Measuring Concept Concreteness from the Lexicographic Perspective *

Oi Yee Kwong

Department of Chinese, Translation and Linguistics, City University of Hong Kong,
Tat Chee Avenue, Kowloon, Hong Kong
Olivia.Kwong@cityu.edu.hk

Abstract. The distinction between concrete and abstract concepts is psychologically valid but so far it can hardly be quantified in any objective way, which prevents it from being further studied in computational linguistics. This paper proposes a systematic way to measure concreteness from the surface structure of dictionary definitions. Comparing the scores from WordNet definitions with human ratings, the method apparently works better for concrete senses than abstract ones, and is not adequate for measuring concreteness on a finer scale beyond the simple dichotomous distinction. Future work thus includes exploring the possibility of deploying a wider range of surface clues from definitions for the purpose and investigating how the method works with definitions from different dictionaries.

Keywords: Concept concreteness, Dictionary definitions, Lexicography, Word senses.

1 Introduction

Most people might find it more than normal that concrete concepts are easier to understand and learn than abstract concepts. Such common sense is supported by ample psychological evidence from lexical decision tasks and children's spoken and reading vocabulary (e.g. Bleasdale, 1987; Kroll and Merves, 1986; Yore and Ollila, 1985). However, few studies have addressed how this apparently trivial observation might imply on the mental storage and organisation of words and their meanings, and whether modelling such distinction between concrete and abstract concepts in computational semantic lexicons might benefit natural language processing tasks like automatic word sense disambiguation. For example, how can concreteness be included in common ontological organisation of semantic lexicons? If concrete concepts are easier for people than abstract ones, should we expect concrete senses to be more easily disambiguated than abstract senses? If this is the case, should we take it into account when evaluating disambiguation performance, perhaps in addition to the similarity among senses (e.g. Resnik and Yarowsky, 1999; Palmer *et al.*, 2006)? Should different information be employed for disambiguating concrete and abstract senses?

One major obstacle to this kind of study is the subjectivity in deciding what is concrete and what is abstract, which is a matter of degree and confounded by many factors. In psychology studies, concreteness (or abstractness) has often been measured by averaging human ratings on a sample of words. Human raters give a score for each word on a scale (e.g. from 1 for highly abstract to 7 for highly concrete), often based on various factors like the familiarity and imageability of the concepts, and frequency of occurrence, but more heavily on their own

* The work described in this paper was supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 1508/06H) and the Department of Chinese, Translation and Linguistics of the City University of Hong Kong.

experience and probably private cognition. This way of measuring concreteness is limited in scalability. In addition, ratings on word samples do not say anything on the effect of polysemy. A concrete word may also have some abstract senses, and vice versa, and the raters may not be thinking of the same sense when rating a word.

Only if we can have a way to objectively measure concreteness will it be possible for us to pursue further studies on the role of such intrinsic nature of word senses in lexical resources and natural language processing. In the current study, we investigate the feasibility of obtaining such a measure from dictionary definitions. The surface linguistic structure of a given definition is analysed and mapped to a concreteness score, and the results are compared to human ratings.

In Section 2, we will explain the rationale underlying our approach. The experiment and results will be discussed in Section 3, followed by a conclusion with future directions in Section 4.

2 Our Approach

The way we propose to measure concreteness is based on a simple intuition. Dictionaries constitute an important source of our lexical knowledge. Language learners often form their perception and understanding of words from dictionary definitions. Professional lexicographers are trained to write definitions informatively and consistently, and it is generally assumed in lexicography that concepts corresponding to tangible objects or intangible things are more appropriately defined by different styles. Dictionary definitions thus reflect how lexicographers (who are also human beings) perceive the concepts being defined.

2.1 Definition Styles

The way in which definitions are structured and phrased has undergone considerable evolution in modern lexicography, from "lexicographese" to full-sentence definitions (Atkins and Rundell, 2008). Traditional models tend to explain words through relatively short and concise definitions, in different defining styles, as summarised in Jackson (2002). Examples for various defining styles are shown in Table 1.

Table 1: Examples for Different Defining Styles

Defining Style	Example
Genus + Differentiae	**[bag]** a flexible container (genus) with a single opening (differentiae)
Synonymous Phrase	**[carbon]** carbon paper
Prototype	**[car]** a motor vehicle with four wheels; usually propelled by an internal combustion engine (prototype)
Usage	**[baby]** sometimes used as a term of address for attractive young women
Full-Sentence Definition	**[audience]** The audience at a play, concert, film, or public meeting is the group of people watching or listening to it.

