A Corpus-based Analysis of the Semantic Relatedness between the American English *-ic/-ical* Adjective Pairs[1)]

Myoung Hyoun Song
(Seoul National University)

Song, Myoung Hyoun. 2008. A Corpus-based Analysis of the Semantic Relatedness between the American English *-ic/-ical* Adjective Pairs. *SNU Working Papers in English Linguistics and Language 7*, 107-120. The English adjective pairs ending in *-ic* versus *-ical* share the stem, implying the possibility of semantic relatedness as well as morphological relatedness. The corpus-based analysis of the bigrams (the *-ic/-ical* adjective with their R1 collocates) shows that some adjectives (*analytic/analytical, classic/ classical, historic/historical, magic/magical*, etc) are semantically related to the degree that they can be interchangeably used. But others (*economic/ economical, politic/political*, etc.) are not semantically related to the degree that they can not be used in place of each other. The semantic relatedness can be extended to the difference in meaning, when the bigrams are taken into account, excluding the shared R1 collocates. **(Seoul National University)**

Keywords: bigrams, corpus-based, semantic relatedness, ESCO

1. Introduction

In English, there are paired adjectives sharing the same base, ending in *-ic* and *-ical* respectively, as shown in (1) below:

(1) *analytic - analytical ; classic - classical ; comic - comical ; economic - economical ; electric - electrical ; geometric - geometrical ; graphic -graphical ; historic - historical ; logistic - logistical ; magic -magical ; numeric - numerical ; politic - political ; problematic - problematical ; egotistic - egotistical*

At first sight, these pairs might be used interchangeably in that they

1) This paper is based on Gries 2001, taking its key concepts and methodology to investigate the semantic relatedness between the adjective pairs to apply to the American English represented by the TIME corpus, available at http://corpus.byu.edu/time/.

share the same base, generally of Latinate origin, e.g. analyt- , class- , com- , econom- , electr- , geometr- , graph- , history -, logist- , mag- , numer- , polit- , problemat -, egotist -, as the case of analytic - analytical pair is illustrated in (2).

(2) a. They are, indeed, intended in the main to provide an analytic framework for just such comparative and historical work.
b. Theoretical constructs such as ideal types, models, and paradigms provide an objective, analytical framework that we can use to study culture and change in institutions.

We see in (2) that the pair analytic - analytical occurs in a very similar context, where they collocate with the same noun 'framework', forming an NP complement, which is taken by the same verb 'provide'. This is true of the other pair adjectives. Thus, Bauer (1983: 122) points out that "Chomsky & Halle implies that pairs such as *economic / economical, electric / electrical, historic / historical* are simply free variants and synonymous, which is manifestly not the case." As the quote suggests, the semantic relatedness of these pairs of adjectives has been an object of linguistic studies.

One major concern about the semantic relatedness is whether and to what degree the two elements of each pair are semantically similar. The pair of *economic and economical* in (3) is admitted to be distinguished semantically by many linguists, with the first one generally meaning 'related to economy', as in (3a), the second one specifically meaning 'money-saving', as in (3b).

(3) a. This was a major economic loss for Florida since the citrus crop alone is worth roughly $3.5 billion.
b. This arrangement seemed to be the most popular and economical **way** to fish.

Then are they totally different in meaning or is there anything between them that is shared in their meaning? Another concern is on the other side of the same coin, that is, about whether and to what degree the two elements are semantically different. The pair of analytic and analytical is admittedly similar in meaning, and even usage in the context, as we see in (2). But is there anything that differentiates them in meaning?

The goal of this paper is to answer the questions posed above. We will inquire into the semantic relatedness between the paired adjectives ending in -ic versus -ical, based on the quantitative analysis of corpus data. For this purpose, we'll examine the previous analyses on the -ic versus -ical adjectives, especially Gries (2001), whose methodology has been a suit we follow here, in section 2. Besides previous analyses, we'll resort to a measure of semantic relatedness, originally developed for the on-line search engine, in order to observe a general tendency of semantic relatedness of those two adjectives. In section 3, we'll show how similar and how different the bigram adjectives are between each other, collecting and analyzing the corpus data with a few statistics tools. Section 4 concludes this paper.

