Corpus approaches to antonymy

Steven Jones
University of Central Lancashire

1. Introduction

The word "antonymy" was coined in 1867 by CJ Smith[1] to describe word-pairs - commonly known as "opposites" - such as *hot/cold*, *girl/boy* and *buy/sell*. Some linguists (e.g. Lyons (1977) and Cruse (1986)) apply the term "antonymy" restrictively and would only identify the first of these three pairs as being truly antonymous. However, this seems counter-intuitive in many respects, and this paper will use "antonymy" in its broader sense to refer to all word-pairs which could reasonably be identified as "opposites" by speakers of English.

Antonymy is often defined simply as "oppositeness of meaning" (Palmer 1976: 94). However, the problem with an exclusively semantic definition is that it fails to explain, or even acknowledge, the tendency for certain words to become enshrined as "opposites" in language while others do not. For instance, *rich* and *poor* would be regarded as antonyms because they occupy opposite ends of the same scale, namely the scale of wealth. *Affluent* and *broke* also occupy opposite ends of this scale, but, intuitively, one would be reluctant to describe them as antonyms. Therefore, antonymy should be defined according to lexical as well as semantic criteria. It is a phenomenon "specific to words rather than concepts" (Justeson & Katz, 1991: 138).

This paper will examine some of the ways in which advances in corpus technology enable antonymy to investigated afresh. Firstly, the categories to which antonymous pairs have been logically assigned will be summarised. Secondly, a set of new, corpus-based categories will be presented. Thirdly, a statistical analysis of co-occurrence rates among antonyms will be offered. And finally, ways of identifying new antonyms, again using corpus data, will be explored.

2. Traditional classes of antonymy

The meanings of antonymous pairs have been logically examined by a number of linguists (e.g. Lyons 1977, Kempson 1977, Cruse 1986, etc.) and antonyms have been classified according to their theoretical differences, perhaps at the expense of their intuitive similarity. Using terminology favoured by Leech (1974), each of the traditional categories of antonymy will now be outlined.

2.1. Binary Taxonomy

The name given by Leech to antonymous pairs such as *man/woman*, *alive/dead* and *married/unmarried* is "binary taxonomy" (1974: 109). Other writers (see Palmer 1972, Jackson 1988, Carter 1987) prefer to speak of "complementarity". Kempson - whose favoured term is "simple binary opposition" - describes examples of Binary Taxonomy as "the true antonyms" (1977: 84). However, this description is particularly confusing in light of the unwillingness of other linguists - namely Cruse (1986) and Lyons (1977) - to acknowledge Binary Taxonomy as a form of antonymy at all. The criterion necessary for an opposition to be considered binary is that the application of one antonym must logically preclude the application of the other. For instance, if X is a *smoker*, X cannot be also a *non-smoker*; if X is *baptised*, X cannot be also *unbaptised*, and so on.

2.2. Multiple Taxonomy

Multiple Taxonomy - also known as Multiple Incompatibility (Carter 1987:19) - is a borderline classification of antonymy that refers to pairs such as *summer/winter* and *north/south*. In some respects, this category is akin to Binary Taxonomy. The pair *male* and *female*, for example, belong to a two-member system, such that X can never be simultaneously more than one member; *solid*, *liquid* and *gas*, by comparison, belong to a three-member system, such that X can never be simultaneously more than one member; similarly, *clubs*, *diamonds*, *hearts* and *spades* belong to a four-member system, such that X can never be simultaneously more than one member. And so on. Thus, Multiple Taxonomy may be seen as Binary Taxonomy extended to three or more terms. Whether such examples remain within the boundaries of antonymy is debatable.

---

[1] See Muehleisen (1997) for more details or "Introductory Matter" in *Webster's Dictionary of Synonyms* (1951: vii-xxxiii).

2.3. Polar Opposition

Polar Opposition differs from Binary Taxonomy because one antonym is not automatically debarred by the other's application. In other words, it is possible to be neither *tall* nor *short* in a way that it is not possible to be neither *male* nor *female*. Thus, *tall/short* are said to be a Polar Opposition, as are the majority of everyday opposites (*old*/*new, cold/hot, wet*/*dry*, etc.). Because Polar Oppositions are not mutually exclusive, they are readily modified (*quite happy, extremely happy, fairly happy*, etc.) and can take both comparative (*happier*) and superlative (*happiest*) form. Indeed, many commentators (e.g. Lyons 1977, Cruse 1986, Jackson 1988) prefer to label such pairs Gradable Antonymy.

2.4. Relative Opposition

An example of Relative Opposition is *tenant/landlord*. The statement *X is the landlord of Y* entails and is entailed by *Y is the tenant of X*. Therefore, *landlord* and *tenant* belong to a reciprocal relationship, also reflected by pairs such as *teach/learn, buy/sell* and *above/below*. The majority of semanticists label this phenomenon "converseness" and Kempson notes that if the variables X and Y are converse verbs, the statement *A X B* implies *B Y A* and the statement *A Y B* implies *B X A*. In other words, *B precedes A* implies that *A follows B*, and *A follows B* implies that *B precedes A* (1977: 85). A fertile area for Relative Antonymy is the field of kinship relations. If X is the *grandparent* of Y, then Y must be the *grandchild* of X; if *X is the husband of Y*, then *Y must be the wife of X*.

2.5. Other Categories

Leech identifies two further categories of antonymy: "hierarchy" (1974: 106) is similar to Lyons's notion of "rank" and describes the relationship between sets of terms such as *January/February/March* and *one/two/three*; "inverse opposition" describes pairs such as *all/some* and *remain/become*[2] which may not otherwise be regarded as "opposites". Other types of antonymy include "orthogonal" and "antipodal" opposition (Lyons, 1977: 286). Orthogonal - meaning perpendicular, at right angles - describes the antonymy holding between the words *man, woman, girl* and *boy*. Each of these four words contrasts with two of the other three. So *man* can be the antonym of *boy* and *woman*, but not *girl*; and *boy* can be the antonym of *girl* and *man* but not *woman*. An example of an antipodal opposition would involve the terms *north, east, south* and *west*. Here, words only contrast in one direction. So *north* is an antonym of *south*, but not *east* or *west*; and *west* is an antonym of *east* but not *north* or *south*.

