# Dictionary Definitions vs. Example Sentences: Learning, Use and Recall of Unknown Adjectives

James Ronald

(Received on May 10, 2001)

This paper, focusing on adjectives, follows up on research on learning verbs from dictionary definitions (JR98a). The basic reasons justifying such a follow-up experiment is to investigate whether learning and retention is observably different for the two parts of speech. The use of adjectives will also affect a number of other related aspects of the research, such as the nature of dictionary definitions for different parts of speech, the choice of example sentences, and the issues raised for the raters of the subjects' responses. These will be considered below.

The follow-up study was felt to be of value mainly because of two features apparent in much vocabulary acquisition research:

i) That a "balanced" set of target words composed, for example, of nouns, verbs, adjectives and adverbs, is often used in vocabulary acquisition research.

ii) That opinion remains divided as to whether vocabulary learning and retention are different for different parts of speech.

One implication of i) is that if all the target words in an experiment assessing vocabulary acquisition were verbs or nouns or adjectives, the sample would be unbalanced. This in turn implies either a belief that learning and retention are different for different parts of speech or, as long as there is no evidence either way, that it is wise to play safe.

As far as ii) is concerned, research where target words were subdivided

into sets of parts of speech, the results for each set is often not discussed, and is often not worthy of discussion if each set only consists of four or five words. For others, the results vary. Rodgers (1969) has found that learning a word is affected by the part of speech; that nouns are easiest, followed by adjectives, then verbs and adverbs. Paribakht and Wesche (1997), for example, in one experimental condition, record subjects making significant gains for nouns but not for other parts of speech.

From a psycholinguistic perspective, concrete nouns tend to be learned earlier, remembered better, and translated faster than abstract nouns (De Groot, 1993). In a similar way, verbs or adjectives may be typically more abstract than, for example, nouns and so less easily learned, remembered, or translated. Different words also, as Paradis (1997:336) observes, lend themselves more or less easily to mental representation. This may be true not only for individual words but also, typically, for different parts of speech. The different parts of speech may also be more or less dependent upon context for meaning – more or less "stand-alone" items. In research such as that described in this paper, where context is one variable, this may well have some bearing on results.

Another possible source of differences in difficulty for different parts of speech has been noted by Laufer (1997). The typical morphological complexity of different parts may be a factor in determining whether, say, verbs are harder to learn than adjectives. If we compare the lemma for one word from each category, the difference is clear:

Verb: *break, breaks, breaking, broke, broken*

Adjective: *large*

So although, as Laufer (1997) points out, it might be misleading to say that verbs are more difficult than adjectives in English, the intrinsic element of morphological complexity may well lead us to the same conclusion.

MacWhinney, too (1997: 120–121), notes the extra load in learning verbs, at least for production, in some languages such as Hungarian.

Apart from the intrinsic hardness of the part of speech of target words themselves, differences in the difficulty of learning and using words may also depend on how they appear in the learning materials used. In the study reported below, for example, the two types materials used are contextless example sentences drawn from a corpus and dictionary definitions from monolingual English learners dictionaries. The ease of guessing meaning from context, either the context of a sentence or the context of a definition in the target language, may also vary according to the part of speech.

Finally, where subject responses are rated for correctness or naturalness, the ease with which responses can be judged correct or natural, and the level of interrater reliability, may also be affected by the part of speech.

For reasons of intrinsic learnability, the extent to which learning materials help, and the ease and accuracy of rating subject responses, it was judged worthwhile to conduct a follow-up experiment focussing on adjectives. However, since these various factors involved in the learning of different parts of speech are interdependent, it is difficult to formulate any simple hypotheses as to how the learning and retention of adjectives might be different from that for verbs. However, it is proposed that there will be differences between the two parts of speech, both in the rate at which they are learned and in the task faced by the raters of the subject responses.

## Method

### *Design*

Following a pretest to ensure that target words were unknown, Japanese intermediate learners of English were given a set of unknown adjectives. Together with the target words, one group (the Dictionary Group) were

given a set of monolingual dictionary definitions for the target words (two definitions per word). The other group (the Example Sentences Group) were given a set of authentic example sentences (three sentences per word). The subjects were instructed, in their own language, to study the definitions or example sentences and to write their own sentence for each of the target words, using the information in the materials provided. As they finished the task, they were instructed to write Japanese equivalents for each of the target words. Three weeks later, the subjects were given a test of vocabulary retention, a kind of gap-fill exercise in which they had to match the sentences or definitions with the correct word.