2. Previous analyses

Linguists generally agree that the -ic versus -ical adjectives show a different degree of semantic relatedness. Plag (2003: 96) points out that "sometimes these forms (-ic versus -ical adjectives) are clearly distinguished in meaning (e.g. *economic* 'profitable' vs. *economical* 'money-saving'), in other cases it remains to be determined what governs the choice of the form over the other." Merchard (1969) makes a suggestion about the very question "what governs the choice of the form over the other" by resorting to the morphological structure of the two bigram adjectives, that is, economical -> economic + al , leading to the proposal that the -ic adjective is closer and more directly related to the base substantives than the -ical adjective. For example, when the adjective economic is compared with the adjective economical, the first -ic adjective is more related to the substantive 'economy', so that it can be defined as "related to economy", whereas the second -ical adjective is more abstracted from and more indirectly related to the substantive, so that it has a definition "money-saving". But this is not always the case. When the same criterion is applied to the pair *historic - historical*, the expected definitions do not match up with the definitions we have on the dictionaries. Even if the directness criterion can be applied to most cases, it is limited in that it still does not inform us about to what degree the adjectives are similar or different in meaning.

Gries (2001), however, makes a significant contribution to the

understanding of the -ic versus -ical adjectives with a help of a quantitative corpus linguistics. He uses a variety of statistical methods to investigate whether and to what degree the two bigram adjectives are similar or different in meaning[2]. There are two procedures in his methodology. One is calculating the percentage of the R1 collocates (the first elements on the right side of each bigram adjective) shared by the two adjectives and assigning their meeting percentages to a dot on the two-dimensional plane, called ESCO2, as a way to confirm the semantic similarity between them. The other is using a kind of t-test to determine the differentiating collocates between the two adjectives and to generalize how different the meanings of the two adjectives are. He concludes in either way. First, the -ic versus -ical shows a variety of degree of semantic similarity as shown below in Figure 1.

2) Technically speaking, the probabilistic statistical methods such as Tversky's similarity model and Biber's Principal Component Analysis contribute to the development of the Gries' ESCO coordinates, which show the degree of the pair's being interchangeably used in naturally occurring contexts.

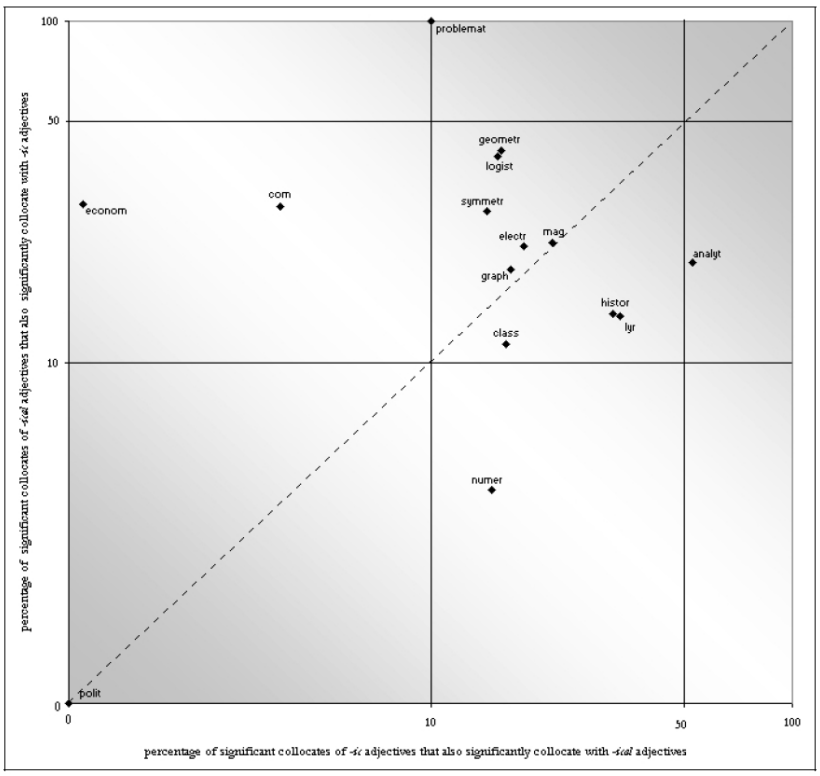

Figure 1: ESCO$_2$ for frequent adjectives ending in -ic *and* -ical *(excluding function words)*[16]