3. New classes of antonymy

The categories outlined above are useful if we wish to look at antonymy from a logical perspective. However, it has now become possible for antonymy to be approached from a corpus-based angle, with new classes being created to describe not what antonyms <u>are</u>, but what antonyms actually <u>do</u> in text.

3.1. Data and Methodology

In order to ascertain and quantify the various textual functions of antonymy, a database of 3,000 sentences was constructed. Each of those sentences were retrieved from a large corpus and features both members of a recognisable antonymous pair. The corpus which I chose to use consists of about 280 million words from *The Independent*. All stories printed in the newspaper between 1 October 1988 and 31 December 1996 are included in the corpus. Journalistic corpora are suitable for studies of this nature because they are large, genre specific and reflect a natural, modern, non-fictional use of written language. Thus, an overview of how antonymy is used in the field of broadsheet newspaper journalism is possible, although it should be acknowledged that antonymy might be found to function differently in other corpora.

Selecting a representative sample of antonymous pairs is more problematic. It is difficult to imagine a list of antonyms which would not raise a single eyebrow, either because of words included but not considered to be "good opposites", or because of "good opposites" which might be conspicuous in their absence from the list.

---

[2] Leech's reasoning is that *<u>some</u> artists have no formal training* is synonymous with *not <u>all</u> artists have formal training* and *she did not <u>become</u> a smoker* is synonymous with *she <u>remained</u> a non-smoker*.

Other corpus-based investigations into antonymy, such as Justeson & Katz (1991) and Mettinger (1994), resolved this problem by using an existing index of antonyms. Justeson & Katz chose to make use of the 40 "historically important" (1991: 142) antonyms identified by Deese (1965). This list of antonyms was based entirely on the results of word association tests. Deese took 278 adjectives[3] and used them to elicit responses from 100 informants. When a pair of contrast words successfully elicited one another more than any other word, they were added to the list of antonymous pairs, which ultimately numbered 40. Though all antonyms cited by Deese fulfilled this requirement, some passed the test with alarmingly low scores. For example, given the stimulus *together*, only 6% of informants replied *alone*; given *alone*, only 10% replied *together*. However, this was evidently enough to make these answers more popular than any other, even though the fact remains that a minimum of 84% of informants failed to give either *alone* as a response to *together*, or *together* as a response to *alone*. Indeed, of the 278 adjectives tested, only one word succeeded in eliciting its antonym on a majority of occasions (*left*, to which 51% of informants gave *right*). Therefore, though there remains a strong tendency for informants to provide antonyms as responses to given stimuli in word associations tests, it may not be wise to treat Deese's 40 antonyms as being in any sense exhaustive or definitive.

A different approach was taken by Mettinger (1997), who used Roget's Thesaurus as his source for antonyms. Created in the middle of the nineteenth century, Roget's Thesaurus attempted to catalogue language, not in alphabetical order, but according to "ideas". This is of relevance to a study of antonymy because Roget chose, where possible, to present these ideas in opposition to one another. Thus, the thesaurus begins by listing words associated with *existence*, then considers words associated with *inexistence*. Following next are *substantiality* and *insubstantiality*, then *intrinsicality* and *extrinsicality*.

Neither using the Deese Antonyms nor turning to thesaural listings is ideal. Essentially, one is still dependent on the intuitions of others to identify antonymous pairs. In the case of Roget, these intuitions are 150 years out of date and "contain a number of lexical items that are hardly used in contemporary English" (Mettinger, 1994: 94); in the case of the Deese antonyms, one is dependent on the criteria for antonymy established by 1960s' schools of psychology. However, it is impossible to rely on anything other than intuition when it comes to a psycholinguistic phenomenon such as antonymy. No exhaustive list of antonyms will ever be produced because the process which gives a pair of words antonymous status is complex and dynamic. Indeed, this status could only really be gauged by consensus, as definitions of antonymy vary not only from one linguist to the next, but also from one mental lexicon to the next. With these limitations in mind, I decided that the best approach would be to create a new list of antonyms, customised to meet the demands of this research and relevant to a 21st Century investigation of antonymy:

active/passive	advantage/disadvantage	agree/disagree	alive/dead
attack/defend	bad/good	badly/well	begin/end
boom/recession	cold/hot	confirm/deny	correct/incorrect
difficult/easy	directly/indirectly	discourage/encourage	dishonest/honest
disprove/prove	drunk/sober	dry/wet	explicitly/implicitly
fact/fiction	fail/succeed	failure/success	false/true
fast/slow	female/male	feminine/masculine	gay/straight
guilt/innocence	happy/sad	hard/soft	hate/love
heavy/light	high/low	illegal/legal	large/small
long/short	lose/win	major/minor	married/unmarried
new/old	officially/unofficially	old/young	optimism/pessimism
optimistic/pessimistic	peace/war	permanent/temporary	poor/rich
private/public	privately/publicly	punishment/reward	quickly/slowly
right/wrong	rightly/wrongly	rural/urban	strength/weakness

Table One: Antonymous pairs selected for inclusion in the database

Any native speakers could be reasonably expected to identify the antonym of each of the above 112 words. Most core antonyms are represented and the list also features a number of lower frequency pairs, including antonymous nouns, adverbs and verbs, which previous investigations have been inclined to overlook.

[3] Deese's list included words such as *above*, *inside* and *bottom* which function as adjectives less often than they function as other parts of speech.

