Difficulty and Ambiguity of Verbs : Analysis based on Synsets in Japanese WordNet

Miyata, Koki
Faculty of Information Science and Electrical Engineering, Kyushu University

Suzuki, Takahiko
Research Institute for Information Technology, Kyushu University

Hirokawa, Sachio
Research Institute for Information Technology, Kyushu University

出版情報：Proceedings of 1st International Conference on Advanced Information Technologies, 2013-11-28. IIAI Publications
バージョン：
権利関係：

Difficulty and Ambiguity of Verbs

－Analysis based on Synsets in Japanese WordNet－

Koki Miyata
Faculty of Information Science and Electrical Engineering, Kyushu university
Fukuoka, Japan
2IE13089S@s.kyushu-u.ac.jp

Takahiko Suzuki, Sachio Hirokawa
Research Institute for Information Technology, Kyushu University
Fukuoka, Japan
{suzuki, hirokawa}@cc.kyushu-u.ac.jp

***Abstract*— When foreign students learn Japanese words, they encounter two types of problems. The first is the difficulty of the word itself. The second problem is related to the situation which it is used. The meaning of the word may differ as the situation changes. It is necessary to understand the ambiguity of the word. In this paper, we propose three simple formulas representing ambiguity of words based on Synset structure in Japanese WordNet. Then we analyze the relationship between the ambiguity and the difficulty of Japanese words, particularly verbs. We use vocabulary level of Japanese-Language Proficiency Test (old version) as the difficulty measure. The result shows that easy (not difficult) words are more ambiguous than difficult words.**

Keyword WordNet, Japanese-Language Proficiency Test (JLPT), ambiguity of verbs, difficulty of words

I. INTRODUCTION

In recent years, the necessity for studying Japanese as a second language has increased with the increase of the foreign students in Japan. There are many researches in English as Second Language (ESL) [1]. Generally, expressing various concepts appropriately using a limited number of basic words is regarded as one of the key issues in second language learning [2]. However, expressions which are composed of basic words do not necessarily become comprehensive. To make comprehensive expressions, it is required to select unambiguous words according to the situation. Sometimes, comprehensive expressions cannot be composed only by using basic words. For example, a Japanese expression using a basic verb 取る (take):

“その　酒を　取って (Take the glass of alcohol)” is ambiguous and not comprehensive. The meaning of the expression can be “Pass me the glass of alcohol”, “Drink it”, or “Remove the glass of alcohol in front of me”.

Generally, even a foreign student who is fluent enough in Japanese conversation when one can reconfirm the meaning of speech on the fly tend to be poor in writing.

In this paper, we will introduce measures that indicate the ambiguity of words and will investigate the relationship between the ambiguity and the difficulty of words. The purpose of this paper is to supply data about what kind of words should be studied by foreign students in Japanese language learning when focusing on the ambiguity of the words.

We propose three kinds of quantitative measures of ambiguities based on the Synset [3] structure in Japanese WordNet[4,5].

The word lists in JLPT (Japanese Language Proficiency Test) [6] is used as references to the difficulty of the words. We use The JLPT level 1 word list as the difficult word set and the level 2-4 as the easy word set.

The relationship between ambiguity and difficulty was investigated for nouns and verbs respectively. Interesting relationships between the ambiguity and the difficulty are found, particularly in verbs.

The rest of the paper is organized as follows:

Section 2 is an introduction of related works. Section 3 introduces WordNet and JLPT. Section 4 explains the measures of ambiguity proposed in this paper. Section 5 is the analysis results. Section 6 is the discussions on the relationships between ambiguity and difficulty. Section 7 is a conclusion and about future works.

II. RELATED WORKS

Maeda [8] evaluated the difficulty of the words based on the frequency of English words. Kawamura [9] estimated the difficulty of the Japanese words by using the level of the JLPT. Kitamura et.al [10] used IDF (inverse document frequency) in estimating the difficulty of words. There are various researches about WSD (word sense disambiguation) [11, 12]. [13] is a research on the ambiguity of the words in WordNet. Lists of advanced English vocabulary words which should be learned by ESL students can be found elsewhere [7].

