A Preliminary Study on the Impact of Lexical Concreteness on Word Senses Disambiguation *

Oi Yee Kwong

Department of Chinese, Translation and Linguistics, and
Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
rlolivia@cityu.edu.hk

Abstract. Psychologists have shown that abstract words are harder to understand and often acquired later than concrete words. In this work, we study how the difficulty of automatic word sense disambiguation (WSD) might be affected by this intrinsic property of words, namely the concreteness of a word and its individual senses. We also explore the feasibility of inducing a numerical index for sense and lexical concreteness from dictionary definitions. Analysis of system performance in previous SENSEVAL exercises suggests that concrete words are often easier to disambiguate. The high overall agreement between human ratings and definition-induced ratings is also encouraging. The concreteness factor is worth the attention of computational linguists, particularly in terms of how it bears on the differential information demand of individual words in WSD and how the knowledge of this property could be employed to fine-tune WSD systems to better deal with the lexical sensitivity of the task.

Keywords: Lexical concreteness, Word sense disambiguation

1. Introduction

The SENSEVAL (and recently SemEval) exercises have revealed a lot of issues on automatic word sense disambiguation (WSD), and allowed researchers to learn more about the linguistic and technical aspects of the task. System performance often depends on many factors, including the feature set, availability of training instances, and language models, amongst others. One important linguistic factor is the fine-grainedness of the sense inventory and the semantic closeness among the senses of a word. To this end, Resnik and Yarowsky (1997) suggested that closely related senses are more difficult for WSD, and therefore systems should be penalised less if they fail to distinguish between similar senses than if they fail to tell distinct senses apart.

Despite being a psychologically valid and intrinsic property of words and senses, *concreteness* is seldom addressed in WSD literature. Psychologists have shown, from lexical decision and naming tasks, that abstract words are harder to understand than concrete words, and are often acquired later (e.g. Bleasdale, 1987; Kroll and Merves, 1986; Yore and Ollila,

* The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 1508/06H).

1985). This thus implies differential underlying mechanisms in the representation, development, and processing of word meanings in the mental lexicon. By analogy, the inclusion of the concreteness information in computational lexicons should also benefit natural language processing (NLP) tasks like WSD, in addition to maneuvering only linguistic and technical factors. It should also allow us to study polysemy and sense similarity in a more comprehensive and cognitively plausible way.

Hence, we start on a preliminary study on the relationship between concreteness and WSD. In particular, given the lexical sensitivity of the task, we are interested to see how this psycholinguistic factor bears on the difficulty of WSD, how it could be employed in fine-tuning WSD systems to accommodate the different information demand of individual target words, as well as its potential impact on evaluative measures.

In the first part of this study, we analyse system performance reported in previous SENSEVAL exercises with respect to the concreteness of the target words. In the current study, we focus on noun samples only. In the second part, we explore the feasibility of automatically inducing a more objective and robust concreteness measure from dictionary definitions, which is needed for enhancing lexical resources and benefiting WSD in the long run. In general it was observed that concrete words are often easier to disambiguate. The high overall agreement between human ratings and definition-induced ratings is also encouraging.

In Section 2, we briefly review related work and further set out the background of this study. In Section 3, we report on our analysis of system performance with respect to the concreteness of target words. In Section 4, we outline the relation between definition styles and concreteness, and describe our work on automatically inducing a concreteness measure from surface syntactic forms of dictionary definitions. The results are further discussed with future directions in Section 5, followed by a conclusion in Section 6. In this paper we use “lexical concreteness” and “sense concreteness” as a generic term for the degree of concreteness of words and senses respectively, from highly abstract to highly concrete.

2. Background

Many psycholinguistic studies on lexical processing confirmed that abstract words are harder to understand than concrete ones. For instance, concrete words are often found to lead to shorter reaction times than abstract words in lexical decision tasks (e.g. Bleasdale, 1987; Kroll and Merves, 1986). Such concreteness effect is concurrently under the influence of various lexical, semantic, and even personal factors, including word frequency, imageability, and context availability (DeGroot, 1989; Kroll and Merves, 1986; Schwanenflugel, 1991).