In total, 2,844 sentences were retrieved at random from the corpus, all of which features both members of one of the above antonymous pairs. A further 156 sentences were retrieved in which a word co-occurred with an *un-* version of itself. As *un-* is the most prolific morphological marker of antonymy in English, this was a useful way to enrich the database with antonymous pairs which, though often less familiar than the 56 pairs selected for analysis, still reflect opposition and are instantly recognisable as antonyms because of their morphology.

3.2. Classifying the database

All 3,000 database sentences have been classified according to their textual function. Eight categories are presented below, in alphabetic order only, together with three illustrative examples.

3.2.1. Ancillary Antonymy

Sentences attributed to this category contain two contrasts: that between the established pair of antonyms (in bold) and that between a pair of words or phrases which would not usually be interpreted contrastively (in italics). Here, it would appear that the antonyms function as lexical signals. They serve an "ancillary" role, helping us to process another, perhaps more important, opposition nearby.

- ind902[4]: At Worcester on Wednesday, Botham - apart from bowling well - was wandering around in a T-shirt with the message: '*Form* is **temporary**, *class* is **permanent**'.
- ind913: Broadly speaking, the community charge was **popular** with *Conservative* voters and **unpopular** with *Labour* voters.
- ind891: Robin Cook, Labour's health spokesman, demanded: 'How can it be **right** to limit the hours worked by *lorry drivers and airline pilots*, but **wrong** to limit the hours of *junior hospital doctors undertaking complex medical treatment*?'

3.2.2. Comparative Antonymy

This category is home to sentences in which a pair of antonyms are set up in comparison with one another. This function of antonymy is often expressed by a lexico-syntactic framework such as *more X than Y* or *X is more [adjective] than Y*.

- ind891: And it is possible to accept both that Dr Higgs was a lot more **right** than **wrong** in her diagnoses, but that it is now impossible for her to return.
- ind923: 'Well,' said Cage, completely unabashed, 'some living composers are more **dead** than **alive** '.
- ind903: Training would be based upon rewarding good behaviour, because behaviourists, Skinner argued, had found that **reward** is more effective than **punishment**.

3.2.3. Co-ordinated Antonymy

The antonymous pair in each of the examples below is presented in a unified, co-ordinated context. The function of such antonyms is to identify a scale, then exhaust that scale. The contrastive power of each pair remains untapped because their purpose is to express inclusivity. Mostly, antonyms which serve this role are conjoined by *and* or *or*.

- ind953: He showed no disloyalty, **publicly** or **privately**, to Virginia Bottomley though it must have irked him that she was in the Cabinet and he was not.
- ind921: Whitehall was yesterday unable to **confirm** or **deny** other simulated devolutions.
- ind941: Again in debates over genetic research it is significant that Christians, Muslims and Jews have united, **implicitly** and **explicitly**, in condemning a low view of the value of embryonic life.

[4] The notation gives detail about where and when each sentence was published (newspaper; year; quarter). For example, this sentence was published in the second quarter of 1990 by *The Independent*.

3.2.4. Distinguished Antonymy

The sentences below refer, in a metalinguistic fashion, to the semantic dissimilarity between antonyms. The framework which houses the antonyms most frequently is *n between X and Y*, where *n* is *difference* or a synonym thereof.

- ind892: But far from that, Mortimer's father had not given him even a basic moral education, such that today he still doesn't know the difference between **right** and **wrong**, or so he said.
- ind931: But it made the point that the division between **gay** and **straight** is one of many rifts in our society.
- ind884: Mr Craxi's fresh-faced deputy, Claudio Martelli, also dissented, saying that 'one must distinguish between **hard** and **soft** drugs'.

3.2.5. Extreme Antonymy

Sentences classified in terms of Extreme Antonymy are similar to Co-ordinated Antonymy examples. The difference is that here a contrast is set up, not between antonyms, but between both ends of a semantic scale, on one hand, and the semantic space in between, on the other. Typical frameworks show antonyms linked by *or* or *and*, and premodified by an extremity-signalling adverb such as *very* or *too*.

- ind892: No-one can afford to go to law except the very **rich** and the very **poor** and it can't possibly get any worse.
- ind903: The advantages are that the track does not need watering, and can be used when conditions are either too **dry** or too **wet** for racing on turf.
- ind964: Freud maintained in Civilization and Its Discontents that human beings feel a deep **hate** and a deep **love** for civilization.

3.2.6. Idiomatic Antonymy

Many antonymous pairs co-occur as part of a familiar expression, proverb or cliché. Such examples have been assigned to category of Idiomatic Antonymy.

- ind944: The **long** and the **short** of it is that height counts.
- ind893: They evidently knew they could teach this **old** dog a few **new** tricks.
- ind892: Whoever said the **female** of the species was more deadly than the **male** hadn't met Lord William Whitelaw.

3.2.7. Negated Antonymy

Arguably the purest function of antonymy, the sentences below each negate one antonym in order to place additional emphasis on the other or to identify a rejected alternative. The most common framework for this class is *X not Y*.

- ind884: Well, without the combination of an arms race and a network of treaties designed for **war**, not **peace**, it would not have started.
- ind912: Democracy means more than the right to pursue one's own self-interest - government must play an **active**, not **passive**, role in addressing the problems of the day.
- ind893: However, the citizen pays for services to work **well**, not **badly**.

3.2.8. Transitional Antonymy

The function of antonyms belonging to this category is to help describe a movement from one state to another. This transition is usually expressed by a framework such as *from X to Y* or hinges around the verb *to turn*.

- ind923: Her film career similarly has lurched from **success** to **failure**, with enormous periods out of work.

- ind934: The atmosphere of the negotiations was tense, discussion uneven, the mood in both camps swung from **optimism** to **pessimism**.

- ind923: Inflation is a tax which redistributes wealth to the **sophisticated** from the **unsophisticated**.