We see in Figure 1 that the nine paired adjectives, e.g. *geometric-geometrical, logistic-logistical, symmetric-symmetrical, electric-electrical, magic-magical, graphic-graphical, classic-classical, historic-historical,* within the grid of 10-50 percentage of X axis and 10-50 percentage of Y axis, show a semantic similarity, though it is moderate, while the other six pairs, e.g. *economic-economical, comic-comical, politic-political, numeric-numerical, analytic-analytical, problematic -problematical,* rather show a degree of semantic difference. Specifically, the pair *graphic-graphical* occupies a dot within 10-50 x 10-50 in the coordinate system, at which more or less than 20% of the shared ones in the whole R1 collocates of the adjective *graphic* meets with almost the same percentage of those of the adjective *graphical*. The almost same percentages of the shared R1 collocates between the two bigram

adjectives lead to the location of the meeting point on the slope line. So the adjectives scattered near around the slope line is concluded to be semantically close. But, by contrast, the meeting percentages of the pair '*economic-economical*' are assigned to the dot on the plane, where the percentage of the shared R1 collocates of the adjective *economic* is slightly over zero %, while the one of those of the adjective *economical* is somewhere within the 10-40 x grid. This lopsided distribution in shared collocates implies that the two adjectives are far away in meaning, and that the adjective *economic* has a predominant status over the other adjective *economical* (See details of methodology in section 3.1).

Secondly, he concludes that the unnoticed regularities of each bigram adjective can be detected through a detailed analysis of the patterns of differentiating collocates. Taking the *magic-magical* pair as an example, he showed properties of discriminating collocates in Figure 2 below, in which the adjective 'magic' can be defined with concrete terms whereas 'magical' can be with abstract terms.

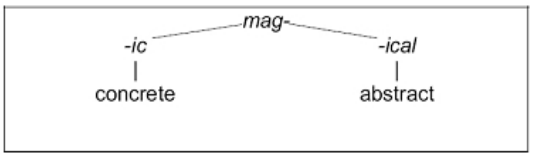

Figure2. Properties of discriminating collocates of magic versus magical adjectives

In spite of a significant contribution to the understanding of -ic versus -ical adjectives, Gries (2001) has a limited advantage in generalizing its results since he uses only British English data, collected from the 90 million word written part of the BNC corpus. So this paper will be complementary to Gries (2001) with the American English data collected from the 100 million word TIME corpus and nearly the same methodology.

Before winding up this section, it is helpful to refer to an on-line measure of semantic relatedness called 'MSR (specifically, LSA CU-tasa)', originally developed to sort out the search words according to their semantic relatedness and available at http://cwl-projects.cogsci. rpi.edu /msr/. This service provides us with a semantic relatedness between

a main term and its related terms, in our case, the -ic adjectives and the -ical adjectives, as shown in Table 1.

stem	analytic	classic	comic	economic	electric	geometric	graphic	historic	logistic	lyric	magic	numeric	politic	Problematic	egotistic	total
-ical	0.19	0.41	0.16	0.07	0.81	0.32	0.27	0.3	None	0.69	0.56	0.47	0.18	None	None	0.36

Table 1. Semantic Relatedness of *-ic* versus *-ical* adjectives (results from MSR)

With the overall semantic relatedness of 0.36 in a 0-1 scale, the pairs *electric-electrical (0.81), lyric-lyrical (0.69), magic-magical (0.56), numeric-numerical (0.47)* show a higher degree of semantic relatedness while the pairs *economic-economical (0.07), comic-comical (0.16), politic- political (0.18), analytic-analytical (0.19)* display a lower degree of semantic relatedness. This degree of semantic relatedness can serve as a reference point to the following quantitative corpus data analysis. The adjective pairs can be arranged in the order of semantic relatedness as in below.

(4)

logistic/ical

problematic/ical < economic/ical < comic/ical < politic/ical < analytic/ical

egotisti/ical

< graphic/ical < historic/ical < classic/ical < numeric/ical

<magic/ical < lyric/lyrical < electric/ical

3. Quantitative analysis of *-ic/-ical* adjective pairs

As we said in the preceding section, we collected data with regard to -ic versus -ical adjectives from the 100 million word TIME corpus (available at http://corpus.byu.edu/time/x.asp). One drawback with the online site is that the program is allowed to return 1,000 types at the maximum, meaning that we have a limited access to the corpus. Keeping this limit in mind, we will explore the semantic similarity of those 15 paired adjectives given in Gries (2001) and in the section 3.2, we will

go on to inquire into the semantic difference of them.