III. WORDNET AND JLPT

A. *WordNet*

WordNet [3] is a large lexical database. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms named as “Synset”, each expressing a distinct concept. Synsets are interlinked by conceptual-semantic and with lexical token relations. WordNet can be used as a thesaurus because words are grouped in Synset by their meanings. In this research, we use Japanese WordNet [4, 5] based on WordNet3.0. Japanese WordNet contains 57,238 concepts (Synset) and 93,834 words. We used Perl interface WordNet::Multi [4] in order to access Japanese WordNet data. Fig.1 shows a part of the structure in Japanese WordNet

for the Japanese word "労働 (meaning work:verb)". The graph in Fig. 1 shows corresponding Synsets (concepts) and synonyms in Japanese WordNet. The word 労働 is shown in the left ellipse. Two gray boxes in the center designate two Synsets for the word 労働. The right ellipses show synonymous words. "労働(work:verb)" has two meanings. i.e. ID:02413480-v: "exert oneself by doing mental or physical work for a purpose or out of necessity" and ID: 02410855-v: "be employed". There are four synonyms 立ち働く, 労働 , 勤労 and 働く including the word itself in the 1st Synset 02413480-v. Six synonyms 労働 , 勤労, 働く, 就労, 務める and 勤務 are in the 2nd Synset 02410855-v. Three of synonyms 労働, 勤労 and 働く overlap in the two Synsets.

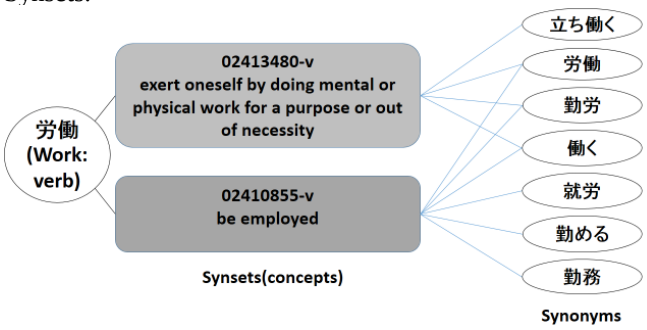

Figure 1. Synsets(concept) and synonyms of the word 労働(work:verb)

B. *Japanese Language Proficiency Test(JLPT)*

Japanese Language Proficiency Test (JLPT) [6] is an examination for the purpose of evaluating and certifying the Japanese proficiency of non-native speakers. JLPT had 4 levels until 2009. Level 3 and level 4 had measured the level of understanding of basic Japanese mainly learned in elementary classes. Level 1 and level 2 had measured the level of understanding of Japanese used in a broad range of scenes in actual everyday life. Vocabulary lists for each level had been published in[15]. There are web pages which contain the lists of the vocabulary words such as[14].

The JLPT level 1 requires linguistic competence necessary to manage general social life such as reading newspaper editorials or critiques comprehensively. It assumes about 900 hours of learning in Japanese language. The standard of the JLPT level 1 had included advanced grammars, 2,000 characters of kanjis (the Chinese characters), and 10,000 Japanese vocabulary words.

Level 2-4 had been less demanding.

JLPT was revised in 2010, N3 was newly established as an intermediate level. From then on the examination consists of 5 levels. The vocabulary lists for each level of JLPT becomes non disclosed information after the revision.

C. *THE COVERAGE RATIO OF JLPT WORDS IN JAPANESE WORDNET*

The current version of Japanese WordNet does not cover all words in JLPT word lists. Table I shows the coverage ratio of JLPT words in Japanese WordNet

TABLE I. The coverage of JLPT words in Japanese WordNet

	# of Words	# in J-WN	Coverage
JLPT 1	2981	2359	79%
JLPT 2-4	4380	3780	86%

In this paper, we use those words both in the JLPT word lists and in Japanese WordNet.

IV. MEASURES FOR THE AMBIGUITY OF WORDS

We assume the ambiguity of words can be identified by three features listed below.

(1)Numbers of the meanings of the polysemous word

(2) Complexity of the meaning of each word

(3)The difficulty in distinguishing a certain meaning of the polysemous word from other meanings

We propose three measures, M1, M2, and M3 for each word in Japanese WordNet as quantitative measures corresponding to (1), (2), and (3).

(M1)The number of Synsets which contain the target word as a synonym

(M2)The average count of synonyms per Synset which contains the target word

(M3) The degree of semantic overlap between Synsets which contain the target word

Precise definitions of M1, M2, and M3 are as follows.