3.3. Frequency of New Classes

The eight classes outlined above collectively account for 2,894 of the 3,000 database sentences. The remaining 3.5% of contexts demonstrate that antonyms can also function in unusual, sometimes innovative, ways about which it difficult to generalise. The table below shows the statistical distribution of all database sentences.

	ancill	co-or	comp	distin	trans	negat	extre	idiom	other	total
active/passive	53	14	9	6	6	6	-	-	2	96
advantage/disadvantage	15	14	2	-	4	1	-	-	-	36
agree/disagree	26	17	3	-	-	-	-	-	3	49
alive/dead	16	26	9	1	-	1	-	-	1	54
attack/defend	10	15	3	-	-	2	-	-	-	30
bad/good	55	47	4	4	3	1	-	2	1	117
badly/well	31	15	4	1	-	2	-	-	-	53
begin/end	24	23	3	-	-	1	-	-	-	51
boom/recession	12	3	4	-	5	-	-	-	-	24
cold/hot	21	23	-	-	2	-	2	11	-	59
confirm/deny	-	34	-	-	-	-	-	-	-	34
correct/incorrect	6	11	-	1	-	-	-	-	-	18
difficult/easy	19	5	-	-	-	-	1	-	2	27
directly/indirectly	21	57	1	-	-	-	-	-	-	79
discourage/encourage	16	8	2	-	-	2	-	-	-	28
dishonest/honest	8	4	-	-	-	-	-	-	-	12
disprove/prove	-	14	-	-	-	-	-	-	-	14
drunk/sober	8	7	-	-	1	1	-	-	1	18
dry/wet	11	9	3	1	3	-	4	-	-	31
explicitly/implicitly	6	19	2	-	-	3	-	-	-	30
fact/fiction	5	5	2	11	2	4	1	-	6	36
fail/succeed	30	27	5	-	-	1	-	-	-	63
failure/success	38	20	10	12	1	6	1	-	-	88
false/true	10	34	3	11	-	1	1	-	2	62
fast/slow	17	7	2	1	-	-	-	-	1	28
female/male	23	43	1	4	1	-	-	2	13	87
feminine/masculine	37	18	2	3	-	-	1	-	7	68
gay/straight	3	20	7	1	1	1	-	-	-	33
guilt/innocence	5	27	3	5	1	2	-	-	1	44
happy/sad	22	17	2	-	2	-	2	-	-	45
hard/soft	17	3	2	1	3	2	3	-	1	32
hate/love	40	44	7	2	1	2	2	-	6	104
heavy/light	46	19	5	1	4	-	2	-	-	77
high/low	20	3	2	3	1	1	1	1	-	32
illegal/legal	10	17	1	-	3	-	-	-	-	31
large/small	17	23	4	2	2	-	2	-	-	50
long/short	22	7	4	1	-	1	-	1	-	36
lose/win	27	25	5	-	-	1	-	-	-	58
major/minor	11	9	-	3	4	-	-	-	-	27
married/unmarried	4	14	8	5	-	-	-	-	-	31
new/old	81	76	21	19	10	3	1	6	37	254
officially/unofficially	14	10	1	-	-	-	-	-	-	25
old/young	20	34	6	5	-	1	3	-	-	69
optimistic/pessimistic	30	12	3	-	1	-	1	-	-	47
optimism/pessimism	9	1	2	-	6	1	1	-	1	21
peace/war	3	5	1	1	1	2	-	-	2	15
permanent/temporary	6	12	5	1	3	1	-	-	-	28
poor/rich	46	16	6	24	1	-	5	-	4	102
private/public	36	68	6	13	5	2	-	-	4	134
privately/publicly	20	24	2	1	-	-	-	-	-	47
punishment/reward	6	5	4	-	-	3	-	-	1	19
quickly/slowly	16	6	2	-	-	-	4	-	-	28
right/wrong	36	13	1	5	1	-	-	-	4	60
rightly/wrongly	1	43	-	-	-	-	-	-	-	44
rural/urban	7	13	-	2	1	-	1	-	-	24
strength/weakness	11	6	6	-	4	4	-	-	4	35
un-words	58	60	15	10	7	3	1	-	2	156
TOTAL:	**1162**	**1151**	**205**	**161**	**90**	**62**	**40**	**23**	**106**	**3000**
TOTAL (%):	**38.7**	**38.4**	**6.8**	**5.4**	**3.0**	**2.1**	**1.3**	**0.8**	**3.5**	**100**

Table Two: Statistical breakdown of database classes

Table One demonstrates that the most popular category is that of Ancillary Antonymy, to which 38.7% of all sentences have been attributed. Recording only 11 fewer sentences is Co-ordinated Antonymy, which accounts for 38.4% of all database sentences. These two classes are significantly larger than any others and collectively account for 77.1% of sentences. The third largest category identified is Comparative Antonymy, but this is only a fraction the size of the two major categories. 205 sentences have been attributed to the class of Comparative Antonymy, less than 7% of the database. Distinguished Antonymy accounts for a further 5.4% of sentences, but no other class of antonymy is represented by more than 3% of the total sample.

Perhaps most remarkable is that the majority of pairs, regardless of their word class, follow a similar pattern of distribution. For example, Ancillary Antonymy and Co-ordinated Antonymy are the most commonly occurring categories, but this is not just because they are each strongly favoured by a small number of pairs. Rather, this pattern is consistent among almost all pairs. Indeed, in the case of 44 of the 56 pairs sampled, Ancillary Antonymy and Co-ordinated Antonymy each account for more sentences than any other category. This suggests that any given word-pair is likely to have a predictable textual profile. However, some pairs may have unusual individual distributions. For example, all 34 sentences retrieved which feature *confirm/deny* are assigned to the class of Co-ordinated Antonymy. This is because refusing to confirm or deny a proposition has become a cliche among politicians and other public figures.