A common way to define a concept is by means of genus (superordinate concept) and differentiae (distinctive features), such as "bag" is defined as a kind of "container" distinguished from other containers by being flexible and having a single opening. For words which are not easy to be defined by a genus term, the definition is often composed with a synonym, a collection of synonyms, or a synonymous phrase. Sometimes it may not be easy to isolate the sufficient and necessary conditions for a sense, and lexicographers will capture the essential constituents in the form of prototype. This is usually combined with genus and differentiae but additionally specifies what is typical of a referent with words like "typically" or "usually". For the rest, where a referent is unlikely to be available, meanings will be explained

through their usage in real text. While tangible objects and physical actions are more easily defined with genus, differentiae and prototype, abstract concepts as well as other aspects of meanings like connotation and collocations often need to be defined by other means. Full-sentence definitions, on the other hand, use more natural prose and often embed the word being defined in its typical syntactico-collocational context in the definition (Hanks, 1987). Although this style has been thoroughly used by the COBUILD series of dictionaries, it has not really dominated the dictionary market as anticipated but continues to co-exist with most conventional defining styles in many other dictionaries. In the current study, we focus on the surface structure of the various traditional defining styles for clues on concreteness.

2.2 Scoring Definitions by Surface Structure

Computational linguists have made use of dictionary definitions to semi-automatically acquire simple ontologies for nouns (e.g. Chodorow *et al.*, 1985; Vossen *et al.*, 1989). For instance, apart from real genus, Vossen *et al.* (1989) have also observed other pseudo genus like empty kernels (e.g. a kind of …) and those shifting the definition to some non-NPs (e.g. a manner of speaking …), which they called linkers and shunters respectively in the LINKS Project.

In this study, we also exploit the regularity exhibited in definitions to reveal how concreteness is perceived in the eyes of lexicographers. Based on extensive observation and analysis of dictionary definitions, we started with the definition patterns roughly outlined in Kwong (2008) and fine-tuned them with respect to the outputs given by a dependency parser, and implemented a scoring system which analyses the surface structure of a definition. The basic assumption is that the more concrete a concept, the more conveniently and convincingly it can be explained with reference to its superordinate concept and distinguishing features. Scores are thus assigned according to the presence or absence of various surface structures. Figure 1 roughly maps various patterns to a 7-point scale of concreteness.

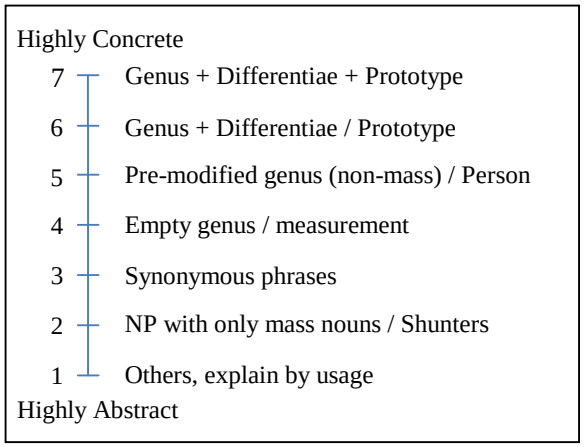


Figure 1: Mapping Definition Patterns to 7-Point Scale

To analyse the structure of the definitions, we make use of the dependency parser from Lund University (Johansson and Nugues, 2008) and detect the various definition patterns by means of the presence or absence of certain dependency relations obtained from the parse results. Only those parses consisting of a root (ROOT) and a predicative complement (PRD) in the form of a noun will be of interest to us. Any of the following dependency relations on the PRD would be considered differentiae: Apposition (APPO), Location (LOC), Modifier of nominal (NMOD), Object complement (OPRD), and Temporal (TMP). If any of such relations has a dependency from words like "usually", "often", "typically", etc., it would be considered a prototype.