3.1 Semantic similarity of *-ic/-ical* adjectives

Given that significant collocates make up a semantic feature of a word, according to Biber (1993), we can look into the semantic similarity by examining the shared significant collocates between words, here -ic and -ical adjectives again. Like Gries (2001), we sorted the types of R1 collocates accompanied by each bigram adjective according to the result of the -2 log Λ and went on to remove the cases where the x2 value exceeds the threshold value 6.63, for the significance of p=o.o1 with df=1. In other words, these cases occur much more times than the expected frequency so that they are suspected to occur by chance. The statistic result is shown in Table 2, where SC is short for significant collocates.

stem	analyt-	class-	com-	economic-	electr-	geometr-	graph-	histor-	logist-	lyr-	mag-	numer-	polit-	Problemat-	egotist-	total
-ic	61	997	1000	1000	990	112	281	991	13	235	494	3	22	41	10	**6189**
SC	59	994	993	979	980	98	272	970	13	227	480	3	22	39	10	**6080**
-ical	122	607	75	141	490	58	10	1000	79	288	264	124	1000	12	31	**4179**
SC	120	603	75	141	484	58	10	972	76	288	263	121	981	12	28	**4112**
shared	18	178	20	39	193	18	4	254	3	68	76	2	8	1	0	**864**
ESC O X	30.5 %	17.9 %	2.0 %	4.0 %	19.7 %	18.4 %	1.5 %	26.2 %	23.1 %	30.0 %	15.8 %	66.7 %	36.4 %	2.6 %	0.0 %	14.2 %
ESC O Y	15.0 %	29.5 %	26.7 %	27.7 %	39.9 %	31.0 %	40.0 %	26.1 %	3.9 %	23.6 %	28.9 %	1.7 %	0.8 %	8.3 %	0.0 %	21.0 %

Table2. The number of R1 collocates after each -ic/-ical adjectives from TIMES corpus

Taking the pair analytic-analytical as an example, the adjective analytic shows significant 59 R1 collocates, by two collocates less than the number of raw data while the adjective analytical has 120 significant R1 collocates after the application of the log likelihood test and the x2 test. These two adjectives have shared 18 R1 collocates, followed by the calculation of the percentages of those shared collocates in each bigram adjective, 30.5% and 15.0% respectively. The percentage of the first adjective on

the X axis (ESCO X) meets with that of the second adjective on the Y axis (EXCO Y) at a dot within 10-50 x 10-50 grid. This way, the semantic relatedness of the adjectives in question is represented in Figure 3.

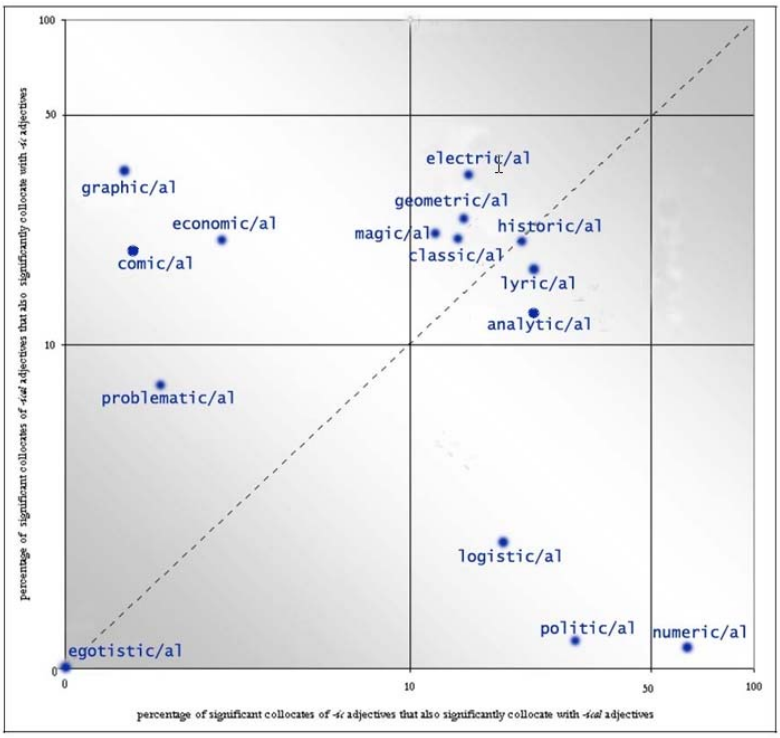

Figure 3 : *ESCO₂ for frequent adjectives ending in* -ic *and* -ical *(excluding function words)*