(M1) The number of Synsets #Syn(w)

The number of the Synsets which contains the word w.

(M2) The average count of synonyms per Synset

#Word(Syn(w))/#Syn(w)

The denominator #Syn(w) is the same as M1. The nominator #Word(Syn(w)) is the total sum of the count of synonyms in each Synset which contains the word w. If the same synonymous word appeared in two or more distinct Synsets, we count them twice or more.

(M3) The degree of semantic overlap Ovlp(Syn(w))

First we define #Word_uniq(Syn(w)). It is the sum of the count of synonyms of w which appear only once throughout all the Synsets which contain w. Ovlp(Syn(w)) is defined as follows.

$$\mathrm{Ovlp}(\mathrm{Syn}(w)) = \left(\#\mathrm{Word}(\mathrm{Syn}(w)) - \#\mathrm{Word\_uniq}(\mathrm{Syn}(w))\right) / \#\mathrm{Word}(\mathrm{Syn}(w))$$

M1,M2, and M3 values for the example 労働 (work:verb) shown in Fig.1 are as follows.

As Two Synsets (02413480-v and 02410855-v) contain 労働 then M1 for 労働 is 2 (the word 労働 has 2 meanings).

The total number of arcs which connect one of the right-hand side synonymous-word nodes to one of the Synsets

equals to #Word(Syn(w)). In Fig.1, the number of arcs (M2=#Word(Syn (労働))) is 10.

#Word_uniq(Syn(w)) equals to the number of the synonymous words which appeared only in one Synset containing the word w.

In Fig.1, such words are 立ち働く, 就労, 勤める, and 勤務. Then #Word_uniq(Syn(w)) is 4 and

M3 =(10-4)/10=0.6

We count all synonyms even when some of them are not in JLPT word lists.

V. RESULTS

We have computed M1, M2, and M3 for diff(icult) words (JLPT1) and for easy words(JLPT2-4). Nouns and verbs are counted separately. The result is shown in Table II.

All words categorized as verbs in WordNet are counted as verbs even when they are not categorized as verbs in JLPT. The result is shown in Table II.

TABLE II. Relationships between the ambiguity and the difficulty for nouns and verbs

	# of words	M1	M2	M3
verb:easy	1289	4.54	11.4	0.400
verb:diff	906	3.26	11.6	0.343
noun:easy	2556	3.41	9.33	0.336
noun:diff	1732	3.03	9.43	0.316

F test shows that there are significant difference between JLPT1(difficult) and JLPT2-4(easy) in M1 and M3 (For M3 of nouns P <= 0.05, for others P <= 0.02).

The comparison of M1 and M3 are consistent (JLPT2-4 > JLPT1 and verbs > nouns).

Fig.2 and Fig.3 show cumulative distribution of M1 and M3 respectively. X-axis is the number of Synset.

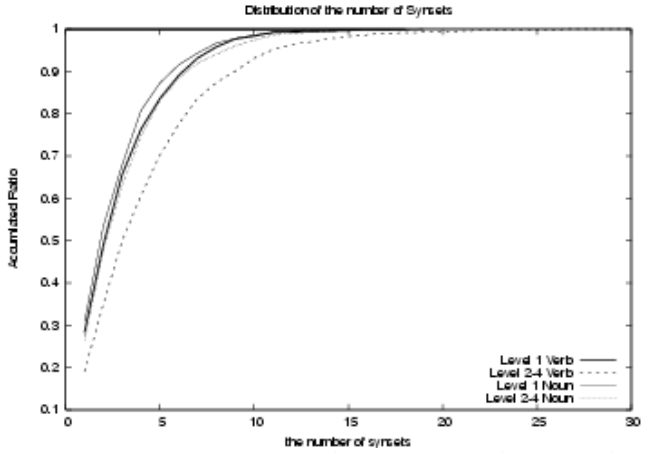

Figure 2. Cumulative distribution of M1 (the number of meanings of the word)