The automatic scoring is thus done this way: Given the dependency parse of a definition (such as "car is a motor vehicle …"), we first look for the ROOT (usually "is") and its dependent PRD. If a PRD as a noun is found, it is treated as the genus and an NP definition is assumed to be identified and further processed. Subsequent analysis of an NP definition includes detecting genitives like "X of Y", checking for differentiae in the form of dependent APPO, LOC, NMOD, OPRD, or TMP to the PRD, and prototypes if such dependencies are marked by words like "usually" or "typically". If no differentiae, prototype, or pre-modifier is present, the definition will be considered a synonymous phrase. The genus is also checked if it is possibly an abstract or mass noun, which is currently approximated by the absence of any indefinite articles and plural markers. The scoring flow is summarised in Figure 2 in the form of pseudocodes with examples.

```
Let Genus X = noun at PRD
Let Score = 4
if (X…of…Y) is found, then
    if X ∈ {kind, type, etc.} and Y is mass, then Score=Score-1        // e.g. a kind of commercial enterprise ...
    else if X ∈ {group, part, etc.} and Y is mass, then Score=Score-1  // e.g. a series of related events …
    else if Y is –ing verb, then Score=Score -2                        // e.g. the manner of speaking to …
    else {if X is mass, Score=Score-1; if Y is mass, Score=Score-1}    // e.g. the content of cognition …
    stop
if X ∈ {person, someone, anyone, etc.}, then Score=Score+1, stop       // e.g. someone who controls resources …
if X ∈ {something, somewhere, etc.}, then stop                         // e.g. something intended as a guide …
if ∃ Dependency D ∈ {NMOD, APPO, LOC, TMP, OPRD} ←X at any             // prototype
    word after X and D is preceded by {usually, typically, often,etc.},   e.g. … usually used for drinking
    then Score=Score+1
if ∃ Dependency D ∈ {NMOD, APPO, LOC, TMP, OPRD} ← X at any            // differentiae
    word after X, then Score=Score+1                                   e.g. [a motor vehicle] with four wheels
if no prototype and no differentiae is found and X is mass,            // synonymous phrase with mass noun,
    then Score = Score -1                                                 e.g. brilliant radiant beauty
else if ∃ Dependency D ∈ {NMOD}← X at any word before X,               // pre-modifiers,
    then Score=Score+1                                                    e.g. a very young mammal
if no prototype and no differentiae and no pre-modifier,               // minimal NP treated as synonym
    then Score=Score-1                                                    e.g. an idea
```

Figure 2: The Scoring Flow

3 Experiment

In the current study, we first used the Lund University dependency parser[1] to parse dictionary definitions, then analysed the parse output according to the above scoring method and compared the results with human ratings.

3.1 Data sources

Kroll and Merves (1986) used 100 abstract and 100 concrete nouns in their lexical study. The words were rated by human subjects for concreteness on a 7-point scale. The abstract and concrete words were matched on the basis of word frequency and word length. The word frequency data were taken from Kucera and Francis (1967). For the current study, a total of 100 word samples (50 concrete and 50 abstract) with frequency greater than 20 were selected from Kroll and Merves' list. The more frequent items were selected so that they would more likely be familiar to our human raters. Sense definitions were collected from WordNet 3.0[2]. WordNet started off as a psycholinguistic project for studying human lexical memory but it turns out to be a large lexical database of English widely used in computational linguistics. It

[1] Downloaded from http://nlp.cs.lth.se/lt_home/
[2] http://wordnet.princeton.edu

contains individual databases for nouns, verbs, adjectives and adverbs. Word senses are grouped as sets of synonyms (i.e., synsets). The synsets are hierarchically organised and linked by relational pointers indicating various kinds of lexical relations such as hyponymy, meronymy, etc. in the noun hierarchy (Fellbaum, 1998). Each synset is associated with a simple gloss like conventional dictionary definitions. For our word samples, the concrete words have 1 to 17 senses and the abstract words have 1 to 9 senses, with 4.36 and 3.44 senses on average respectively.

3.2 Procedures

In addition to the average human ratings available from Kroll and Merves (1986), four human judges were asked to rate the selected samples on a 7-point scale (with 7 for highly concrete and 1 for highly abstract). They were asked to first rate each given word as a whole, and then give a score for individual senses of the words.

The WordNet definitions were first made complete sentences by adding the headword as subject such as "car is a motor vehicle …", and then parsed. The parsing results were analysed by a program implementing the scoring method above, giving a score for each definition and a detailed log for the analysis.

The scores for individual senses were compared to the human ratings we obtained for this study. For the overall concreteness of a word, we tried two ways for estimating it from the individual senses. One is to take the average of the scores from all senses (AvgDef), assuming all component senses of a word contribute equally to its overall concreteness; and the other is to take the score for the first sense (FirstDef), assuming the most frequent or familiar sense dominates one's perception of the word in general. These two measures were compared with the original ratings in Kroll and Merves' study and the ratings of the human raters in this study.