The observed difference between the two kinds of words also implies a somewhat different mechanism by which they are stored, represented, connected, and processed in the mental lexicon. While there were studies investigating the relationship between lexical access and polysemy (e.g. Swinney, 1979), few have addressed the relation between concreteness and polysemy, and WSD. Analysis on word association responses, for instance, has suggested that tangible concepts seem to be more easily activated than abstract concepts; and in the case of polysemy, tangible senses appear to be more accessible than abstract senses (Kwong, 2007). However, concreteness is often discussed only at the lexical level. We must also look into concreteness at the sense level in order to study its impact on the information demand of individual words in WSD, and hence varied information susceptibility (Kwong, 2005).

Concreteness is often measured by means of human ratings on an ordinal scale from highly abstract to highly concrete (e.g. Paivio *et al.*, 1968). Scalability is essential for its application in WSD and other NLP tasks, and to this end, we need to find ways to automatically induce an objective measure of concreteness which is comparable to human judgements. Lexical data reflecting human lexical processing is possibly available from various resources, including dictionary definitions, word association norms, lexical and knowledge bases, as well as corpus data from authentic texts. In the current study, we explore the feasibility of simulating human judgements on concreteness from dictionary definitions.

It has been suggested that WSD systems should be less penalised if they fail to distinguish between closely related word senses than if they fail between distinct senses. This issue of sense similarity is addressed by Resnik and Yarowsky (1999) with quantitative characterisation in terms of sense proximity, and by Chugur *et al.* (2002) in terms of sense stability.

WSD is often considered a lexically sensitive task, in which individual target words might vary in their difficulty and require different treatment. Pedersen (2002) assessed the difficulty of test instances in the SENSEVAL-2 English lexical sample task by analysing the agreement among participating systems.

We base our analysis on the target nouns for the English lexical sample tasks and system performance reported in SENSEVAL-1 and SENSEVAL-4 (officially SemEVAL-2007). Different systems might vary in the disambiguating information and computational approaches they use. Nevertheless, in terms of performance on individual target words, they sometimes complement one another and sometimes achieve similar results. It thus suggests that there is hardly a universal set of parameters which will work in precisely the same way and satisfactorily for all target words. That is to say, words have different *information susceptibility*.

Moreover, with the experience from four SENSEVAL exercises, systems should have matured in one way or another to cope with the lexical sensitivity of WSD. If a difference still persists among target words within individual systems, there must be something else intrinsic to the words and senses themselves that has not been adequately recognised and effectively addressed by automatic WSD systems. We are thus interested in how lexical concreteness, as a psycholinguistic factor and an intrinsic property of words, bears on the difficulty of WSD and information demand of individual target words, in addition to other linguistic and technical factors. More importantly, we should explore how we could capitalise on such a relationship to fine-tune WSD systems and shed light on WSD evaluation.

3. Concreteness and WSD Difficulty

In this section, we outline the procedures in selecting word samples and discuss the results on comparing human concreteness ratings with system performance on individual target words as reported in previous SENSEVAL exercises.

3.1.Materials

Target nouns from the English lexical sample tasks in SENSEVAL-1 and SENSEVAL-4 (Task 17) were selected. There are 15 and 35 target nouns in these two exercises respectively, as listed in Table 1. The column "S" refers to the number of senses in WordNet 3.0, and the column "C" refers to the average of human ratings on lexical concreteness, as discussed below.

Sense definitions were collected for these words from WordNet 3.0. The average number of senses per word for SENSEVAL-1 nouns is 3.67, and the words have 1 to 9 senses. The average for SENSEVAL-4 nouns is 7.94 senses per word, and the words have 1 to 26 senses. Note that the HECTOR sense inventory was used for SENSEVAL-1, and some had very different degrees of polysemy as reported here. For example, "knee" has as many as 22 senses. On the other hand, OntoNotes senses were used for sense distinction in SENSEVAL-4 (Task 17). For the current study, we use WordNet 3.0 senses as a common reference for both sets of words.

For data on WSD difficulty, we made use of the task and system reports, and results summaries from SENSEVAL, assuming WSD difficulty of test words is reflected from system performance on individual words. For SENSEVAL-1 data, we refer to the official scores under "fine-grained, all systems, average" as available from http://www.senseval.org. Precisions and recalls were reported, and we computed the F1 measure for convenience in comparison. System performance for SENSEVAL-4 (Task 17) is based on the average results on individual target words from all systems, as well as results from two individual systems (System 1 and 4, both using Support Vector Machines), reported in Pradhan *et al.* (2007).