4. Antonym Co-occurrence

One of the questions raised by my research is this: exactly how widespread is antonymy in language? Gauging the true answer is very difficult. Firstly, there is the problem of defining antonymy: the stricter the definition one uses, the less pervasive the phenomenon will appear. Then there is the even greater problem of counting: in order to arrive at an estimate of the proportion of sentences which feature antonyms, one would need to identify every single antonymous pair in use, then retrieve all sentences which features both of those words. But that would not be all - one would then need to edit all of these sentences manually (which would number over a million in my corpus) to eliminate those in which the word-pair do not function antonymously (coincidental co-occurrence is common among higher frequency pairs, especially those which feature a polysemous term, such as *well*). Only then could one arrive at an approximation of the proportion of corpus sentences which feature antonyms, and this approximation would still fail to account for inter-sentential antonymous usage.

An easier way to estimate the prevalence of antonymy in text is to compare the expected co-occurrence rate of antonyms with their observed co-occurrence rate. Therefore, I shall now examine each of the 56 antonymous pairs in my sample to determine whether those antonyms co-occur more or less than would be expected by chance.

Listed below are the 56 antonymous pairs selected for study in this thesis. Each pair is followed by five columns of figures: columns one and two simply record the raw frequency of each antonym in the corpus; column three records the number of sentences one would expect to feature both antonyms if those words co-occurred at random; column four records the number of sentences in the corpus which, in reality, contain both antonyms; column five records the Observed/Expected ratio, which is generated by dividing the figure in column four by the figure in column three.

word one (W1) / word two (W2)	raw frequency of W1	raw frequency of W2	Expected Co-occurrence	Observed Co-occurrence	Observed / Expected
active/passive	11411	2033	1.8	172	95.6
advantage/disadvantage	21531	2483	4.2	69	16.4
agree/disagree	18196	2472	3.5	153	43.7
attack/defend	43395	9198	31.0	273	8.8
cold/hot	16466	16026	20.5	751	36.6
correct/incorrect	10529	1484	1.2	34	28.3
dead/alive	32214	11661	29.2	565	19.3
deny/confirm	7514	6595	3.9	335	85.9
difficult/easy	54244	31395	132.4	434	3.3
directly/indirectly	14172	1377	1.5	492	328.0
drunk/sober	4730	1878	0.7	56	80.0
dry/wet	10978	5109	4.4	348	79.1
encourage/discourage	12586	1614	1.6	77	48.1
end/begin	145438	19682	224.6	740	3.3
explicitly/implicitly	1320	813	0.1	32	320.0
fact/fiction	78900	7391	45.3	503	11.1
fail/succeed	10963	8258	7.0	131	18.7

fast/slow	22625	17374	30.6	350	11.4
feminine/masculine	1191	903	0.1	140	1400.0
good/bad	181876	47247	668.1	4804	7.2
guilt/innocence	4229	3804	1.3	162	124.6
happy/sad	28217	9420	20.7	140	6.8
hard/soft	68635	11960	63.8	526	8.2
high/low	93232	41088	297.8	2847	9.6
honest/dishonest	6922	1084	0.6	28	46.7
legal/illegal	40832	11208	35.6	302	8.5
light/heavy	36832	22898	65.6	297	4.5
long/short	131582	52119	533.2	2168	4.1
love/hate	42541	6108	20.2	511	25.3
major/minor	45452	10624	37.5	432	11.5
male/female	16930	14883	19.6	2556	130.4
married/unmarried	25581	1033	2.1	101	48.1
new/old	341832	113065	3004.9	9426	3.1
officially/unofficially	6025	394	0.2	33	165.0
old/young	113065	83247	731.8	2704	3.7
optimistic/pessimistic	7123	1984	1.1	96	87.3
optimism/pessimism	5717	1163	0.5	91	182.0
prove/disprove	20968	258	0.4	35	87.5
permanent/temporary	10413	7878	6.4	351	54.8
poor/rich	34054	20999	55.6	2027	36.5
public/private	133056	61202	633.1	6741	10.6
publicly/privately	8108	6406	4.0	282	70.5
punishment/reward	6363	6152	3.0	38	12.7
quickly/slowly	25129	8958	17.5	83	4.7
recession/boom	22707	8678	15.3	334	21.8
right/wrong	125712	42376	414.2	2677	6.5
rightly/wrongly	4558	2681	1.0	182	182.0
rural/urban	8600	7923	5.3	515	97.2
small/large	86908	69219	467.7	2928	6.3
straight/gay	21672	9734	16.4	277	16.9
strength/weakness	19866	5971	9.2	441	47.9
success/failure	47816	24438	90.8	971	10.7
true/false	35357	10245	28.2	227	8.1
war/peace	81293	38258	241.8	2586	10.7
well/badly	178431	15772	218.8	712	3.3
win/lose	76372	27771	164.9	1125	6.8
TOTAL:	**2662409**	**955994**	**8441.8**	**55411**	
AVERAGE:					**6.6**

Table Three: Co-occurrence of Antonymous Pairs

Table Three shows that all antonymous pairs selected for study co-occur at a statistically significant rate, at least three times more often than would be expected by chance. Some pairs record an enormous Observed/Expected ratio, but this is often attributable to their low individual frequencies. For example, *feminine* and *masculine* only arise on about 1,000 occasions each in the entire corpus. Therefore, they can be expected to co-occur in only 0.1 sentence. In fact, they co-occur in 140 sentences, generating an Observed/Expected ratio of 1400. This is anomalous, but even pairs of words with relatively high individual frequencies (*female/male, cold/hot, poor/rich*, etc.) are able to record healthy co-occurrence figures. Indeed, according to this experiment, antonyms co-occur 6.6 times more often than would be expected by chance.