3.3 Results and Discussion

We had assumed the validity of the concrete/abstract distinction from psycholinguistic evidence to start with, and aimed to investigate whether such a distinction would be similarly reflected in dictionary definitions through various defining styles. It is thus important to see if the two groups of words will result in significantly different scores as estimated by different ways (e.g. AvgDef and FirstDef) for the word-level concreteness.

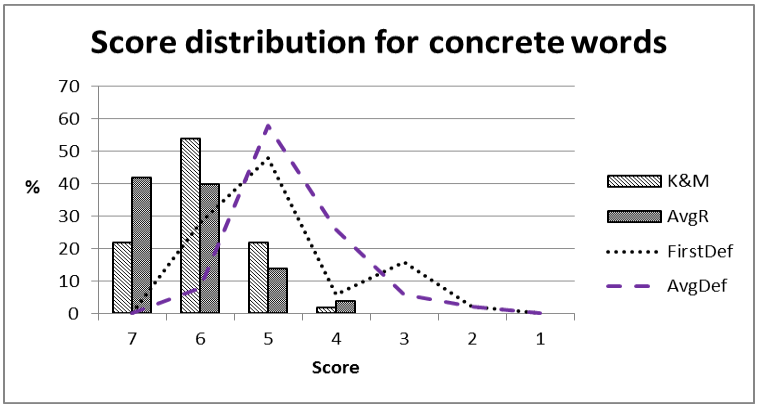
Table 2 shows the mean and standard deviation of the concreteness ratings for the 100 word samples (and for the 50 abstract and 50 concrete samples separately) obtained from human raters and WordNet definitions. The column K&M refers to the original rating from Kroll and Merves (1986). AvgR refers to the average of the four human raters in the current study. AvgDef and FirstDef refer to the scores obtained by the two ways respectively from WordNet definitions.

Table 2: Summary of Concreteness Ratings at Word Level

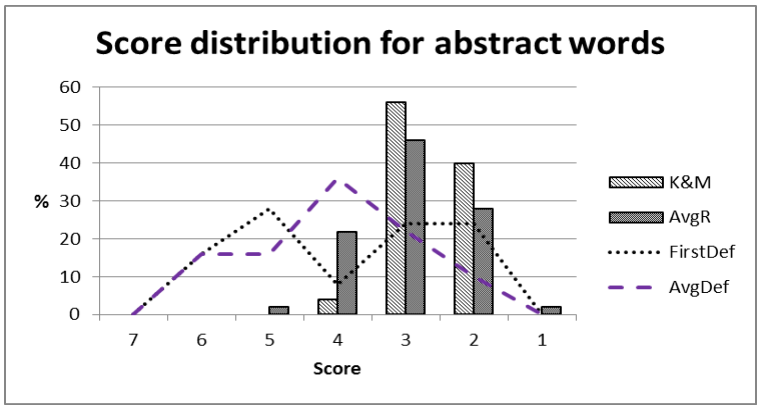
	K&M	**AvgR**	**AvgDef**	**FirstDef**
All Word Samples (N=100)				
Mean	4.27	4.58	4.34	4.36
SD	1.76	1.79	0.92	1.37
Abstract samples (N=50)				
Mean	2.63	2.96	4.11	3.88
SD	0.58	0.73	1.03	1.47
Concrete samples (N=50)				
Mean	5.92	6.19	4.58	4.84
SD	0.63	0.77	0.74	1.08

The distinction between the two groups of words is particularly clear from the human ratings. With K&M, which is our data source, the abstract samples have a mean rating of 2.63 with standard deviation of 0.58, and the concrete samples have a mean rating of 5.92 with standard deviation of 0.63. A similarly apparent distinction is also seen in AvgR, with mean at 2.96 and standard deviation at 0.73 for the abstract words, and mean at 6.19 and standard deviation at 0.77 for concrete words. Interestingly, the human raters in this study seem to be more lenient in their scores than those in K&M, as shown by the higher mean ratings for AvgR in Table 2 and the distribution of ratings shown in the bar charts in Figure 3. The variation might be related to the small number of human raters in this study and the difference between native and non-native speakers. Nevertheless, the distinction between concrete and abstract words is apparent among human raters.

The difference between the mean scores for abstract words and concrete words is much smaller with the estimation from definition scores. The mean is above 3 for abstract words, and below 5 for concrete words, with both AvgDef and FirstDef. The Mann-Whitney rank test shows that the difference between the means for abstract words and concrete words is statistically significant for both cases. The dispersion of scores for AvgDef and FirstDef is shown in the line charts in Figure 3. There is obviously a mode at 5 for concrete words, and the majority did fall on the high side above 4. The scores estimated from WordNet definitions by FirstDef are roughly bi-modal for the abstract words, while those by AvgDef are somehow averaged out giving a mode at 4. This suggests that words deemed abstract also possess considerable concrete senses, and the first senses of about half of the abstract words are not necessarily abstract or are not defined as if they are abstract in WordNet.



