A corpus-linguistic analysis of English *-ic* vs *-ical* adjectives

Stefan Th. Gries*
University of Southern Denmark at Sønderborg

1 Introduction

A very interesting and productive phenomenon of English word-formation is that of *-ic* and *-ical* adjectives such as *economic*/*economical*.[1] This group of adjectives poses several problems to theoretical linguists on the one hand and applied linguists as well as language teachers on the other hand:

1. It turns out to be very difficult to detect any pattern governing the distribution of suffixes: when does an adjective end in *-ic* only (cf *acrobatic*/**acrobatical*) and when does it end in *-ical* only (**zoologic*/*zoological*)?
2. There are cases where one adjective root takes both suffixes (*electric(al)*, *historic(al)* etc), which raise further, even more complex questions:
   a. are the two forms that constitute a pair synonymous? Put differently, to what degree are the adjective forms differentiated today? – if they are not synonymous,
   b. does each suffix contribute some constant meaning component according to which the adjectives constituting a pair can be reliably distinguished or is there some other possibility to distinguish between the different adjective forms?

One look at only a very limited number of standard reference works shows that, unfortunately, unequivocal answers to these questions do not seem to exist. The questions that I would like to address (though not fully answer) in this study are 2a) and 2b).

The paper is organised as follows. Section 2 will summarise previous studies that have concerned themselves with these questions and will (i) point out many frequently-occurring difficulties in some detail and (ii) relate these difficulties to methodological shortcomings that nearly all studies share. Section 3 will then suggest a different strategy to overcome many of these problems. Since the par-

ticular strategy to be proposed has not been used often in lexicography, sections 3.1 and 3.2 briefly introduce its theoretical foundations, which are then brought together in section 3.3. Section 4 is concerned with the practical application. Section 4.1 demonstrates how this method can be fruitfully applied to question 2a), while section 4.2 is devoted to a brief demonstration of how extended techniques can be used to tackle the issue raised in 2b). Section 5 provides some additional results. Section 6 concludes the study.

2 Previous analyses

The question of semantic differentiation between the two adjectives of such a pair has been repeatedly addressed throughout the last two centuries. Most analyses proceed in two steps, roughly corresponding to questions 2a) and 2b). That is to say, they start out from discussing several (typically well-known and frequent) adjective pairs, such as *classic(al), economic(al), electric(al), historic(al), politic(al)* etc, in terms of their semantics (and, sometimes, in terms of other grammatical properties). In a second step, the analyses are extended by also focussing on the issue of whether there is some common component of meaning that the two suffixes add to the meaning of the adjective root to which they are attached. This review will proceed in the same way. Section 2.1 will provide a brief overview of common classifications of frequently discussed adjective pairs. Section 2.2, then, will be devoted to presenting proposed semantic and distributional generalisations, and section 2.3 will point out a variety of difficulties in previous analyses.

2.1 Particular adjectives

-ic/*-ical* adjectives have been investigated by many dictionary makers, grammarians, language teachers and (applied) linguists. It is neither interesting nor possible (given lack of space) to discuss all of the individual claims in great detail. I will therefore only provide an overview of results from a small, though representative, sample of references. For brevity's sake, I will do so in tabular form. Consider Table 1.[2]

Table 1: Findings and claims on *-ic*/*-ical* adjectives

Adjective	Semantic feature	Reference(s)
politic	artful, crafty, prudent, sagacious, wise, scheming, sensible (given the circumstances), well-adapted (to a particular purpose)	OJ, QGLS, CEDT, OED, CoCD, CCED, NJR
	crafty and unscrupulous, shrewd, cunning (sinister)	CEDT, OED, CoCD
	political, constitutional (archaic)	CEDT, OED
political	of, relating to, dealing with or pertaining to politics and/or the science of government/state/ administration	OJ, CEDT, OED, CoCD, CCED, NJR
	policy-making as distinguished from administration, law, military	CEDT, OED
	relating to the way power is achieved and used	CCED
economic	of, relating to or concerned with economics and finance	HWF, HM, QGLS, CEDT, OED, CoCD, CCED, NJR, MK
	concerning or affecting the organization of material resources, industry, money and trade	CEDT, OED, CoCD, CCED
	pertaining to the management of a household or private affairs	OED
	relating to services, businesses etc, producing a profit by being produced or operated	CEDT, OED, CoCD, CCED
	inexpensive, cheap	CEDT
	not resulting in money being lost	CoCD
	practical, utilitarian	CEDT, OED
	a variant of *economical*	CEDT
economi-cal	thrifty, money-saving, frugal	HM, HH, CEDT, OED, NJR, MK