*Subjects*

Initially, 32 Japanese university students were recruited as subjects for this experiment. For the retention test, eight of the original subjects were not present, leaving a total of 24 subjects. They were all second year university students, aged between 19 and 21, majoring in English at a private university. They had all received about seven years of formal instruction in English. Despite this, their TOEFL scores would range between about 440 and 520.

*Target words*

There were 20 target words, all adjectives, and all unknown to the subjects. The words were selected according to the following criteria:

1. To exclude words that were either too rare or too likely to be known, only words in the Cobuild 'two diamond' band (Sinclair et al., 1995) were included; words in the 3,400 – 6,600 band.
2. Basically, any words with more than one sense identified in the dictionaries used were excluded.
3. Words for which there were at least 30 occurrences in the *Cobuild*

*Direct* 50 million word corpus.

4. A list of 85 words conforming to 1. and 2. was presented to the subjects as a pretest. Adjectives correctly identified or frequently mistaken for other words were removed from the list.
5. The 20 target words were chosen from the remaining 54 words. The aim of selection at this final stage was to have a variety of adjectives: adjectives only followed by nouns (*illicit*), adjectives only preceded by link-verbs (*afoot*), adjectives for which either of these main patterns is available (*colossal, defunct*), adjectives followed by a preposition (*averse, akin*), and others (*galore*).

The number of target words was set at twenty to allow enough time in a 90-minute class period for subjects to complete the tasks as required. The final 20 target words are as follows:

*afoot, akin, averse, bereft, blatant, callous, colossal, defunct, eerie, fleeting, furtive, galore, gaudy, illicit, inviolate, lenient, morbid, obese, poignant, quaint*

*Learning materials*

As described above, the experiment required the subjects to be divided into two groups and use one or other of two different resources for the target words. The groups were divided randomly into the Dictionary Group and the Example Sentences Group.

Dictionary Group: This group received two dictionary entries for each of the target words, taken from the Longman Dictionary of Contemporary English 3rd Edition (1995) and the Collins COBUILD English Dictionary 2nd Edition (1995). The order in which the definitions in this group's materials appeared alternated between the two dictionaries. The dictionary entries were stripped of any example sentences but included the definitions and grammatical information.

Example Sentences Group: This group received three example sentences

for each of the target words, drawn from the 50 million word COBUILD Direct corpus. Sentences were chosen to display typical syntactic patterns and collocations, and for their comprehensibility. Wherever possible, sentences were taken directly from the corpus without changing them in any way. In a few cases, however, parts of sentences were deleted if they were too long but otherwise ideal. Where more than one grammatical pattern is frequent, the choice of examples reflect this.

*Vocabulary Retention Test*

Three weeks after the main vocabulary learning session in which the above materials were used, a test of vocabulary retention was conducted. For this test, subjects were given a multiple choice answer sheet, together with the same materials as three weeks before, except that the target word was deleted wherever it occurred in the definitions or example sentences and the test items were randomly reordered. This is an example of one item in the retention test, in which subjects had to identify the word which matched the example sentences or definitions:

| A. haphazard | quaint | defunct | idyllic |
|---|---|---|---|
| *Example Sentences*: | | *Definitions*: | |
| All the shops are closed due to a _____ Roman tradition. | | Unusual and attractive, especially in an old-fashioned way. | |
| I am aware of a number of _____ pastimes that are perfomed in rural parts of Britain. | | Something that is _____ is attractive because it is unusual and rather old-fashioned. | |
| Fingleton, in one of his many books, made it clear how _____ he found all this stuffiness. | | | |

## Rating

Two native speaker teachers of English rated the English sentences produced by the subjects and two highly proficient Japanese teachers of English rated the subjects' translation equivalents for the target words. For the rating of the sentences, a set of guidelines from a previous experiment was used to determine what would count as an acceptable error or an unacceptable answer. A set of concordance lines for each target word was available as a resource to help raters decide when it was difficult to decide. For the raters of the translation equivalents, two monolingual English dictionary entries were provided for each target word; dictionaries other than those used in the experiment: Oxford Advanced Learners Dictionary, 5th Edition (Crowther et al., 1995) and Cambridge International Dictionary of English (Procter et al., 1995). The raters also used a widely respected English-Japanese dictionary (Kenkyusha English-Japanese Dictionary for the General Reader, 3rd edition, (1999)).