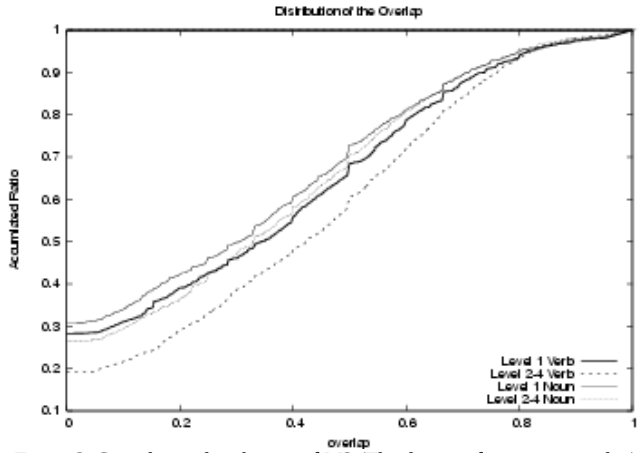

Figure 3. Cumulative distribution of M3 (The degree of semantic overlap)

Both Fig. 2 and Fig. 3 indicate that the distributions for verbs in JLPT2-4 (dot line) differ from the other distributions. It means the ambiguity (measured by our definition) of JLPT2-4 (easy) verbs are higher than other vocabulary words.

Although there always are difference between nouns and verbs, no significant difference is found in M2 between easy and difficult vocabularies. M2 (the average number of synonyms in Synsets containing the target word) is not an appropriate measure of ambiguity.

The values of M1 and M3 are not independent. The value of M3 (semantic overlap) will increase as the number of Synset for the target word increases.

Fig. 4 shows the average values of M3 for different value of M1 (M1 from 2 to 10).

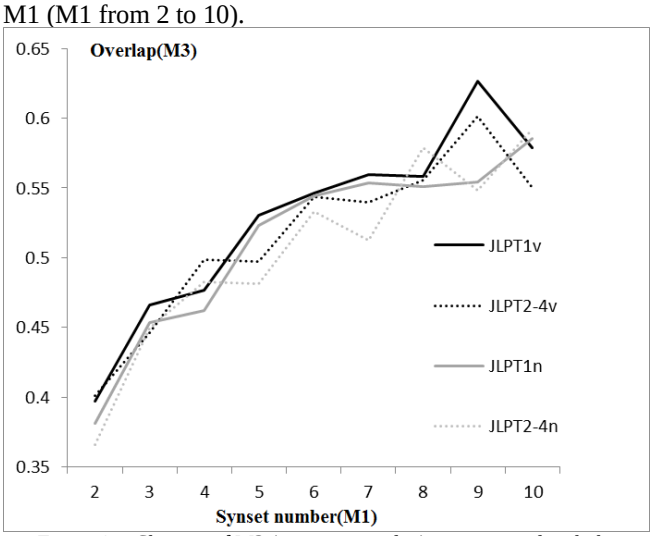

Figure 4. Changes of M3 (semantic overlap) accompanied with the changes in M1

There are no significant differences in M3 between easy and difficult words when grouped with the same M1 value.

VI. THE RELATIONSHIPS BETWEEN DIFFICULTY OF WORDS AND THEIR AMBIGUITY

In section 5 we have shown that the easy words, particularly easy verbs are more ambiguous than other

vocabulary words. There are some hypotheses which may explain this phenomenon.

Maeda et.al. [8] argued that the difficulty of a word is related to how frequent the word is used, and that the difficult words are less frequently used. As a result, two things may happen. First, since the difficult words are used in the limited situations, the original meanings of the words are preserved well thus remains unambiguous. Second, since there are few examples of use for difficult words, the granularity of a classification of the meaning (Synset) in Japanese WordNet may differ from those of easy vocabulary words.

In the case of difference in verbs, we have the following hypotheses. Since a meaning of a verb is determined by the dependency to the subject and the object corresponding to the verb, the meaning of the verb may have a plentiful variety as it can take various kinds of subjects and objects. Therefore, the number of meanings (M1) of an easy verb becomes larger. A difficult verb is used with limited kinds of subjects and objects, so the number of its meanings becomes small.

Japanese WordNet has the common Synset structure with original English WordNet. In English, there are phrasal verbs such as 'go to', 'get out', etc. Such phrasal verbs in English which are composed from easy words may influence the result in this paper.