	using the minimum required, not wasteful, spending money sensibly, not requiring much resources etc, cheap to operate/use, associated with economy	HWF, QGLS, CEDT, CoCD, CCED
	pertaining to pecuniary position	OED
	a variant of *economic* (some senses)	CEDT, OED
historic	famous, important, memorable in history, with a history, makes history	HWF, OJ, HM, QGLS, CEDT, OED, CoCD, CCED, NJR
	to be part of history (as opposed to fiction or legend)	HH, OED, CoCD
	with a history	QGLS
	of verb tenses used for the narration of past events	HWF, CEDT, OED
	see also *historical*	OED, CoCD
historical	pertaining to or dealing with (the science/study of) (events in) history	OJ, HM, HH, QGLS, CEDT, OED, CoCD, CCED, NJR
	having existed (as opposed to fiction or legend)	CEDT, OED, CoCD, CCED
	of verb tenses used for the narration of past events	OED
	celebrated or noted in history (now *historic*)	OED
classic	exhibiting all expected characteristics, typical, representative	CEDT, CoCD, CCED, MK
	of high/first class, outstanding, serving as a model or standard or following standard principles	HWF, PHM, QGLS, CEDT, OED, CoCD, CCED, NJR, MK
	characterised by a simple, pure, traditional form and unaffected by changes of fashion, thus often of lasting significance	CEDT, OED, CoCD, CCED, NJR, MK
	(more widely) belonging to Greek/Roman antiquity	HWF, OED, MK

classical	of, relating to (properties of) Greek and Roman antiquity	HWF, PHM, QGLS, CEDT, OED, CoCD, CCED, NJR, MK
	exhibiting a (traditional and simple) form, style or content that is characterized by emotional restraint and conservatism	CEDT, CoCD, CCED
	of the first rank or authority, of lasting value	OED, CoCD, CCED
	constituting a standard, esp. in literature	CEDT, OED
	referring to classicism	MK
	orchestral (of music)	CoCD, MK
lyric	relating to or (genuinely) exhibiting the characteristics of lyric/poetry	HWF, PHM, CEDT, CoCD
	(poetry) written (in a simple and direct style) and expressing emotions	CEDT, CoCD, CCED
	having the form and manner of a song (also accompanying a lyre)	CEDT, OED
lyrical	suggestive and/or imitative of or resembling lyric verse	HWF, PHM, OED
	poetic, romantic, musical	OED, CoCD, CCED
	enthusiastic, effusive	CEDT
	(also) *lyric*	CEDT
magic	supernatural, of or relating to magic	HWF, HH, CEDT, OED, CoCD, CCED, NJR
	wonderful, exciting, enchanting, term of commendation	CEDT, OED, CoCD, CCED
	important in a particular situation	CoCD, CCED
	also *magical*	CEDT
magical	of or involving or pertaining to magic	HWF, OED, CCED
	resembling magic or as if by magic, amazing	HH, OED
	wonderful, exciting, enjoyable	CoCD, CCED, NJR
	cf *magic*	CEDT, CoCD

comic	of and/or intended as (artistic) comedy (aiming at humorous effect)	HWF, HH, QGLS, CEDT, CoCD, OED, CCED, NJR
	humorous, funny, laughable (whether intended or not)	CEDT, OED, CoCD, CCED
comical	causing laughter, having the effect of comedy (unintentionally)	QGLS, CEDT, OED, CoCD, CCED, NJR
	queer, strange, silly	OED, CCED
electric	powered by or working on or using electricity	QGLS, OED, CEDT, CoCD, CCED, NJR, MK
	producing, carrying, transmitting or supplying electricity	CEDT, OED, CoCD, CCED, MK
	the actual power, the thing itself	HM, HH
	exciting, emotionally charged	OED, CoCD, CCED, MK
electrical	of, relating to or concerned with electricity	HWF, OJ, QGLS, CEDT, OED, CCED, MK
	not the power itself, has to do with electric things	HM, HH
	working by, supplying or using electricity	OED, CoCD, CCED, MK
	a less direct, more general connection with electricity	CoCD, NJR, MK
	thrilling	OED
analytic	of, pertaining to, concerned with or in accordance with analysis	CEDT, OED
	consisting in, or distinguished by, the resolution of compounds into their elements	OED
	short for *psychoanalytic*	OED
	analytical (logical reasoning)	CEDT, OED, CoCD, CCED
	true or false by virtue of the meanings of words alone	CEDT