The raters of the English sentences judged each sentence as Acceptable, Unacceptable, or Questionable. Interrater reliability for the sentences including non-answered items was 83%. However, when non-answered items are excluded, interrater reliability falls to 77.5%. All differences were resolved at a joint meeting, at which the Questionable category was discarded: items in this category being rerated as either Acceptable or Unacceptable.

The raters of the Japanese translation equivalents rated each translation as Correct, Partially Correct, or Incorrect. One example of how the equivalents were judged is for the word *obese*. Translations judged correct include those meaning 'too fat' or 'very fat', while meanings approximating to 'fat' were judged partially correct. Partial superordinates, such as ones

meaning 'big' or 'unhealthy', would be judged incorrect. Interrater reliability overall was 78.5%; again, when non-answered items are excluded, interrater reliability falls to 71.2%. As for sentence rating, differences were resolved at a joint meeting of the two raters.

## Results

There were three sets of results: for the subjects' English sentences, for their Japanese translation equivalents, and for the retention test. These will be presented one by one then discussed both individually and in relation to each other. Following this, there will be a comparison of results from this experiment for adjectives with those on verbs described in JR98a.

*English sentences*

In this analysis omissions and questionable or unacceptable English uses were collapsed into one category – not correct sentences. Acceptable uses were correct sentences. There was a large difference between the two groups for their production of correct English sentences. A T-test was conducted on the ratings of the English sentences per word for the subjects in each of the two groups to confirm that there was a significant difference between the two groups. The results are shown in Table 1.

Table 1. Analysis of results for English sentences.

| | Definitions Group | Example Sentence Group |
|---|---|---|
| Av. Correct (out of 20) | 10.2 | 6.6 |
| S.D. | 2.5 | 3.1 |
| Probability | .001 | |

The Japanese translation equivalents were rated as matching the English word, partially matching, or not matching at all. (No answer was included in

this last category but will be discussed further below.) Weighting of the ratings was done in two different ways. For the 'collapsed categories', matching and partially matching equivalents were collapsed into one category as 'acceptable answers', while with the 'differential weighting', matching equivalents were given two points and partially matching equivalents one point. As for the English sentences, there was a large difference between the two groups for their production of acceptable translation equivalents, regardless of the weighting used. Again, T-tests were conducted on the ratings of the Japanese translation equivalents for both types of weighting in each of the two groups to confirm that these difference were significant. The results are shown in Table 2.

Table 2. Analysis of results for Japanese translation equivalents.

| | Collapsed categories | | Differential weighting | |
|---|---|---|---|---|
| | Defs Gp | Ex Grp | Defs Gp | Ex Grp |
| Av. Score | 14.3 | 4.0 | 22.3 | 5.7 |
| S.D. | 3.4 | 2.1 | 5.9 | 3.1 |
| Probability | .0001 | | .0001 | |

Finally, the results for the Retention Test are shown below. In this case there was no rating of answers; answers were either right or wrong. The results, shown in Table 3, show that was between the two groups' scores on the retention test were almost identical, and a T-test confirmed that there was no significant difference between the two groups.

Table 3. Analysis of Retention Test scores.

| | Definitions Group | Example Sentence Group |
|---|---|---|
| Av. Correct | 9.2 | 9.3 |
| S.D. | 3.9 | 4.3 |
| Probability | .922 | |

## Discussion

In this section, I will begin by discussing the results of the three tasks – production of English sentences, of translation equivalents, and retention scores – and then compare the three sets of results for the two groups. This will be followed by a discussion of the results of the experiment with adjectives described above with the results of the experiment conducted using verbs.