When Justeson & Katz conducted a similar test on the Deese antonyms, they found that antonyms co-occurred in the same sentence 8.6 times more often than chance would allow (1991: 142). Their results were based on a corpus much smaller than the one from which the above statistics are drawn and this inevitably affects their reliability. For example, Justeson & Katz calculate an Observed/Expected ratio of 19.2 for *happy/sad*, based on individual frequencies of 89 and 32 respectively and observed co-occurrence in just one sentence. My corpus yields an Observed/Expected ratio of 6.8 for *happy/sad*, based on individual frequencies of 28,217 and 9,420 respectively and observed co-occurrence in 140 sentences. It seems fair to conclude that statistics derived from the latter corpus will be more trustworthy. However, despite difference in corpus size, the two average Observed/Expected rates (6.6 and 8.6), are close enough to prove that antonyms do co-occur in text at a relatively high rate.

5. New Antonyms

A further question which may be considered with the help of corpus data is: how and why do certain pairs of words become enshrined as antonyms? To address this issue, the productivity of

frameworks associated with some of the new classes of antonymy will be investigated. Productivity here refers to the "statistical readiness" (Renouf & Baayen, 1996) of lexico-syntactic constructions to incorporate other related terms. In other words, if antonyms occupy certain lexical environments in text, which other words also occupy those environment and could some of those words be seen as new, developing antonyms?

Most new classes of antonymy were found to favour certain lexical environments in text. Three lexico-syntactic frameworks will be used to assess the antonymous profiles of given words:

both X and Y	[Co-ordinated Antonymy]
between X and Y	[Distinguished Antonymy]
whether X or Y	[Co-ordinated Antonymy]

The productivity of these frameworks will be tested by placing a word in the X-position, extracting all concordances which feature that word-string from the corpus, then examining which items occupy the Y-position. Three words will be placed in X-position for each framework. I shall begin by investigating an antonym from my sample, *good*. If these frameworks are in any sense productive, intuition demands that they should retrieve *bad* and, to a lesser extent, *evil* in Y-position with high frequency. I shall then examine two new words (*natural* and *style*) to discover whether this strategy can be used to identify potential antonyms for words which do not have established antonyms.

5.1. Seed Word: *good*

- ***both good and ...***

The lexical word-string *both good and* appears in a total of 63 corpus sentences. In 45 of those 63 sentences, it is followed immediately by *bad*. A further 4 of the 63 sentences recorded *evil* appearing immediately after *both good and*. This leaves 14 occurrences of *both good and* which are followed by neither *bad* nor *evil*. These are listed below, together with the noun-head they modify:

- both good and flawed (King Hassan's reign)
- both good and pathetic (years)
- both good and wicked (people)
- both good and nasty (youths)
- both good and hard (times)
- both good and not green (God)
- both good and true (a story)
- both good and lasting (friends)
- both good and powerful (patriotism)
- both good and new (a paper)
- both good and friendly (a service)
- both good and inimical to the Labour Party (Conservative belief)
- both good and non-sexually explicit (a novel)
- both good and great (wines)

The concordances above make interesting reading. Some Y-position words are very useful contrast terms for *good* (*flawed* and *pathetic* are synonymous with *bad*; *wicked* and *nasty* are synonymous with *evil*). One would not intuitively identify *hard* as an antonym of *good*, but this contrast is perfectly valid within its given context - *hard times* are quite the opposite of *good times*. However, the phrase *both X and Y* does not always reflect an obvious contrast. For example, a story is described as being *both good and true*; one would not want to consider these terms as potential antonyms. In such contexts, it would appear that the framework signals unlikely inclusiveness. Similarly, no contrast is generated between *good* and either *powerful*, *new*, or *friendly*. Finally, *both good and great* is an interesting example because a distinction is made, but that distinction is not at the usual point on the scale on quality (i.e. between *good* and *bad*). Rather, this context distinguishes between *good* and something better than *good*.

- ***between good and ...***

The framework *between good and Y* occurs in 140 corpus sentences, more than double the number of *both good and Y*. However, the distribution of *bad* and *evil* is very different. Of the 140 examples of *between good and Y*, 50 feature *bad* in Y-position and 78 feature *evil* in Y-position. Only 12 sentences feature neither *bad* nor *evil* in Y-position. These contexts are listed below, together with their corresponding noun-head, where appropriate:

- between good and poor (schools)
- between good and poor (performance)
- between good and lousy (comprehensives)
- between good and harmful (foods)
- between good and greed (a struggle within Lewis)

- between good and suspicious (toadstools)
- between good and good to soft (the going)
- between good and very good
- between good and very good
- between good and excellent (Melbourne's eateries)
- between good and really great (wine)
- between good and the best

Once again, some of the occurrences at the lower end of the frequency scale are valid contrast terms and others are not. Two contexts show *poor* occupying Y-position in the *between good and Y*. This is an excellent contrast term, as is *lousy*. Equally interesting are the distinctions made between *good* and other, more extreme points on the scale of quality. On two occasions, *very good* is contrasted with *good*, and *excellent*, *really great* and *the best* each appear in opposition on one occasion. This is reminiscent of the *both good and great* word-string retrieved earlier. Although one would expect *good* to contrast exclusively with negative items in language, it would seem that many writers choose to exploit its latent contrast with "super-positive" terms instead.

- ***whether good or ...***

Of the three lexico-syntactic frameworks analysed, *whether X or Y* is the least common. In the corpus, only 8 sentences feature the word-string *whether good or Y*. In 7 of those sentences, *bad* fills the Y-position; in the eighth, *evil* fills the Y-position.

5.2. Summary of *good*

This analysis of *good* has demonstrated that it is possible to retrieve contrast words from the corpus using productive lexico-syntactic frameworks. Collectively, the three frameworks examined occur on 211 occasions in the corpus. On 102 of those occasions (48.3%), the given word-string is followed by *bad*. This is compatible with our intuitions - one could predict that *bad* would be set up in opposition against *good* most commonly. Indeed, one could also predict that *evil* would be runner-up; *evil* fills the Y-position in frameworks analysed on 83 occasions (39.3%). However, the purpose of this experiment is not to prove that *bad* and *evil* are antonymous with *good*; rather, it is to show that the three frameworks identified are fertile enough to be deemed productive. This seems indisputable.