We see in Figure 3 above that the 7 adjective pairs within the grid of 10-50 on the X axis and 10-15 Y axis, (e.g. *electric-electrical, geometric-geometrical, magic-magical, historic-historical, classic- classical, lyric-lyrical, analytic-analytical*) has a semantic overlap on the basis of the behaviors of R1 collocates, even though moderate, whereas the other pairs outside the grid above, (e.g. *graphic-graphical, economic-economical, comic-comical, problematic-problematical, egotistic -egotistical, logistic-logistical, politic-political, numeric -numerical*) has rather a low degree of semantic overlap. A more detailed observation of each case

returns more interesting results. Before looking into each case, we should note that the larger ratio of shared significant collocates reflects more similarity to the other of the smaller ratio than vice versa. The first case in point is that of the pair *politic-political*. Its dot in the lower left-hand grid indicates that the adjective *politic* is more similar to the adjective *political* than vice versa. It follows that the R1 collocates of the adjective *politic* are subsumed by those of the adjective political. That is, the R1 collocates after the adjective politic have no trouble in occurring with the adjective *political*, but not vice versa. This tendency shows a stark contrast to that of Gries (2001), in which the paired adjectives lack any significant R1 collocates, meaning that the two adjectives are totally different in meaning. This is due to the difference in the usage between American English and British English. The pair *egotistic - egotistical* is considered in this study instead of *symmetric - symmetrical* pair since the first item symmetric in the removed pair has no tokens in the TIME corpus. Any way, *egotistic - egotistical* pair is positioned on the zero point in the coordinate, indicating that the two adjectives have no semantic relatedness.

Another case worth mentioning is that of the *historic-historical* pair. In Gries (2001), the pair is positioned somewhat far away from the diagonal line inside the semantic overlap grid, meaning that the meaning of the adjective *historic* is subsumed by the adjective *historical* in British English. But this study shows that the pair is almost or rightly on the diagonal line, indicating that the two adjectives have symmetric meanings between each other so we can tell that they are very similar in meaning in American English. Also, the *economic - economical* pair is moved slightly to the right, compared to Gries (2001), which means the extreme dominance of *economic* over *economical* in British English is weakened by such moving distance in American English.

Finally, the movement of the pair *logistic-logistical* from the semantic overlap grid in Gries (2001)'s figure to the lower right-hand side in this study is noteworthy. Unlike the case in British English where the pair has some degree of semantic similarity, the pair is rather different in meaning in American English with the adjective *logistical* stronger in taking collocates.

Compared to the semantic relatedness results between -ic versus -ical adjectives from MSR in Table 1 and (4) above, almost all the adjective pairs with the semantic related value of above the average value 0.36

are positioned near around the slope line in the figure above. The adjective pair *numeric - numerical* is an exception. Although the bigram adjectives has a relatively high value 0.47 of semantic relatedness, they are placed at the dot on the right-hand bottom of the coordinate. This implies that the two adjectives, at least in the TIMES corpus, are not used interchangeably and that the meaning of the adjective *numeric* is subsumed by the adjective numerical. One more interesting comparison is that the adjectives with a value of 'none' in the MSR results are located far away from the slope line, rather on the corners of the coordinate, showing the symmetry between MSR and ESCO results. But the adjective pair *analytic - analytical* pair is positioned 10-50 x 10-50 grid in spite of the relatively low value (0.19) in the MSR test.

3.2 Semantic difference of *-ic/-ical* adjectives

Given that discriminating collocates differentiates the meanings of the words, we can explore the semantic difference by examining the discriminating R1 collocates between -ic and -cal adjectives. For illustration, we sorted the types of R1 collocates in the order of the 41 R1 discriminating collocates of the adjective *analytic*, the 20 shared collocates of theses adjectives, and the 102 R1 discriminating collocates of the adjective analytical from the left to the right, as shown in Table 3 below.

actor, brain, circles, coldness, congresses, mode, movement philosophers, practice, principles, treatment, procedure, processes(41)	capabilities, course, criticism, geometry, intelligenced,...(20)	abilities, ... vision ... work (102)

Table 3. Discriminating and Shared Collocates of *analytic* versus *analytical* from TIMES corpus

Then we conducted a t-test to pick out the insignificant case where the observed frequency is over the expected frequency, leading to the

high probability that it occurs by chance. A kind of t-test, called SISA, serves our purpose in that it returns a t-value (or z-value) with the significance of 95% for each case, as shown in Table 4.