Table 1: Target Nouns from SENSEVAL Data

SENSEVAL-1			**SENSEVAL-4** (Task 17)					
Word	**S**	**C**	**Word**	**S**	**C**	**Word**	**S**	**C**
accident	2	4.33	area	6	4.00	network	5	3.00
behaviour	4	3.00	authority	7	3.33	order	14	2.67
bet	2	4.50	base	19	3.67	part	12	3.33
disability	1	2.00	bill	10	5.00	people	4	6.67
excess	4	1.50	capital	6	4.00	plant	4	6.67
float	7	4.50	carrier	11	4.67	point	26	4.67
giant	7	4.50	chance	5	1.33	policy	3	2.00
knee	3	7.00	condition	8	1.33	position	16	3.00
onion	3	6.50	defense	11	2.33	power	9	3.00
promise	2	2.00	development	9	2.67	president	4	6.00
rabbit	3	6.50	drug	1	6.33	rate	4	1.33
sack	9	4.50	effect	6	2.00	share	5	2.67
scrap	4	4.00	exchange	10	3.67	source	9	2.33
shirt	1	6.50	future	3	1.33	space	9	3.67
steering	3	4.00	hour	4	2.33	state	7	3.33
			job	9	4.00	system	9	2.67
			management	2	3.33	value	6	1.33
			move	5	4.00			

3.2.Method

Three human judges were asked to rate the words and senses in the sample on a 7-point scale of concreteness, with 1 for highly abstract, and 7 for highly concrete. Ratings were to be given to all words (ignoring individual senses) first, and then independently to each sense. They were asked to do the rating according to their intuition and subjective evaluation, although it was also suggested that imageability could be used as a criterion in their judgement without precluding other relevant factors. One of the judges was an undergraduate student and the other two were graduates. All have studied linguistics before.

Based on the average human ratings, we divided the words into three categories along the concreteness continuum: Abstract (with average rating below 3.0), Medium (with average rating between 3.0 and 5.0 inclusive), and Concrete (with average rating above 5.0).

Analysis and comparison were done with respect to the following: (a) agreement among the human judges at both the word level and sense level; (b) difference in mean performance with respect to word categories based on lexical concreteness; and (c) difference in mean performance with respect to word categories based on sense concreteness.

3.3.Agreement among Human Raters

The Kendall's Coefficient of Concordance W was computed to assess the agreement among the human raters. At the word level, an overall W of 0.767 was found among our three judges on SENSEVAL-1 words, 0.706 on SENSEVAL-4 words, and 0.727 on all words, all statistically significant. This suggests that at the word level, the raters in general agree with one another on positing the word samples on the lexical concreteness continuum, although the absolute ratings they have assigned to individual samples might differ. The agreement, however, is less strong at the sense level, but overall medium to high agreement has been observed.

In the subsequent analysis, we used the average of the rating from the three raters as a measure for concreteness of the words and senses, and divided the words into three categories as explained in Section 3.2.

3.4.Impact of Concreteness on WSD

We compared the mean performance with respect to the three categories of words along the concreteness continuum. Given that we only used secondary data sources for the analysis, we have not controlled for the number of samples in each category, but only depend on the ratings assigned by the human judges on the target nouns. Hence the number of samples in each group is small and the distribution may not be even. Nevertheless, it happens that the datasets in SENSEVAL-1 and SENSEVAL-4 do contain examples for all three groups, and they allow us to start on a preliminary analysis of the relationship between concreteness and WSD difficulty.

The comparison results are shown in Tables 2 and 3. There were only 30 words from SENSEVAL-4 used in the comparison instead of 35, since results for five words were omitted in Pradhan *et al.* (2007).

We also compared the mean performance with respect to the concreteness of the first sense, which is assumed to be the predominant sense according to WordNet ordering, and also the average of the concreteness of all senses. Note that the first sense and the average of all senses may or may not correspond to the concreteness rating at the lexical level, as the lexical concreteness depends on the human raters' intuition, personal experience and their understanding of the word in general without considering individual senses. The results are shown in Tables 4 and 5.

With the datasets and small number of samples, we were not able to establish a statistically significant difference for the mean performance among different word categories on the concreteness continuum. Nevertheless, we had some interesting observations. From SENSEVAL-1 data, it appears that nouns at both ends of the continuum, i.e. either very concrete or very abstract, are better disambiguated than those lying in the mid range of the continuum. With reference to the impact of the first and supposedly predominant sense or the average concreteness of all senses, a similar trend was found. Moreover, words with an abstract predominant sense or more abstract senses in general tend to be even better disambiguated than those with more concrete senses. This is an interesting phenomenon which deserves more in-depth investigation and qualitative analysis.