*English Sentences*

Two aspects of the results for the subjects' English sentences stand out; the large difference between the results for the two groups and the size of the Standard Deviation for the Example Sentences group. While the former may simply be accounted for by concluding that the example sentences are less informative or less accessible than the dictionary definitions, the question of the Standard Deviation needs more consideration. The first thing a large Standard Deviation may suggest is that subjects' L2 proficiency levels vary considerably. The lower Standard Deviation for the Definitions group, however, suggests that this may not be the cause. Or, rather, it may be that the task with the example sentences is a more sensitive indicator of L2 proficiency than that with the dictionary definitions. This in turn leads us to ask what, in this context, proficiency may mean. For the successful completion of the English sentences task, it appears that at least three areas of L2 knowledge or skills are required; a vocabulary wide enough to understand the English example sentences, an ability to pick up clues from the example sentences as to meaning and usage, and an ability to write coherent English sentences. We will consider this question further below, when we compare results for the different tasks.

*Japanese translation equivalents*

The results of the Japanese translation equivalents do not seem to show any more than that the Definitions group was much more successful than the Example Sentences group. The use of differential weighting appears to magnify the difference between the two groups. Despite the attraction of the differential rating, however, as there are problems of interpreting the scores for this, the collapsed scores are more reliable.

*Retention*

In a desire to use a sensitive measure of vocabulary retention, the retention test that is used does not in fact test retention of an area of word knowledge that was previously tested: it tests sensitivity to the contexts in which the target word was presented. (An alternative might be to use the original pre-test and ask for subjects to rate their own knowledge of the target words.) The remarkably high scores, especially for the Example Sentence group, does however confirm the test's sensitivity. We will consider below why the results of this test are so much at variance to the results of the English Sentences and Japanese Equivalents tasks.

*Comparison of task scores*

The results for the three tasks – the English sentence task, the Japanese translation equivalent task, and the retention test – are shown in the Figure 1. First, if we compare the results of the English sentence task with the Japanese translation equivalents, we can see that for the Dictionary Definitions group scores for the translations are higher than for the English sentences. This is not surprising, given that dictionary definitions focus on conveying meaning and do not (other than, to some degree, those of Cobuild) show how the word is used. For the Example Sentences group,

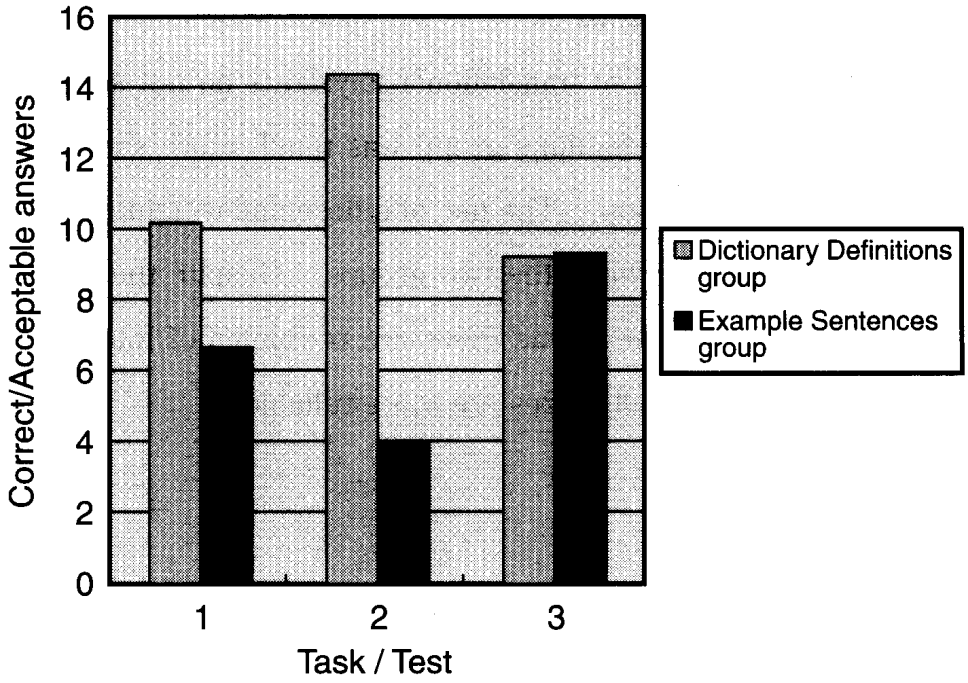


Fig. 1. Summary of combined results for three tasks (1 = English sentences, 2 = Japanese translation equivalents, 3 = retention test).

the position is reversed; their average scores for the English sentences task are higher than those for the Japanese translation equivalent task. Again, this is perhaps what we might expect; the example sentences show how the target words are typically used while any clues they provide as to meaning are, in a sense, incidental. Given this, that this group's scores for English sentences were lower than the Dictionary Definition group's may be due to the difficulty of writing sentences around particular words when understanding of those words is unclear or mistaken.