(a)



(b)

Figure 3: Score Distribution for (a) Concrete Words and (b) Abstract Words

Table 3 shows the Spearman rank correlation (ρ) and the Kendall's coefficient of concordance (W) between the two sets of human ratings and those between the human ratings and definition scores on all word samples. The former measures the interdependence between two sets of scores and the latter is used for assessing the agreement among raters, or the different rating sources in this study. Values which are statistically significant at least at the 0.05 level are indicated in bold. The human ratings in K&M and those in the current study (AvgR) are quite strongly correlated (ρ=0.848), and the overall ranking is very much similar (W=0.924). The strong agreement among human raters provides evidence that a gradient of concreteness is shared in human cognition.

Comparing with the definition scores, it turns out that the first sense of a word in WordNet is a relatively better indicator of the overall word concreteness than the average of all senses. FirstDef shows better correlation with human ratings (0.301 with K&M and 0.415 with AvgR, both are statistically significant). It suggests that sense frequency might play a very important if not predominant role in one's perception of the concreteness of a word in general, given that the most frequent sense is listed as the first sense in WordNet. However, as seen in Figure 3, FirstDef still has a wider and more even spread of scores across the scale, especially for abstract words. Hence words deemed concrete may more likely have a concrete first sense, while words deemed abstract may also have frequently used concrete senses. Alternatively, many abstract words (and their first senses) might be unexpectedly describable in the lexicographers' minds.

Table 3: Degree of Association between Word Concreteness Ratings from Various Sources

	AvgR	**AvgDef**	**FirstDef**
Spearman Rank Correlation (ρ)			
K&M	**0.848**	0.193	**0.301**
AvgR	--	**0.363**	**0.415**
Kendall's Coefficient of Concordance (W)			
K&M	**0.924**	0.597	**0.650**
AvgR	--	**0.681**	**0.707**

For the sense level ratings, there are 390 senses altogether, 172 are from the abstract word samples, and 218 are from the concrete word samples. Table 4 shows the mean and standard deviation of the concreteness ratings for all senses and separately for senses deemed concrete and abstract respectively by human raters. AvgR is the average rating of the four human raters in this study, and is used as a reference here since no comparable ratings at the sense level are available from previous studies. WN refers to the scores from WordNet definitions.

Table 4: Summary of Concreteness Ratings at Sense Level

	AvgR	**WN**
All Sense Samples (N=390)		
Mean	4.81	4.33
SD	1.31	1.27
Concrete Senses (AvgR≥4, N=270)		
Mean	5.51	4.47
SD	0.89	1.18
Abstract Senses (AvgR<4, N=120)		
Mean	3.24	4.03
SD	0.44	1.49

Similar to word-level concreteness, the mean of human ratings on the concreteness of individual senses is higher than those obtained from definitions. Figure 4 shows the distribution of scores estimated from the definitions, compared with the average human ratings. Apparently our categorisation of the defining styles and surface structures of the definitions has not been able to realistically distinguish concrete concepts on a finer scale (e.g. highly concrete vs mildly concrete) as naturally as human raters.

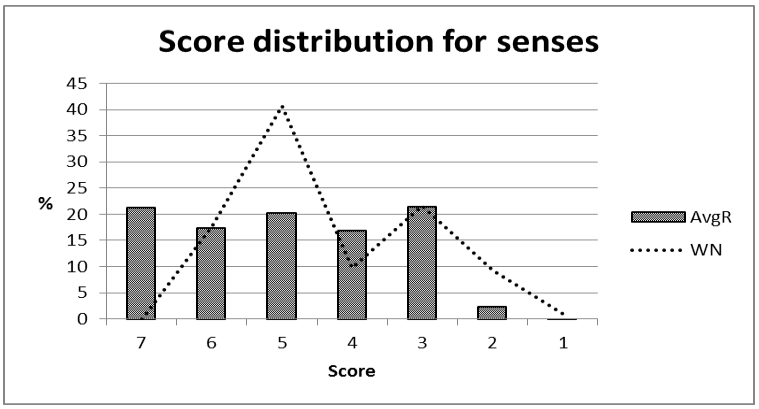


Figure 4: Score Distribution for Senses

Unlike the word level, human ratings for individual senses show less agreement. The variation is particularly apparent for the abstract concepts. For instance, one sense of "devil" is "an evil supernatural being", for which the four raters gave 5, 3, 6, and 2 respectively. The concept is obviously intangible, but individual raters might easily imagine the relevant images as found in fictions and movies. The definition, however, is one of genus and differentiae.