analytical	of or pertaining to analytics	OED
	pertaining to analysis and/or algebra	OED
	employing an analytic/logical method or process (eg in chemistry, logic or linguistics)	OED, CoCD, CCED
	analytic	CEDT, OED
	psychoanalytic	OED
logistic	= *logistical*	CoCD, CCED
	relating to the organization of something complicated, to logistics	OED, CoCD, CCED
	pertaining to reckoning, disputation or (mathematical) calculation or logic	OED
	logarithmic	OED
	no entry	CEDT
logistical	relating to the organization of something complicated, to logistics	OED, CoCD, CCED
	pertaining to reckoning, disputation or (mathematical) calculation, *logistic*	OED
	logarithmic	OED
	no entry (cf *logistics* or *logistic*)	CEDT, CoCD, CCED
geometric	relating to or following from the principles of geometry	HWF, CEDT, CCED, NJR
	consisting of or formed by (regular) circles, lines curves etc	CEDT, OED, CoCD, CCED
	or *geometrical*	CEDT, OED, CCED
geometri-cal	relating to or following from the principles of geometry	HWF, CEDT, OED, CoCD, CCED, NJR
	consisting of or formed by (regular) circles, lines, curves etc	CEDT, CCED
	or *geometric*	CEDT, OED, CoCD, CCED

numeric	'cf numerical' or '= numerical'	CEDT, OED
	no entry	CoCD, CCED
numerical	of (the nature of), relating to or written as numbers/figures	CEDT, OED, CoCD, CCED
symmetric	= *symmetrical*	OED
	of a binary relation: such that when two terms for which it is true are interchanged, it remains true	OED
	no entry	CEDT, CoCD, CCED
symmetrical	possessing or displaying symmetry (due to regular organisation)	CEDT, OED
	mathematically constant in spite of changes of variables	OED
	having two halves which are exactly the same, except that one half is the mirror image of the other	OED, CoCD, CCED
	when entities are equally distributed about a dividing line, plane, or point so that they are at equal distances on opposite sides of those	OED
graphic	clear, detailed, vividly descriptive (of descriptions of negative things)	CEDT, OED, CoCD, CCED
	concerned with or resembling drawing, writing and/or graphs/curves	CEDT, OED, CoCD, CCED
graphical	*no entry*	CoCD
	pertaining to writing	CEDT, OED
	using graphs	CEDT, CCED
	= *graphic*	CEDT, OED
problematic	full of problems, complicated, difficult to answer, uncertain, doubtful	HH, CEDT, OED, CoCD, CCED
	possible, but not necessarily true	OED

proble-matical	perhaps without answer, but posed for, eg, discussion	HH
	involving or of the nature of a problem, disputable, doubtful	OED, CEDT
	(*no entry*), = *problematic*	OED, CoCD, CCED

Note that the order of adjectives in Table 1 is not random – rather, the vertical position of an adjective pair (only approximately) reflects the degree to which studies have considered the two adjectives to be clearly distinguishable (extending upon Marsden 1985:31). In the top rows of Table 1, we find the adjective pairs that are easiest to distinguish while lower rows represent cases of decreasing distinguishability.

In a laudable attempt to investigate the degree of semantic differentiation more thoroughly and from a different methodological perspective, Marsden (1985) conducted a forced-choice selection test with native speakers of English at English institutions of higher education: 'Respondents were requested to insert the appropriate form of the given adjective in typical noun phrases of the kind used as illustrative material by Fowler and the GCE. Thus, for instance, the slot in "___ languages" was to be filled with "*classic*" or "*classical*"' (1985:31). On the whole, Marsden (1985:32) found that 'the elicited usage concurs with Fowler as regards both the overall sequence of the items and the variation in degree of differentiation. […] Thus the general picture that emerges confirms the relevance of the formal definitions to contemporary usage'. Before discussing Table 1 in slightly more detail, let us now look at proposed generalisations governing the suffixes' distribution.