VII. Conclusion and Future Works

In this paper, we proposed three measures for ambiguity of a word based on the Synset structure of Japanese WordNet. In M1 (the number of Synsets) and M3 (semantic overlap), there were correlations between the measures and the difficulty of the word. Remarkable correlations are found in verbs and there is a consistent tendency that the difficult verbs have less ambiguity.

If we apply the above result directly in the field of language learning, the conclusion "difficult verbs with little ambiguity should be used in order to describe a comprehensive Japanese sentence"" can be drawn, which is contrary to the common sense.

The difficulty of Japanese words has not been a thoroughly researched subject. Oshio et.al. [16] tried to define a new difficulty ranking by making JLPT level as an initial rating and extending it by using dictionaries and Web texts.

JLPT was revised in 2010. It had been announced [17] the publication of the new vocabulary list classified by new difficulty level, but the list has not been published until 2013. There is a room for discussion whether the JLPT level used by our research indicates the difficulty of words.

However there is a simple way to utilize the result of this paper in Japanese language learning for foreign students. That is, one should learn at least one unambiguous synonym for each meaning of the verb when he/she learns easy but ambiguous verbs.

The ambiguity of language is not decided only by ambiguity of each word. The composition, dependency and the other factors in whole texts affect.

We are trying to apply a word sense disambiguation system (WSD) to texts and determine the ambiguity of words in real context. It will supply the other materials for the discussion in this paper "whether we should use difficult words (verbs) in order to compose comprehensive texts."

In this paper, we analyzed only Japanese WordNet. It is interesting to investigate the other languages such as English.

References

[1] M. R. Salaberry, The use of technology for second language learning and teaching: A retrospective, Modern Language Journal 85 (1), pp. 39-56, 2001

[2] Basic-English Institute, BASIC ENGLISH: International Second Language, http://ogden.basic-english.org/isl.html

[3] Princeton University "About WordNet." WordNet. Princeton University. 2010, http://wordnet.princeton.edu

[4] NICT Information Analysis Laboratory, National Institute of Information and Communications Technology, Japanese WordNet, http://nlpwww.nict.go.jp/wn-ja/index.en.html

[5] Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki, Enhancing the Japanese WordNet, The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP 2009

[6] Japan Foundation, Japan Educational Exchanges and Services, Japanese Language Proficiency Test, http://www.jlpt.jp/e/index.html

[7] Murray Bromberg, Melvin Gordon, 1100 Words You Need to Know, 6th edition, Barrons Educational Series Inc, 2013

[8] Joyce Maeda, Frequency Level Checker, http://language.tiu.ac.jp/flc/

[9] Yoshiko Kawamura, Analysis of text for reading by using the vocabulary checker, Lectures in Japanese Language Education (in Japanese), 34, pp.1-22, 1999

[10] Tatsuya Kitamura, Yousuke Tomioka, and Yoshiko Kawamura, Construction and verification of the Word Level Decision System by using IDF, Journal of Japanese Language Education Methods (in Japanese). 16(1), 52-53, 2009

[11] Roberto Navigli, Word Sense Disambiguation: A Survey, ACM Computing Surveys, .41 (2), No.10, 2009

[12] Claudia Leacock , George A. Miller, Martin Chodorow, Using Corpus Statistics and WordNet Relations for Sense Identification, Computational Linguistics - Special issue on word sense disambiguation archive, 24 (1), pp.147-165, 1998

[13] Roberto Navigli, Meaningful clustering of senses helps boost Word Sense Disambiguation performance COLING/ACL 2006 - 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference 1, pp.105-112, 2006

[14] Jonathan Waller, http://www.tanos.co.uk/jlpt/

[15] International exchange fund, Japanese Institute of International Education, "The Standard of the Japanese Language Proficiency Test Questions", Bonjin Sha (in Japanese) , 1994

[16] K. Oshio, J. Ishige, et.al., Toward creation of the vocabulary list for New Japanese Proficiency Test, Japanese Language Education Bulletin, Japan Foundation (in Japanese), pp.71-86, 4, 2008

[17] H. Nakanishi, N. Kobayashi, H. Shina, Estimation method of difficulty rating of Japanese words based on clustering by using dictionary data and Web data, Proc. 19th Annual Conference of The Association for Natural Language Processing (NLP19) (in Japanese), pp.682-685, 2008