Patient object | a. perceived, understood or appreciated as tactile, gentle or sympathetic touch – *pai, peng*; b. might be slightly harmed or feel pain – *chu* |
| | B. To non-human Patient object | a. no result can be seen – *chu, pai, peng*; b. sound effect – *pai*; c. moved slightly –*peng*; d. activated – *chu, pai* |

Table 1: Specification of Semantic Properties for Verbs of Touching [1]

## 2 Methodology

A near-synonym database needs to be set up first. To make the semantic representations abstract enough for computing purpose, a rule based model will be applied to quantify the semantic properties of each word. The following is a rule showing how the semantic meaning of the word peng 'touch' is written in relation to its other class members. The rule representation follows a standardized English Machinese according to certain conventions in machine translation [2].

```
% near-synonym peng of chu and pai %
peng3(smc(c,_9), pav) --> (bp(3/3),{bp <3/3>,[hand]}),
(ins(2/3),{ins<0/3>,[possibly not]}),
(for(3/3),{for<2/3>,[gently]};{for<2/3>,[lightly]}),
(md(3/3),{md<3/3>,[downward]};{md<2/3>,[forwards]};{md<1/3>,[any direc-
tion that the hand may easily move]}),
(sp(3/3),{sp<1/3>,[quickly]}),
(du(3/3), {du<3/3>,[instantaneous]}),
(po(3/3),{po<2/3>,[animal]};{po<2/3>,[human]};{po(3/3),[any object that is
touchable]}),
(int(3/3);{int<1/3>,[drawing attention]}),
(eff(3/3),{eff<2/3>,[to human: perceive, understood or appreciated as tactile,
gentle or sympathetic touch]};eff<3/3>, [to nonhuman: no visible result];
{eff<1/3>,[to nonhuman: moved slightly]}).
% pinyin {peng [peng4]} %; % tone marked by number 1,2,3 or 4 %
```

The rule above can be rendered in words as follows: There is a Chinese lexical item *peng* classified as a physical action verb (*pav*). It has three near-synonyms (*peng3*) with 9 semantic property types categorized in relation to its synonyms (*smc(c,_9)*). The detailed rule for the 9 categories begins with the agent's body part involvement (*bp*), followed by the instrument used in performing the action (*ins*), degree of the force used (*for*), the motion direction of the action (*md*), the speed of the motion action (*sp*), duration of the action (*du*), possible patient object that the action is executed on (*po*), the subjective agent's possible intention to perform the action (*int*), and the possible effect caused to the objective patient (*eff*).

The numbers given in the brackets such as (*3/3*) or (*2/3*) and in the angle brackets such as <*3/3*> or <*1/3*> are mathematical expressions where the digits in the denominator in both ( ) and < > indicate the number of near-synonyms in the class. The digit in the numerator in the bracket ( ) refers to the number of the class members that can be labeled as bearing the same semantic property category, while the digit in the numerator in < > refers to the number of the class members that share a particular semantic property specified in square brackets [ ]. For instance, the actions depicted by all three near-synonyms of touching verbs involve hand contact as the body part involvement and thus this part of rule is written as {*bp<3/3>,[hand]*}.

The content in { } has a conditional relation: If the number in the numerator is 3 in brackets < >, then the property specified in [ ], such as [hand] in the body part involvement, must be met by all its class members. If the class member bears or shares more than one semantic properties within one category, the alternatives are given with a semicolon in between the bracket { }, as typically shown in the motion direction (*md*) of the action. In the beginning and at the end of the rule representation, brief comments or reminders of the next step rules are given after % as conventionally used in Prolog programming [2].

The second step is to calculate the average degree of uniqueness of each word in each of the 9 categories specified in the rule, representing the relative degree of easiness or difficulty to differentiate the word from the rest of the group within the category. Algorithm for calculating the average degree of uniqueness is proposed below:

Step 1: For each of the 9 semantic property categories, a priority check is conducted first to look for the digit "1" in the numerator in < >. If "1" is found in the numerator in at least one < >, go to Step 2; otherwise go to Step 3.

Step 2: In case of "1" found in the numerator in a < >, the average degree of uniqueness is marked as unit "1" , the highest degree of the semantic representation in revealing the uniqueness of its kind in terms of semantic categories.

Step 3: If "1" is not found as a unique property within any category, assign the value of the numerator in each < > as $A_1$, $A_2$,..., $A_n$, where n indicates the total number of properties that the word owns in each respective category. The average degree of uniqueness represented by this category can be expressed in the following formula: $(1/\text{n})\sum_{i=1}^{n} 1/A_i$.

The average degree has a maximum value of 1. For those words whose degree of uniqueness is lower than 1, the higher the degree of uniqueness, the lower the value for n or $A_i$. With respect to distinguishing themselves from the rest of the class members, they have fewer competitors, leading to larger probabilities of differentiating the near-synonyms by the corresponding semantic property. The near-synonym database will contain one main database linked to a number of sub-tables, which can be more or less than the 9 categories illustrated above, depending on the number of semantic property categories that each class falls into, or whether the specifications are specified enough to be able to discriminate a particular class of near-synonyms under consideration in the future. Values of uniqueness owned by each word found in each semantic property are stored in the sub-tables. The average values of uniqueness owned by each word on per category basis are recorded both in its sub-table and the main table.

In creating the algorithm for the interface, the steps are in reverse order of building up the database. The total value of uniqueness computed will be compared on per category basis. The category with the highest sum will be chosen as the sub-table for the first attempt of differentiation. Similarly, a primary check of the existence of value "1" is conducted in order to give priority to the category where most "1"s are found. Only when no value "1" is found in any field could the arithmetic summation be carried out. Multiple choices questions will be prompted to allow users to select one property from the category that matches his/her context most closely. The number of the word(s) that match such property will be reflected by the value of the numerator in the <> of the chosen property. The algorithm will continue to search for the next category with the highest degree of uniqueness among the remaining near-synonyms, until only one word is selected from the class.

## 3 Conclusion

With a solid ground of linguistic analysis and application of rule based computational methodologies, the e-learning program provides advanced L2 learners with an effective interface in acquiring the nuances among near-synonyms with more intuitional and cognitive understanding. This computational system serves as a good starting point for exploring e-teaching and e-learning tools for advanced learners.

## Acknowledgements

Research grants to H. H. Gao and President Research Scholarship to S. Ouyang from Nanyang Technological University supported preparation of this paper.

## References

[1] H. H. Gao. *The physical foundation of the patterning of physical action verbs.* Lund, Sweden: Lund University Press, 2001.

[2] B. Sigurd and H. Gao. Outline of a computerized chinese grammar enabling english and swedish translation. In *Working Papers*, volume 47, pages 181–199. Lund University Press, 1999.