Table 2: Mean Performance by Word Concreteness for SENSEVAL-1 Data

Group	# Samples	Mean Performance
Abstract	3	0.7235
Medium	8	0.6150
Concrete	4	0.7384

Table 3: Mean Performance by Word Concreteness for SENSEVAL-4 Data

Group	# Samples	Avg	Sys 1	Sys 4
Abstract	12	77.42	85.67	75.00
Medium	15	79.27	87.70	78.00
Concrete	3	92.67	96.00	90.33

Table 4: Mean Performance by Sense Concreteness for SENSEVAL-1 Data

Group	# Samples	Mean Performance
Concreteness Based on First Sense		
Abstract	3	0.7757
Medium	5	0.5815
Concrete	7	0.6871
Concreteness Based on Average of All Senses		
Abstract	4	0.7573
Medium	5	0.6073
Concrete	6	0.6632

Table 5: Mean Performance by Sense Concreteness for SENSEVAL-4 Data

Group	# Samples	Avg	Sys 1	Sys 4
Concreteness Based on First Sense				
Abstract	12	83.83	88.92	80.75
Medium	14	75.29	85.36	74.43
Concrete	4	87.00	89.75	85.50
Concreteness Based on Average of All Senses				
Abstract	5	78.00	86.20	75.00
Medium	20	78.55	86.65	77.00
Concrete	5	87.00	91.40	87.60

On the other hand, analysis on SENSEVAL-4 data seems to yield results closer to our expectation. With reference to word concreteness and average concreteness of all senses, words toward the concrete side tend to be better disambiguated than those in the mid range, which are in turn better disambiguated than those on the abstract end. However, looking at the predominant sense, a similar situation to SENSEVAL-1 data was found, i.e., those with the first sense at either end of the continuum are better disambiguated. One possible reason is that very concrete or very abstract senses are expected to occur in more characteristic linguistic contexts, which could be more successfully captured by the features used in WSD systems. Hence, it remains for us to see in more qualitative terms how this concreteness effect could affect the effectiveness of various kinds of disambiguating information.

4. Concreteness from Definitions

According to McKeown (1991), "a definition can be seen as an attempt to capture the essence of a word's meaning by summarizing all of its applications and possible applications". Although nouns are expected to be relatively easy to define, as compared to other parts-of-speech, various defining styles are observed (Jackson, 2002). A common type is by means of genus (superordinate concept) and differentiae (distinctive features). For words which are not easy to be defined by a genus term, the definition is often composed with one or more synonyms or a synonymous phrase. Another kind of definitions is by means of prototype, which is similar to the genus and differentiae type but in addition specifying what is typical of a referent with words like "typically" or "usually". For others, where a referent is unlikely to be available, lexicographers will capture their meanings in a dictionary by explaining their usage in real text. It is also commonly realised that tangible objects and physical actions are more easily defined in dictionaries, while abstract concepts and other aspects of meaning including connotation, sense relations, and collocations are less readily and often only partially covered by the definitions.

Hence, we assume that the concreteness of a concept will make a difference on the most appropriate defining style. Specifically it will be more difficult to define abstract concepts by means of genus and differentiae, and prototype, and they are more likely to be defined by synonyms and other means. We therefore analysed dictionary definitions and distinguished them into seven categories based on their surface syntactic forms, corresponding to a 7-point scale (7=highly concrete, 1=highly abstract) which is assumed to correlate with various levels on the concreteness continuum from human judgements. The definitions used in this study were obtained from WordNet 3.0. The seven categories are listed and explained in Table 6[1].

Each sense definition for the SENSEVAL-1 nouns was classified into one of the seven types of definitions exemplified in Table 6. The category assigned to each sense definition was thus taken as a numerical indication of the concreteness of the respective meaning on a 7-point scale.

The definition-induced concreteness measures agree relatively strongly with the average human ratings at the sense level, with a statistically significant Kendall's W of 0.680.

With the definition category assigned to each sense definition, lexical concreteness was induced from two conditions. One is to use the category value from the first and presumably dominant sense of a word. We call this condition DefOne. The other is to take the average of the category values from all senses of a word, and we call this condition DefAll. We tested for the correlation with the Spearman rank correlation ρ and agreement with Kendall's W between average human ratings and the definition-induced values. The results are shown in Table 7.

