# Investigation into Human Preference between Common and Unambiguous Lexical Substitutions

**Andrew Walker**
University of Aberdeen
Department of
Computing Science
r05aw0@abdn.ac.uk

**Advaith Siddharthan**
University of Aberdeen
Department of
Computing Science
advaith@abdn.ac.uk

**Andrew Starkey**
University of Aberdeen
School of Engineering
and Physical Sciences
a.starkey@abdn.ac.uk

## Abstract

We present a study that investigates that factors that determine what makes a good lexical substitution. We begin by observing that there is a correlation between the corpus frequency of words and the number of WordNet senses they have, and hypothesise that readers might prefer common, but more ambiguous words over less ambiguous but also less common ones. We identify four properties of a word that determine whether it is a suitable substitution in a given context, and ask volunteers to rank their preferences between two common but ambiguous lexical substitutions, and two uncommon but also unambiguous ones. Preliminary results suggest a slight preference towards the unambiguous.

## 1 Introduction

Paraphrasing is a sub-field of natural language processing (NLP) which aims to modify utterances from one form into another, without changing their meaning. One particular application of paraphrase is text modification to improve information access for low-level readers; e.g., syntactic simplification (Siddharthan, 2006; Siddharthan, 2003), paraphrase (Inui et al., 2003) and lexical simplification (Devlin and Tait, 1998).

Lexical simplification is typically defined as the task of replacing difficult words with simpler ones. However, there are many open question about when one word would be a good substitute for another in context. Our analysis of WordNet 3.0 entries (Miller, 1995) demonstrates an inverse correlation between word frequency rank in the BNC[1] and number of senses it has in WordNet (Pearson = -0.20; $p < 0.001$). In other words, more common (and perhaps simpler) words are also likely to be more ambiguous. This raises an interesting question about whether, given the choice between a common (and perhaps simpler) but ambiguous word and a less common but unambiguous word, readers would prefer one over the other.

## 2 Related work

Hayes (1988) found common patterns of word-usage in various textual genres, indicating that there may be some empirically derivable factors that predict lexical choice in speech and writing. His work focussed on word-frequency statistics, and in that work he highlighted that polysemy issues were important but difficult to analyse due to the limited technology of the time.

The PSET project (Devlin and Tait, 1998; Carroll et al., 1998) looked at simplifying news reports for aphasics and was perhaps the first computational work to focus on lexical simplification (replacing difficult words with easier ones). The PSET project used *WordNet* (Miller, 1995) to identify synonyms and the Oxford Psycholinguistic Database (Quinlan, 1992) to determine the relative difficulty of words (Devlin and Tait, 1998). Elsewhere, there has been interest in *paraphrasing*, including the replacement of difficult words (especially verbs) with their dictionary definitions (Kaji et al., 2002).

The tradeoff between brevity (and perhaps fluency) and clarity (or ambiguity) was studied by Khan et al. (2008) in the context of generating refer-

[1]The British National Corpus, version 3, 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk



*Proceedings of the 13th European Workshop on Natural Language Generation (ENLG)*, pages 176–180,
Nancy, France, September 2011. 

ring expressions with the specific form "Adj Noun and Noun" (e.g., *old men and women*) where the scope of the adjective is ambiguous. They found that hearers prefer to read clear phrases over brief ones. Our study is similar in spirit, and we ask whether hearers prefer clarity to simplicity.

The 2007 SemEval lexical substitution task (McCarthy and Navigli, 2009) created a small corpus of manually selected lexical substitutions for 30 words in 10 contexts each. Participating systems had to submit lists of acceptable substitutions in these 300 contexts and were evaluated on recall and precision relative to the manually compiled gold standard. We reuse data from this corpus, focussing on the question of which of the valid substitutions would be preferred by readers.

## 3 Methodology

This paper has two parts. First, we briefly investigate factors that make a word substitution valid in context and present a machine learning approach to deciding the validity of word substitutions (§3.1). Then, in the second part (§4), we study whether readers prefer simpler but more ambiguous words. We use data from the 2007 SemEval lexical substitution task both parts.

### 3.1 Lexical substitution

In order to investigate the tradeoff between ambiguity and commonness, we need an algorithm to:

1. Discover possible lexical replacements, and
2. Rank the suitability of these replacements according to parameters such as ambiguity and commonness.

Our interest is really in the second step, but we need to identify valid replacements before we begin to rank them. For this purpose, we restricted ourselves to WordNet 3.0 (Miller, 1995) as a source of substitutions. The first step involved extraction of the "synsets" (synonym sets) that contain the word being replaced and then listing all of the elements in those synsets to find synonyms. For verbs and nouns, we also include any synsets bearing a "hypernym" relation to one of the originals; and similarly for adjective synsets via a "similar to" relation.

For this paper, we focus on the second step. This involved determining and weighting various properties of the words deemed as possible replacements. We identified the following properties:

1. **context**: a distributional measure of the likelyhood of each word in the context of the sentence;
2. **recognisability**: an estimation of how likely the word is to be recognised; i.e., whether the word is in the reader's vocabulary;
3. **suitability**: an estimate of whether the word is a suitable replacement, given the sense of the original word; and
4. **ambiguity**: how polysemous the word is.