2.2 Abstracting away from particular adjectives

Given the scale of semantic differentiation found in Table 1, it comes as no surprise that both general opinions concerning generalisability and specific proposals differ strongly. On the one hand, we find researchers who are very pessimistic as to whether there are any discernible patterns since, apart from some frequent adjectives having clearly different meanings, many other cases are more problematic(al?). According to Fowler (1926:249), the choice of one adjective form over the other is often immaterial. Similarly pessimistic is Snell (1972:57): 'If and when similarly formed adjectives end in *-ic* or *-ical* cannot be determined by rules'. With respect to many other (though not all) adjectives, this attitude is echoed by Ross (1998:41–43).

On the other hand, some scholars have attempted to formulate several tendencies (as opposed to watertight rules). Referring to Maxwell (personal communication), Jespersen (1942:391) suggested that 'the forms in *ic* may indicate either the quality or the category of thing, but that those in *-ical* always, or almost always, indicate the quality only […] I dare say this is no more than a tendency, but I think it exists'. However, this is too vague and self-contradictory (*always, or almost always* vs *no more than a tendency*) to be put to a serious test.

Marchand (1969) proposed a few interrelated factors as influencing the distinction between *-ic* and *-ical* adjectives. He argued that *-ic* adjectives derive from the 'basic substantive', whereas *-ical* adjectives in turn derive from *-ic* adjectives. Thus, by some form of analogy that is (unfortunately) not explicitly motivated, the meaning of *-ic* adjectives is notionally more directly connected to the idea expressed by the root than the meaning of *-ical* adjectives (1969:242). For instance, Marchand (1969:242) attempted to buttress this argument by stating that '[a] sound is *metallic*, as it is like metal'. The same proposal was put forward by Hawkes (1976:95): 'the adjective in *-ic*, derived from the root substantive, has a semantically more direct connexion with that root idea; the adjective in *-ical*, a derivative of itself from an adjective form, has a looser connexion with the root idea and often takes on a a [sic] correspondingly looser meaning'. Similar suggestions are made in contemporary dictionaries. According to the OED (sv *-ical*), the form in *-ic* is 'often restricted to the sense 'of' or 'of the nature of' the subject in question', while that in *-ical* 'has wider or more transferred senses, including that of 'practically connected' or 'dealing with' the subject'. However, it is also pointed out that 'in many cases this distinction is, from the nature of the subject, difficult to maintain, or entirely inappreciable'. Similarly, according to CEDT (sv *-ical*), *-ical* is 'a variant of *-ic*, but [has] a less literal application than corresponding adjectives ending in *-ic*', but no example is discussed.

Another distinction introduced by Marchand (1969) is that scientific terms end in *-ic* more often (cf also Fournier 1993) since the scholar is more interested in the inherent quality of things than the layman. In addition, words in wider common use tend to end in *-ical* (1969:242), a suggestion that ties in with the purported specialised scientific use of *-ic* adjectives but is unfortunately not supported by any empirical evidence.

Ross (1998:42) argued that a variety of adjective pairs 'follow a similar pattern with the *-ic* form being more specific, the *-ical* form more general', a proposal that could perhaps be related to Marchand's and Hawkes's 'direct-indirect' distinction mentioned above.

Marsden (1985:30) also suggested several interrelated dimensions according to which the adjectives can be distinguished: 'intrinsic/neutral-value judgement' and 'genuine-resembling/imitation'. Still though, he pointed out that '[w]hat makes usage *appear* unsystematic is the fact that the morphology is at variance with the semantics: sometimes it is the shorter form (*economic*) which has the 'unmarked' function, sometimes that selfsame function is taken over by the longer form (*historical*)'.