Table 7 shows that the correlation between human ratings and definition-induced ratings is not particularly strong and linear, but the overall agreement is nevertheless quite high. With this particular dataset, apparently the average of sense concreteness models lexical concreteness more reliably. However, this is only a preliminary attempt and more study and refinement remain to be done.

5. Discussion and Future Work

Our results thus show that concreteness bears some relation with WSD difficulty as a psycholinguistic factor superimposing on the linguistic factors like sense distinction and sense similarity, which are often believed to directly affect the difficulty of WSD.

Since psychological evidence suggests that words of different concreteness are represented and accessed by possibly different mechanisms, this factor is worth the attention of computational linguists, particularly in terms of how it bears on the differential information demand of individual words in WSD and how the knowledge of this property could be employed to fine-tune WSD systems to better realise and deal with the lexical sensitivity of the task, as well as its potential impact on evaluative measures. For example, future evaluation might consider balancing the number of concrete and abstract test items.

A potential limitation of our current categorisation of the dictionary definitions is that abstract concepts might be defined by genus and differentiae more often than expected. For instance, one meaning of "accident" is "an unfortunate mishap; especially one causing damage or injury". This may be an artifact of WordNet definitions since WordNet places each sense in a hierarchy of hyponymy relation, which covers both concrete and abstract concepts. Words like "mishap" are nevertheless abstract even when they are used as the genus term for other words. We plan

[1] Only a simple version is given here. Refer to Kwong (2008) for more detailed descriptions.

to check against other dictionaries and explore possible ways to deal with various kinds of genus terms, to refine the concreteness index induced from definition categories.

Table 6: Categorisation of Dictionary Definition Styles

Category	Patterns	Explanation and Examples
7	Surface pattern: *Determiner* + (*Modifier*) + *Genus* + *Differentiae* + *Prototype* e.g. *car* – a motor vehicle with four wheels; usually propelled by an internal combustion engine	Concrete concepts are usually defined in terms of genus and differentiae. High imageability is assumed if a prototype could also be described.
6	Surface pattern: As above with either *Differentiae or Prototype* but not both e.g. *bag* – a flexible container with a single opening; *cup* – a small open container usually used for drinking	Assume slightly less concrete if no distinctive feature or prototype is captured.
5	Surface pattern: 1. a + *(Modifier)* + {kind,type} of + *Genus* + *Differentiae/Prototype* 2. *Determiner* + *(Modifier)* + *Genus* 3. someone + *Differentiae/Prototype* e.g. *husband* – a married man; *officer* - someone who is appointed or elected to an office and who holds a position of trust	A less detailed description of the concepts but at least a person or some known membership
4	Surface pattern: 1. a + (Modifier) + {kind, type} of + Genus 2. {somewhere, something, etc.} + Differentiae/Prototype 3. a + (Modifier) + {set, number, collection, etc.} of + Genus + Differentiae/Prototype e.g. body – a collection of particulars considered as a system; mercy – something for which to be thankful	Empty kernels or underspecified objects, but still describable in terms of distinctive features
3	Surface pattern: 1. Det + (Modifier) + {state, instance, etc.} of + Genus(mass noun) 2. (Det) + (Modifier) + Genus(mass noun) + (Differentiae/Prototype) e.g. hour – clock time; glory – brilliant radiant beauty	Unlike tangible objects and physical actions, more abstract concepts are less feasibly and less likely to be defined in terms of countable genus and differentiae.
2	Surface pattern: {your, the} + mass noun + of/to + (Modifier) + mass noun / countable noun in plural form / a gerund e.g. hatred – the emotion of intense dislike; idea – the content of cognition	Mass nouns are often more abstract, and the abstraction often doubles up in patterns in this category involving two mass nouns.
1	All others, including explanation of usage e.g. baby - sometimes used as a term of address for attractive young women	Presumably highly abstract concepts need to be explained more verbosely in other forms.

Given that human ratings on concreteness may be a result of the interaction of many factors including word frequency, context availability, imageability and access to sensory referents, etc., it will be appropriate for us to resort to other sources of external evidence such as word association norm data, authentic linguistic context from corpus data, and domain information, in addition to dictionary definitions, for a more realistic and complete model of lexical concreteness.