In this way, words that are very common in the context should be more likely to be chosen, but might still be ranked lower than another less common word that is also less ambiguous. There should be a strong preference in the system output for any options that are both common and unambiguous.

#### 3.1.1 Context

For the context, we produce a unit vector of the words surrounding the target item (maximum of 5 either side) weighted in proportion to their distance from it. To use an example from the task:

"We cannot **stand** as helpless spectators while millions die for want in a world of plenty"

would be encoded as:

$$\begin{pmatrix} cannot, & 0.208\bar{3} \\ as, & 0.208\bar{3} \\ we, & 0.166\bar{6} \\ helpless, & 0.166\bar{6} \\ spectators, & 0.1250 \\ while, & 0.083\bar{3} \\ millions, & 0.041\bar{6} \end{pmatrix}$$

An entry in the corpus matching one of the substitutions (e.g., "remain") will have its surrounding vector similarly derived. The dot-product of the two is then calculated. The context score for that substitution option ("remain" here) is the sum of all such vector dot-products for entries in the corpus.

#### 3.1.2 Recognisability

The recognisability score is an estimation of how likely a word is to be in a reader's lexicon. We observed that the form of a graph plotting word frequency against word rank does not appear to be plausible as a model of an individual's likely vocabulary. The Zipfian distribution of language would make such a simplistic model predict that the second most common word would only have a 50% chance of being recognised. We predict that a large number of the most common words are almost guaranteed to be recognised, and then a long-tail of the less frequently used words with diminishing recognisability. We model this with the logistic regression function $\frac{1}{1+e^{-z}}$ with $z = 6 - \frac{rank}{10000}$.

This model is so far unjustified though. It predicts a vocabulary of 60,000 words, as per Aitchinson (1994), following a logistic regression curve plateauing with the most common 30,500 words returning recognisabilities greater than 0.95 and then describing a long-tail of words with reducing recognisabilities.

#### 3.1.3 Suitability

As there is no word-sense disambiguity process involved, all we can be sure of is that one of the original word's senses was the intended sense. The suitability score is calculated as the portion of the original word's senses that the substitution shares. Thus the suitability of a substitution ($subs$) given the original ($orig$) is

$$\frac{|senses(subs) \cap senses(orig)|}{|senses(orig)|}$$

#### 3.1.4 Ambiguity

The ambiguity score is simply the inverse of the number of senses held by the substitution word:

$$\frac{1}{|senses(subs)|}$$

### 3.2 Lexical substitution task results

The 2007 SemEval lexical substitution task corpus consists of 30 selected words appearing in ten sentences each, giving 300 sentences in total. For each of these 300 sentences, there is a manually compiled list of valid lexical substitutions for the selected word. The challenge is to computationally derive suitable alternatives for the selected word in each of the 300 sentences. Results were scored for precision and recall relative to the manually compiled gold standard.

The SemEval 2007 task authors described a baseline for WordNet systems that achieved a precision of 0.30 and recall of 0.29. Our implementation (that multiplies the values of the four features defined above) scores a precision of 0.35 and a recall of 0.35. But, it should still be noted that solutions designed at the time used a much richer set of sources for replacements, including automatically constructed paraphrase corpora, and subsequently scored much better, with the best system achieving precision and recall of 0.72.

### 3.3 Learning a model to fit the data

The solution described above assumes (without justification) an equal weighting for each attribute. We also trained a machine learner to classify replacements as valid or invalid based on these four features. To create labelled data, we collated all of the possible replacements as found by our method described in §3.1. We then labelled the replacement word as "valid" if it was one of those found in the manually compiled gold standard for the task, and "invalid" otherwise.

A number of modifications were made to the attributes in order to make them more suitable for the machine-learning process. It was found that the context score had an extremely long tail, and taking the logs of each context score gave a much more reasonable distribution. The polysemy scores were, by their inverse-integer nature, skewed towards $0.0$ with large gaps between each fractional value (e.g. no score could possibly be in the range $(0.5, 1.0)$). For this reason, we instead just used the number of senses the word could be used in, directly, rather than taking the inverse. The overlap scores were modified to be the raw number of senses shared (or the cardinality of the intersection of the two words' sets of senses), demonstrating that in the vast majority of cases only a single sense was shared, suggesting it might not be a very useful metric.

This data was then split into ten parts, each with the results and scores for three words. (Each section therefore did not have the same number of entries.) We tested each set on an IBk classifier (Aha et al.,

1991) trained on the other nine. After extracting the predicted "valid" results we scored them as we described in §3.2 with precision and recall of 0.291. The poor performance of machine learning is possibly due to the low number of words available for training.

## 4 Study on reader preference

We presented human volunteers with 21 sentences. Each sentence had a word singled out and four possible substitutions for it. These four substitutions are the most common and the least ambiguous words from the manually compiled list of valid substitutions in the 2007 SemEval lexical substitution corpus, and the most common and least ambiguous words from the list of words suggested by our algorithm (§3.1). The full matrix is presented in Table 1.