The association between human ratings and definition scores is also weaker at the sense level (ρ=0.257, W=0.628), though still statistically significant. Figure 4 shows a very different distribution of scores obtained from definition structures and human raters. Our categorisation of the definition styles yielded most score 5, which covers only a special kind of genus-and-differentiae definitions. Scores on the two ends of the scale are relatively rare.

Thus our results show that measuring concreteness from dictionary definitions based on the different defining styles apparently works better for concrete senses than abstract ones. Although the two groups of senses can still be told apart in general, the differentiation is not as distinct as is found in human ratings, and definition scores lack the ability to further delineate concepts of different degrees of concreteness on a finer spectrum. There are several factors which might account for this outcome. The range of surface structures we have attended to is quite narrow. On the one hand, the concise conventional defining styles may pose a limit on including further information with differentiae and prototypes. On the other hand, very common concrete concepts may not need so much detailed explanation as others in a dictionary. The assumption that different defining styles are more suitable for concrete concepts and abstract concepts is generally valid. For instance, we do find more definitions involving shunters among the abstract senses. Nevertheless, how a concept is defined may also depend on how describable the concept is, which might just be one dimension amongst others contributing to the concreteness of the concept perceived by humans. As we have observed from the data, apart from parsing errors (e.g. one definition for "dollar" is "a United States coin worth one dollar", but "coin" was tagged as VBP resulting in subsequent parsing error), abstract concepts may also be defined via genus and differentiae, and this might be even more serious with WordNet given its hierarchical organisation of the senses. For instance, the word "concept" is defined as "an abstract or general idea inferred or derived from specific instances" which scored 6 despite the word "abstract" explicitly appears in the definition. However, the

Collins COBUILD Advanced Dictionary[3] defines it as "an idea or abstract principle" which will only score 3 in our system as it is in the form of a synonym. Hence in addition to the dependency structures, there are other surface clues such as lexical choice and length of definition which may suggest on the concreteness of the concept being defined.

Our results do not only echo the subjectivity in human perception of concreteness (and in lexicographers' perception too), but also suggest the multi-dimensionality of concreteness and the importance of finding a systematic measure of this psychologically valid construct so that it can be operationalised and objectively studied in computational linguistics.

4 Conclusion

Hence we have introduced a systematic way of measuring concreteness from how lexicographers understand and explain concepts, by analysing the surface structure of definitions. This is a useful first step to enable further study of a subjective but psychologically valid factor in computational linguistics, especially the construction of semantic lexicons and evaluation for tasks like word sense disambiguation. Current results show that the method works better for concrete senses, and is in many cases able to make a dichotomous distinction between concrete and abstract concepts. However, it is possible that the different defining styles may reveal how describable a concept is, amongst the many factors influencing one's perception of concreteness, and it happens that some abstract senses might be more describable than others. Consequently our method is not adequate for distinguishing concreteness on a finer scale. Our next step is therefore to refine the analysis of definitions in order to deploy a wider range of surface clues, and to investigate how the method works with definitions from other dictionaries.

References

Atkins, B.T.S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press.

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.

Chodorow, M.S., R.J. Byrd and G.E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, Chicago, pp.299-304.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Hanks, P. 1987. Definitions and Explanations. In J.M. Sinclair (Ed.), *Looking Up: An account of the COBUILD Project in lexical computing*. London: HarperCollins Publishers.

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

Johansson, R. and P. Nugues. 2008. Dependency-based Syntactic-Semantic Analysis with PropBank and NomBank. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL 2008)*, Manchester, pp.183–187.

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.

Kucera, H. and W.N. Francis. 1967. *Computational Analysis of Present-day American English*. Providence: Brown University Press.

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.

[3] http://mycobuild.com/homepage.aspx

Palmer, M., H.T. Ng and H.T. Dang. 2006. Evaluation of WSD Systems. In E. Agirre and P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and Applications*. Dordrecht: Springer.

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.

Vossen, P., W. Meijs, and M. den Broeder. 1989. Meaning and structure in dictionary definitions. In B. Boguraev and E.J. Briscoe (Eds.), *Computational Lexicography for Natural Language Processing*. London: Longman.

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.