As suggested earlier, words at either end of the concreteness continuum might occur in relatively characteristic linguistic contexts. So, more importantly, further studies will be conducted to examine the effect of lexical and sense concreteness on the information demand of automatic word sense disambiguation and the use of concreteness for indicating potentially confusable senses for better evaluation of disambiguation performance.

Future work could thus go in several directions: (1) further investigation in automatically inducing an objective concreteness measure from lexical resources which simulates human ratings, (2) studying the contextual nature of words with different concreteness from natural texts to explore the difference in information demand by individual target words and thus their information susceptibility in WSD, (3) applying the findings on the relation between concreteness and disambiguation performance in WSD in turn, to fine-tune the systems, and (4) enriching lexical resources with the intrinsic property of words in terms of concreteness level.

Table 7: Reliability of Dictionary-Induced Ratings

Condition	ρ	W
DefOne	0.240	0.619
DefAll	0.418	0.709

6. Conclusion

In this paper we have discussed our preliminary study on the impact of lexical concreteness on the difficulty of automatic word sense disambiguation. By comparing system performance on target words of different concreteness from previous SENSEVAL exercises, it was found that concrete words are more easily disambiguated in general, and words at either end of the concreteness continuum are better disambiguated than those in the middle. We also explored the feasibility of simulating human judgements on the concreteness or abstractness of words via dictionary definitions. The overall agreement found between human ratings and definition-induced ratings is encouraging, and more language resources will be employed in future work on the simulation of a numerical index for lexical and sense concreteness. Such an index is believed to inform not only lexicography but also WSD and other NLP tasks. Since psychological evidence suggests that words of different concreteness are processed by possibly different mechanisms, this factor is worth the attention of computational linguists as it bears on the information demand of individual target words in WSD.

References

Bleasdale, F.A. 1987. Concreteness dependent associative priming: Separate lexical organization for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 582-594.

Chugur, I., J. Gonzalo and F. Verdejo. 2002. Polysemy and Sense Proximity in the Senseval-2 Test Suite. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, pp.32-39.

DeGroot, A.M.B. 1989. Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 824-845.

Jackson, H. 2002. *Lexicography: An Introduction*. London and New York: Routledge.

Kroll, J.F. and J.S. Merves. 1986. Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 92-107.

Kwong, O.Y. 2008. A Preliminary Study on Inducing Lexical Concreteness from Dictionary Definitions. In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS-2008)*, Barcelona, Spain, pp.84-93.

Kwong, O.Y. 2007. Sense Abstractness, Semantic Activation and Word Sense Disambiguation: Implications from Word Association Norms. In *Proceedings of the 4th International Workshop on Natural Language Processing and Cognitive Science (NLPCS-2007)*, Funchal, Madeira, Portugal, pp.169-178.

Kwong, O.Y. 2005. Word Sense Classification Based on Information Susceptibility. In A. Lenci, S. Montemagni and V. Pirrelli (Eds.), *Acquisition and Representation of Word Meaning*. Linguistica Computazionale, pp.89-115.

McKeown, M.G. 1991. Learning Word Meanings from Definitions: Problems and Potential. In P.J. Schwanenflugel (Ed.), *The Psychology of Word Meanings*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., pp.137-156.

Paivio, A., J.C. Yuille and S.A. Madigan. 1968. Concreteness, Imagery, and Meaningfulness Values for 925 Nouns. *Journal of Experiment Psychology, Monograph Supplement, 76(1, Pt.2)*, 1-25.

Pedersen, T. 2002. Assessing System Agreement and Instance Difficulty in the Lexical Sample Tasks of SENSEVAL-2. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, pp.40-46.

Pradhan, S.S., E. Loper, D. Dligach and M. Palmer. 2007. SemEval-2007 Task 17: English Lexical Sample, SRL and All Words. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, pp.87-92.

Resnik, P. and D. Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for Word Sense Disambiguation. *Natural Language Engineering, 5(2)*, 113-133.

Resnik, P. and D. Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington D.C.

Schwanenflugel, P.J. 1991. Why are Abstract Concepts Hard to Understand? In P.J. Schwanenflugel (Ed.), *The Psychology of Word Meanings*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., pp.223-250.

Swinney, D.A. 1979. Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18*, 645-659.

Yore, L.D. and L.O. Ollila. 1985. Cognitive development, sex, and abstractness in grade one word recognition. *Journal of Educational Research, 78*, 242-247.