analytic		analytical	
R1 collocate	t ; p	R1 collocate	t ; p
treatment(5)	-1.526 ; p=o.12	approach(5)	-1.438; p=0.1505
technique(2)	-0.447 ; p=0.65	epilogue(2)	-0.343; p=0.73
brain(1)	0.135; p=0.89	couch(1)	0.24; p=0.8101
school(1)	0.135; p=0.89	charts(1)	0.24; p=0.8101
theory(1)	0.135; p=0.89	wizard(1)	0.24; p=0.8101

Table 4. Significant Discriminating R1 Collocates of *analytic* versus *analytical*

We see from the Table 4 that the t-values of each of the sample in the R1 collocates of the adjective analytic are all beyond the significance breakpoint of 0.05. It is true of those of the adjective *analytical*. This failure is mainly due to the small size of the samples. In fact, this analysis is done with 49 tokens of discriminating collocates of the adjective analytic while it is done with 140 tokens of those of the adjective analytical. Going back to (4), when we compare the discriminating collates of the adjective *analytic* with those of the adjective *analytical*, we can obtain a tendency that the adjective *analytic* has a property of the academic / scientific field (e.g. school, theory, technique, treatment, etc), the adjective *analytical* has a property of general fields (e.g. ability, couch, approach, wizard, etc.)

4. Conclusion

Through this study on the semantic relatedness of the -ic versus -ical adjectives, we conclude that they are somewhat similar and somewhat different, on the basis of the quantitative data and statistical test. First,

we could confirm the degree of semantic similarity between those pairs by calculating the percentages of each bigram adjective and comparing the location of dots in ESCO2. The pairs *electric - electrical, geometric - geometrical, historic - historical, magic - magical, classic - classical, lyric - lyrical, analytic - analytical* are interpreted to have close semantic relatedness between the pair items. The pairs, on the other hand, *graphic - graphical, economic - economical, comic - comical, problematic - problematical, egostic - egostical, logistic - logistical, politic - political, numeric -numerical* are semantically far away from each other and one item of the pair has more dominant status in the distribution in the TIME corpus. Second, we were able to look briefly into the semantic difference of the bigram adjectives, actually one case of them for illustration by comparing each R1 collocate with the total number of the R1 collocates by means of t-test. Unfortunately, we failed to attain significant data set of discriminating collocates due to the small size of the samples and the lack of allowed time. But this study is beneficial to show a significant difference in the usage of -ic versus -ical adjectives by examining the different R1 collocates that is modified by each item of the adjective pair. Also, this study helps us to grasp the understanding of the difference in meaning of -ic versus -ical adjectives. Into the bargain, this study gives a detailed explanation on the cut off of the insignificant data and the way to draw the ESCO coordinate, which was only briefly mentioned in Gries (2001).

This study has two limitations by itself. The first problem is that the data collected for this study came from the TIME corpus, a collection of weekly news magazine articles. So this is not straightforwardly comparable to the British National Corpus in Gries (2001). With the American National Corpus, which will be available in February, 2008, this kind of study could enable us to compare the lexical items in meaning and usage between British English and American English. The second restriction is that when examining the semantic difference between the bigram adjectives, we select the pair *analytic-analytical* for illustration, but we couldn't obtain statistically significant data, for shortage of the samples. It would have been better to choose the pair with even more samples, such as *economic - economical* or *historic - historical* pair, since they are likely to return R1 collocates, the t-scores of which are below the breakpoint of 0.01.

References

Bauer, Laurie. 1983. *English Word Formation.* The Press Syndicate of the University of Cambridge, 220-225.

Biber, Douglas. 1993. Co-occurrence Patterns among Collocations: A Tool for Corpus-Based Lexical Knowledge Acquisition. *Computational Linguistics* 19:531-538.

Gries, 2001. A corpus-linguistic analysis of English -ic vs -ical adjectives, *ICAME journal* No. 25: 65-108

Merchand, Hans. 1969. *The Categories and Types of Present-Day English Word-Formation. A Synchronic-Diachronic Approach.* 2nd ed. Muenchen: Beck.

Plag, Ingo. 2003. *Word-Formation in English.* The Press Syndicate of the University of Cambridge, 86-97.

Tversky, Amos. 1977. Features of Similarity. *Psychological Review* 84: 327-352.

Myoung Hyoun Song
y720four@yahoo.com