When we compare the results of the retention test with the scores for the English sentences and Japanese translation equivalents, the retention test results are outstanding in two respects. One is that the retention test scores for the two groups are almost identical, when for the other two tasks they have been so different. The other is that for the Example Sentences group retention test scores are higher than those for the other two tasks. In both respects, however, we need to bear in mind that the reten-

tion test is not testing recall of the word knowledge measured in the two tasks three weeks earlier but recall of the environment in which the target words had been presented at that time. Even so, we need to consider why the two groups' scores for the recall test are so similar when their scores for translating the target words or using them in sentences are so different. Read (2000: 101) notes that multiple-choice cloze tests, of which the recall test is a variant, are more 'learner-friendly' than standard format cloze tests. In other words, they ask less of the test taker. Related to this, it could be argued that the vocabulary pretest, in which subjects were asked to give meanings for words they recognized, was harder than the recall test. A consequence of this may be that the recall test was not only testing recall from the two tasks but also measuring word knowledge that was not identified through the pretest. A third interpretation is that the similar recall test scores for the two groups are not directly related to their scores for the two tasks three weeks previously but are, rather, a reflection for the two groups of the depth of processing (Craik and Tulving, 1975) involved in completing the tasks.

Having considered the combined results of the adjectives experiment, we will now compare and consider the results of this experiment and the experiment with verbs.

## Verbs and adjectives

Finally, we will compare the present experiment with that described in (JR98a). We will start with a comparison of the results of the two experiments. Following this, we will consider three related issues; the selection of example sentences for the two experiments, the problems faced by raters of the sentences for the adjectives, and the proportions of 'no answers' for the two experiments.

*Results*

The results are shown below. For the English sentences task, while the Example Sentences groups for both experiments had lower scores than their respective Dictionary Definition groups, the difference between the two groups in the present experiment was much greater than that for verbs. We will consider possible causes for this below. For the translations, the results were fairly similar for both experiments, with the Example Sentences groups both having significantly lower scores than their respective Dictionary Definition groups. The results for the retention tests are all very similar, the only remarkable feature being the unexpectedly high score for the Example Sentences group in the adjectives experiment. Given the similarity between all the results for this test, it appears likely either that the results are, as suggested above, a reflection of the depth of processing involved in completing the previous tasks (Craik and Tulving, 1975) or that the test is, in a sense, too sensitive; revealing previous word knowledge that was not detected by the pretest.

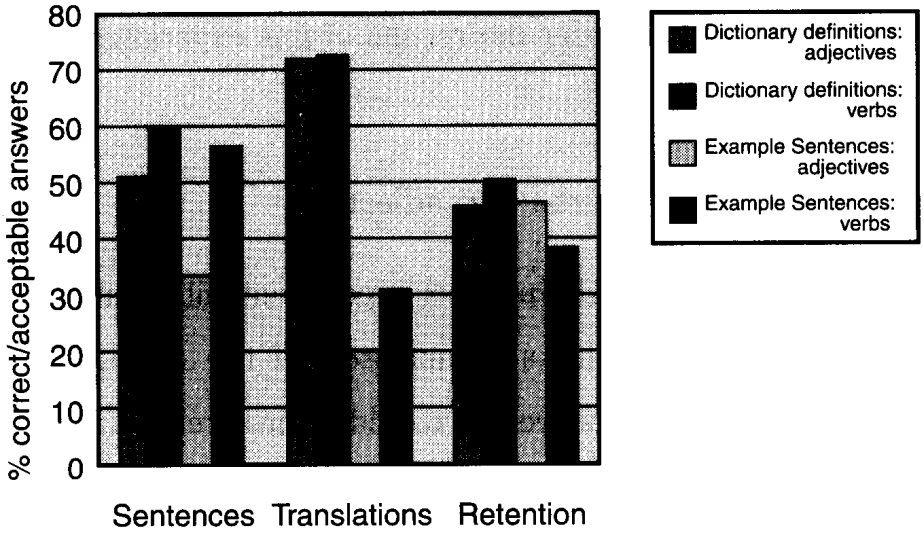


Fig. 2. Conparison of results for adjectives and verbs.