5.3. Seed Word: *natural*

- ***both natural and ...***

- both natural and accurate (their response to the camera)
- both natural and artificial (light)
- both natural and artificial (lighting)
- both natural and artificial (light)
- both natural and artificial (the essence of man)
- both natural and artificial (everything that exists)
- both natural and assisted (fertility)
- both natural and beneficial (high altitude)
- both natural and coloured (light)
- both natural and heraldic (devices)
- both natural and human (perturbations)
- both natural and inevitable (process)
- both natural and inevitable (that ...)
- both natural and lucid (her acting)
- both natural and man-made (components)
- both natural and man-made (beauty)
- both natural and man-made (beauty)
- both natural and man-made (disasters)
- both natural and man-made (facilities)
- both natural and man-made (polymers)
- both natural and market (forces)
- both natural and prudent (paying debts)
- both natural and safe (white sugar)
- both natural and sensible (idea)
- both natural and social (sciences)
- both natural and social (sciences)
- both natural and spiritual (creatures)
- both natural and superb (a history of vodka)
- both natural and synthetic (fibres)
- both natural and taboo (a child's sexuality)
- both natural and technical (the effect)
- both natural and violent (causes)
- both natural and vital (USA action)

The output generated by a search for *both natural and* comprises 33 concordances, all of which are listed above. It can be seen that *both natural and* occurs less frequently in text than *both good and* , but that the Y-position output is more diverse. However, this is not to say that no patterns emerge: of

the 33 occurrences of this lexico-syntactic framework, 6 are followed by *man-made* and 5 are followed by *artificial*. Both of these terms make excellent contrast words for *natural*. Some of the words retrieved in Y-position on one occasion only are non-contrastive, but interesting and valid oppositions of *natural* include *market* (in terms of forces), *synthetic* (in terms of fibres), *violent* (in terms of death) and *assisted* (in terms of fertility).

- ***between natural and ...***

  - between natural and artificial (ozone)
  - between natural and artificial (worlds)
  - between natural and artificial (worlds)
  - between natural and created (forms)
  - between natural and cultivated (areas)
  - between natural and juridical (persons)
  - between natural and man-made (assets)
  - between natural and metal (packaging)
  - between natural and moral (evil)
  - between natural and supernatural

Ten corpus sentences feature the phrase *between natural and Y*. The only word to occupy Y-position more than once is *artificial*. It is interesting to note that *between natural and man-made* also appears. This suggests that *man-made* also shares a strong contrastive profile with *natural*. Significantly, it also confirms that similar words occupy the Y-position in the frameworks *both natural and Y* and *between natural and Y*. All of the six words retrieved on one occasion reflect contrast to a lesser degree, with *supernatural* perhaps being the most interesting because it ties in with *spiritual*, which was picked up by *both natural and Y*

- ***whether natural or ...***

  - whether natural or artificial (hormones)
  - whether natural or electric (light)
  - whether natural or imposed (punishment)
  - whether natural or man-made (environment)
  - whether natural or man-made (beauty)
  - whether natural or otherwise (phenomena)
  - whether natural or step (parents)
  - whether natural or through external intervention (chemical changes)

Eight contexts were found to include the word-string *whether X or Y*. Pleasingly, both *artificial* and *man-made* arise in Y-position. This means that all three lexico-syntactic frameworks have successfully retrieved both of these words. One-off contrast terms again include valid and useful examples. For example, *step* is not the kind of word one would intuitively identify as a potential opposite of *natural*. However, within the given context of parentage, this contrast is not only legitimate but very interesting. It is also reassuring to note the appearance of *otherwise* in Y-position. Though this term is not a valid opposite of *natural* in itself, *otherwise* effectively functions as a proform for unspecified contrast words in text.

5.4. Summary of *natural*

Collectively, the three frameworks examined as part of this study feature *natural* in X-position in a total of 51 sentences. In 9 of those sentences, *artificial* occupies Y-position and, in a further 9 of those sentences, *man-made* occupies Y-position. This strongly suggests that those two words are the primary textual contrast terms of *natural*. Indeed, over one third of all frameworks examined feature either *artificial* or *man-made* in opposition with *natural*.

This output is particularly interesting if analysed in light of the range of antonyms which lexicographers have paired intuitively with *natural*. For example, Webster's Dictionary of Synonyms (1951) lists three antonyms: *artificial*, *adventitious* and *unnatural*. The inclusion of the first-mentioned of this trio is supported by this experiment, but *adventitious* does not occupy Y-position at all. This is not surprising given that the word occurs only seven times in the entire corpus (or about once per 40 million words in text).

More notable is the non-appearance of *unnatural* in textual opposition with *natural*. This could be interpreted as a flaw in the retrieval strategy or it could be interpreted as revealing an interesting aspect of natural: namely, that it prefers to contrast with lexical opposites rather than its morphological opposite.

Collins Cobuild Dictionary[5] (1985) cites six antonyms of *natural*, beginning with *unnatural*. The other contrast words suggested are *surprising* (not retrieved in text), *contrived* (not retrieved), *artificial* (retrieved 9 times), *man-made* (retrieved 9 times), and *processed* (not retrieved). Chambers Dictionary of Synonyms and Antonyms (1989) suggests *unnatural, artificial, man-made, affected* and *contrived*.

Therefore, some correlation emerges between intuitively identified antonyms and antonyms identified by productive lexico-syntactic frameworks: all three dictionaries cite *artificial* as a good opposite and only the oldest of the three fails to cite *man-made*. However, I would suggest that other recommended antonyms (*adventitious, processed* and even *unnatural*) are not placed in textual opposition against *natural* as often as may have been anticipated. Moreover, it could be argued that such words are less valid contrast terms of *natural* than *synthetic, supernatural, assisted* and other Y-position words which have not been cited by lexicographers, but which have been retrieved in this experiment.