The 21 sentences used in this study were selected as follows:

1. The manually compiled gold standard contained at least two substitutions

2. The classifer predicted at least two different substitutions to the gold standard

Thus our data for the study comprises just the sentences for which there are four distinct lexical substitutions available, two each from the gold standard and the classifier. Our method for selecting data for this study filters out sentences for which the system recommendations overlap with the gold standard. Thus it is of interest to see whether these system recommendations are liked by readers.

Ten human volunteers, recruited by word-of-mouth, were presented with each original sentence in a random order, and offered the four possible replacements, again randomly ordered. They were asked to rank all four in order of preference as a substitution for the original word in context.

| | Manual | System |
|---|---|---|
| Most Common | 21 | 21 |
| Least Ambiguous | 21 | 21 |

Table 1: Matrix of word option types

**Context:** There are sound reasons for concluding that the long-run picture remains **bright**, and even recent signals about the current course of the economy have turned from unremittingly negative through the late fall of last year to a far more mixed set of signals recently.

| Judge | Options | | | |
|---|---|---|---|---|
| ID | good | brilliant | gleaming | hopeful |
| 1 | 2 | 4 | 3 | 1 |
| 2 | 3 | 4 | 1 | 2 |
| ... | | | ... | |
| 9 | 2 | 3 | 4 | 1 |
| 10 | 2 | 3 | 4 | 1 |
| Totals: | 21 | 34 | 34 | 11 |

Table 2: Example of result tabulation for lexical substitutions of the word "bright" in context.

| | Pearson | p-value |
|---|---|---|
| frequency | 0.087 | 0.216 |
| log(frequency) | -0.164 | 0.073 |
| polysemy | -0.196 | 0.037 |

Table 3: One-tailed correlations and p-values between average rankings and the listed word properties

### 4.1 Results

For each sentence we added up the ranks from all volunteers for each of the four replacements to get a final score. In Table 2, for example, "hopeful" was ranked as the most preferred replacement, with "good" following, and "brilliant" equalling "gleaming" as the least preferred.

These were analysed against each option's frequency (in the BNC) and its level of polysemy (in WordNet). The correlations are listed in Table 3. Table 3 shows a significant inverse correlation ($p < 0.05$) between the preferred words and their level of polysemy; i.e., readers prefer less ambiguous words. We did not find a significant correlation ($p > 0.05$) between word preference and corpus frequency.

Recalling the matrix in Table 1, we are intereste in the effect of two factors with two conditions each:

1. **source:** manual or system generated

2. **criterion:** most common or least ambiguous

The replacements from the manual gold standard were ranked significantly higher ($p < 0.05$) than the replacements from the system output. However,

there were 7 out of 21 cases where a system suggestion was ranked the highest. This is interesting because we specifically selected sentences where the system recommended words that were not in the gold standard; these novel recommendations were preferred in a third of cases.

We did not find any effect of the criterion factor on preference. Indeed, there were 11 cases where an unambiguous word was preferred and 10 where a common word was. We suspect that this is because the words in the manual gold standard tended to be fairly common ones; the SemEval annotators did not have access to a thesaurus or lexical database when suggesting substitutions. Thus, our ranking of words by frequency was not very informative.

## 5 Conclusions

Our primary result was the significant inverse correlation between the word preference and level of polysemy; i.e., our participants showed a preference for less ambiguous words. We found no correlation between word frequency and preference. We might suppose that the critical matter is simply if a word is familiar or not, and so a more more common familiar word has little or no benefit to a reader over a slightly less common, but still familiar one.

The classifier performed well at predicting which words would be suitable, in line with expectations, and more investigation may be warranted to see if other attributes of words could factor into such a task. We suppose that there might be distinctions between the different parts-of-speech, or that more details about the word being replaced would also aid the classification process.

## References


David W. Aha, Dennis Kibler, and Marc K. Albert. 1991. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66.

Jean Aitchinson. 1994. *Words in the mind: An introduction to the mental lexicon*. Blackwell, Oxford, England.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of AAAI98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10, Madison, Wisconsin.

Siobhan Devlin and John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. In J. Nerbonne, editor, *Linguistic Databases*, pages 161–173. CSLI Publications, Stanford, California.

Donald P. Hayes. 1988. Speaking and writing: Distinct patterns of word choice. *Journal of Memory and Language*, 27(5):572 – 585.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing - Volume 16*, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nobuhiro Kaji, Daisuke Kawahara, Sadao Kurohash, and Satoshi Sato. 2002. Verb paraphrase based on case frame alignment. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 215–222, Philadelphia, USA.

Imtiaz Hussain Khan, Kees van Deemter, and Graeme Ritchie. 2008. Generation of referring expressions: managing structural ambiguities. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 433–440.

Diana McCarthy and Roberto Navigli. 2009. The english lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, November.

Philip Quinlan. 1992. *The Oxford Psycholinguistic Database*. Oxford University Press, U.K.

Advaith Siddharthan. 2003. Preserving discourse structure when simplifying text. In *Proceedings of the European Natural Language Generation Workshop (ENLG), 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, pages 103–110, Budapest, Hungary.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.