*Conclusion and recommendations*

It was initially assumed that the replication of the experiment with verbs (JR98a) using adjectives instead would be a relatively straightforward matter, and with one or two exceptions the graph of results, above, appears to support these assumptions. In terms of selecting example sentences, administering the experiment and rating the results, however, various unexpected problems arose which may have affected the results. For selecting the example sentences, the frequencies of the target words for the two experiments were similar, and so was the 'pool' of 30 to about 200 sentences from which they were selected. Despite this, it was much harder to find typical yet comprehensible example sentences. A sample test with native speakers showed that less than a quarter of the target words could be guessed exactly from the gapped example sentences, while for almost half the guesses were either completely wrong or not made at all. In other words, the L2 subjects were being asked to perform tasks which native speakers could not perform. This difficulty accounts for the poorer results of the Example Sentences group for the adjectives experiment.

Another factor that may have affected results in administering the adjectives experiment was there was no absolute insistence that example sentences should be written for all target words. This may have been because the example sentences were more difficult than for verbs or because it seemed pointless to insist on this when the words' meanings could not be guessed at.

Finally, the rating of the English sentences for the adjectives experiment was especially difficult when compared with that for verbs. Although the same two raters as before did the rating, there was a lower level of interrater reliability. One major reason for this is that while in English there is rarely any difficulty in identifying a word as an adjective, in Japa-

nese, which has no articles, it is often confusing; the difference between a N N pattern and an ADJ N pattern often seemed to be largely a matter of opinion. For the raters, one of the criteria for acceptability was that the Japanese translation equivalent should be the same part of speech; where this was hard to determine, there was confusion.

In conclusion, in comparing the results and circumstances of the two experiments with verbs and with adjectives, it appears that dictionary definitions remain a more reliable and accurate source of information for determining meaning of unknown words in a foreign language. How useful example sentences are appears to depend to some degree on the part of speech. Finally, we will return to whether different parts of speech are different as far as the learning of words is concerned. On the basis of the research described above, it would appear that where context is a major factor, and especially where authentic examples are used, Rodgers' (1969) hierarchy of part of speech difficulty (nouns – adjectives – verbs and adverbs) does not seem to apply.

## References

Craik, F. I. M and Tulving, E. 1975 Depth of processing and the retention of words in episodic memory. *Journal of experimental psychology* 104: 268–284.

de Groot, A. M. B. 1993 Word type effects in bilingual processing tasks. In R. Schreuder and B. Weltens (Eds.) *The bilingual lexicon.* Amsterdam: John Benjamins.

Laufer, B. 1997 What's in a word that makes it hard or easy: some intralexical factors that affect the learning of words. In N. Schmitt and M. McCarthy (Eds.) *Vocabulary: description, acquisition and pedagogy.* Cambridge: Cambridge University Press.

MacWhinney, B. 1997 Second language acquisition and the competition model. In A.M.B. de Groot and J.F. Kroll (Eds.) *Tutorials in bilingualism: psycholinguistic perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Nation, I. S. P 1990 *Teaching and learning vocabulary.* Boston, MA: Heinle and Heinle.

Paradis, M. 1997 The cognitive neuropsychology of bilingualism. In A.M.B. de Groot and J. F. Kroll (Eds.) *Tutorials in bilingualism: psycholinguistic perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Paribakht, T. S. and Wesche, M. 1997 Vocabulary enhancement activities and reading for meaning in second language acquistion. In J. Coady and T. Huckin (Eds.) *Second language vocabulary acquisition.* Cambridge: Cambridge University Press.

Read, J. 2000 *Assessing vocabulary.* Cambridge: Cambridge University Press.

Rodgers. T. S. 1969 On measuring vocabulary difficulty: An analysis of item variables in learning Russian-English vocabulary pairs. *IRAL* 7:327–343.