5.5. Seed Word: *style*

- ***both style and ...***

  - both style and a demonstration of reaching speed
  - both style and achievement
  - both style and commercial space
  - both style and content (x4)
  - both style and date
  - both style and emotion
  - both style and fashion
  - both style and feeling
  - both style and heart
  - both style and history
  - both style and performance
  - both style and personality
  - both style and personnel
  - both style and policy
  - both style and prices
  - both style and qualifications
  - both style and reputation
  - both style and standards
  - both style and substance (x5)

The word-string *both style and* arises in 26 corpus sentences. Two words are set up in opposition to *style* more commonly than anything else - *substance* appears in Y-position five times and *content* appears in Y-position four times. These contexts reflect a trend for *style* to be seen as meaningless or superficial, and licenses its opposition with more "weighty" terms. Other words which reflect this trend but only occur once include *performance, policy, achievement* and *standards.* From examining its antonymous profile, one might infer that *style* has developed a pejorative sense in the language.

- ***between style and ...***

  - between style and content (x4)
  - between style and disorder
  - between style and grape
  - between style and political ideology
  - between style and quality (x2)
  - between style and subject
  - between style and substance (x5)

Four of the 15 *between style and* constructions are followed by *content* and five are followed by *substance*. This means that only 6 of the 15 occurrences of this lexico-syntactic framework feature other terms. On two occasions, this term is *quality*, which conforms to the underlying trend for *style* to be treated negatively in text and be synonymous with emptiness or absence of quality. However, a reminder that these frameworks are not exclusively inhabited by contrast words is provided by *grape*. The sentence from which this word-string is taken actually explores the relationship between the style of a given wine and the nature of the grape used in its production. Of course, *grape* could hardly be seen as a valid or useful opposite of *style*, merely an unlikely instantial collocation.

[5] This dictionary made use of a corpus which was smaller than my own (about 20 million words), but which was not newspaper specific.

- ***whether style or ...***

This lexico-syntactic framework was not found in the corpus at all, probably because *style* functions most commonly as a noun and nouns do not lend themselves readily to this construction.

5.6. Summary of *style*

The textual profile of *style* shows that *substance* is most commonly retrieved in Y-position (10 hits; 24.4%). In second place is *content* (8 hits; 19.5%). Between them, these two words are retrieved in Y-position in 43.9% of frameworks. It is interesting to note that *style* is never seen to contrast with concepts such as *inelegance* or *tastelessness*, as Chamber's Dictionary of Synonyms and Antonyms (1989) suggests.

5.7. Analysis of output

The aim of this exercise was to identify where new antonyms come from. The process by which "opposites" are created is complex, but it is reasonable to speculate that in order for a pair of words to become enshrined as antonyms in any language, they must first receive a significant amount of exposure. This exposure will be in contexts which are more often associated with established pairs of antonyms. Based on this rationale, *man-made* and *artificial* have been identified as potential antonyms of *natural;* and *substance* and *content* as potential antonyms of *style*. This output may be seen as initial evidence that it may be possible to automatically identify embryonic antonyms in text.
Arguably the most dramatic antonymous formation of recent times involved *gay* and *straight*, which held no obvious semantic relation in the middle of the last century, but have now achieved clear antonymous status. The increasing attention given to different sexual preferences must have contributed to the establishment of *gay* and *straight* as new "opposites", though this antonymity was surely enshrined by repeated co-occurrence in lexical environments similar to those examined here.

6. Conclusions

This paper has demonstrated that corpus-based approaches are relevant to an investigation of antonymy. Based on evidence from broadsheet newspaper corpora, I have argued that:

- In addition to logical distinctions, antonymous pairs are also receptive to classification according to their textual function. Data show that the two most common text-based classes of antonymy are Co-ordinated Antonymy (in which antonyms are joined by *and* or *or* and express exhaustiveness or inclusiveness) and Ancillary Antonymy (in which antonyms act as a lexical signal of a further, nearby contrast).
- All 56 antonymous pairs examined co-occur intra-sententially at least three times more often than chance would allow. On average, antonyms record an Observed/Expected ratio of 6.6.
- Using lexico-syntactic frameworks associated with the co-occurrence of established antonymous pairs, it is possible to identify new textual oppositions. Such research may shed light on the process by which a pair of words achieve antonymous status in language and allow us to identify new antonyms in their infancy.

Bibliography

Carter R 1987 *Vocabulary*. London, Allen & Unwin.
Chambers Dictionary of Synonyms and Antonyms 1989. Cambridge, Chambers.
Collins Cobuild English Dictionary 1995. London, HarperCollins.
Cruse DA 1986 *Lexical Semantics*. Cambridge, Cambridge University Press.
Deese J 1964 The Associative Structure of Some Common English Adjectives. *Journal of Verbal Learning and Verbal Behaviour*, 3: pp 347-357.
Jackson H 1988 *Words and their Meaning*. Cambridge, Cambridge University Press.
Justeson JS, Katz SM 1991 Redefining Antonymy: the textual structure of a semantic relation. *Literary and Linguistic Computing* 7: pp 176-184.
Kempson RU 1977 *Semantic Theory*. Cambridge, Cambridge University Press.
Leech G 1974 *Semantics*. Middlesex, Penguin.
Lyons J 1977 *Semantics: Volume 2*. Cambridge, Cambridge University Press.
Mettinger A 1994 *Aspects of Semantic Opposition in English*. Oxford, Oxford University Press.
Muehleisen V 1997 *Antonymy and Semantic Range in English*. Unpublished PhD dissertation, Northwestern University.
Palmer FR 1972 *Semantics*. Cambridge, Cambridge University Press.

Renouf A, Baayen H 1996 Chronicling the Times: Productive Lexical Innovations in an English Newspaper. *Language*, volume 72, number 1.
Roget's Thesaurus 1952. London, Sphere Books.
Webster's Dictionary of Synonyms 1951. Menasha, Merriam.