Finally, Kaunisto (1999:347), apparently unaware of a similar though less general claim by Marsden (1985:29), suggested that, if an *-ic*/*-ical* adjective is preceded by a prefix, then the *-ic* suffix should be more frequent (in order to keep forms shorter). However, no empirical evidence is offered to support this claim. Table 2 summarises the proposed distinctions:

Table 2: Findings and claims on *-ic*/*-ical* adjectives

-ic	***-ical***
quality and category	quality
direct connection to root substantive	less direct connection to root substantive (wider senses)
specific	less specific / more general
genuine	resembling / imitation
positive	less positive or negative
scientific terms	wider common use
prefixed forms	

2.3 Theoretical and empirical problems of previous studies

Unfortunately, the above findings and claims based upon them do not all hold up to scrutiny. Of course, we find that some adjectives are clearly and unanimously distinguished by virtually all scholars (cf eg *politic(al)* and *economic(al)*). However, once we move down along the continuum represented in Table 1, we find less conformity both within a single source and across different sources. Consider, for instance, *magic(al)*. The semantic description by Fowler implies that there is a clear difference but is somewhat unhelpful since it does not allow for a principled differentiation. The OED's entry for *magic* supports Fowler's defini-

tion (adding the meaning of 'exciting'), but the OED's entry for *magical* does not. However, the adjectives are supposed to be virtually identical in meaning. The situation is made even more complicated by the claims of Hawkes (1976) and Ross (1998), whose definitions of *magic* are comparable to those of the previous studies, but whose definitions of *magical* introduce the meaning component 'exciting' that the OED has attributed to *magic*. If we then turn to a very recent and corpus-based dictionary, CoCD, it becomes still more confusing. On the one hand, the dictionary includes *magic(al)* in a special list of adjective pairs where the two adjectives are claimed to exhibit a 'difference in meaning or use', but on the other hand, it states 'You use magic in front of a noun to indicate that an object or utterance does things or appears to do things by magic', 'Magical can be used with a similar meaning', and 'Magic and magical can also be used to say that something is wonderful and exciting' (CoCD sv *magic-magical*). In other words, the two adjectives are not so different in meaning after all, and the meaning component of 'exciting' etc, which has recently been attributed to *magic* by some and to *magical* by others, is now, for the first time, attributed to both. Similarly confusing results can be obtained with other adjectives from the above list (and in other dictionaries not quoted above), and as is obvious from Table 1, in some other cases researchers admit not to be able to discern any consistencies in meaning and/or usage of the two adjectives.

Also, not all dictionaries seem to apply their own criteria consistently. As mentioned above, the OED and CEDT consider *-ical* a less literal variant of *-ic*, but, for instance, the extra usage entry for *classic(al)* in the latter reference does not relate to the proposed 'literal-less literal' dimension:

> The adjectives *classic* and *classical* can often be treated as synonyms, but there are two contexts in which they should be carefully distinguished. *Classic* is applied to that which is of the first rank, esp. in art and literature [...] *Classical* is used to refer to Greek and Roman culture. (sv *classic* usage)

A further peculiarity of the entries is that the division of senses provided in some sources seems somewhat unprincipled or arbitrary.[3] Consider, for instance, the entry for *political* in the OED, consisting of the following subsenses:

(1) a. Of, belonging, or pertaining to the state or body of citizens, its government and policy, esp. in civil and secular affairs; public, civil; of or pertaining to the science or art of politics.

b. Of persons: Engaged in civil administration; civil, as distinct from military; spec. in India, having, as a government official, the function of advising the ruler of a Native State on political matters, as political agent, resident, etc (now Hist.).

(2) Having an organized government or polity. †Said also of animals such as bees and ants (obs.).

(3) Relating to, concerned or dealing with politics or the science of government.

(4) Belonging to or taking a side in politics or in connexion with the party system of government; in a bad sense, partisan, factious. Also (freq. in derogatory use), serving the ends of (party) politics; having regard or consideration for the interests of politics rather than questions of principle.

(5) = politic A. 2. Obs.

It is sometimes difficult to recognise on what basis the decision to have different subentries (for what at times appear to be cases of hyponymy) was made (cf eg senses 1a, 3 and 4).

Finally, this completely confusing situation is not even improved by Marsden's experiment. Given the diversity of opinions and complexity of patterns found on the basis of literature and dictionary data, I am the first to welcome additional methods of analysis. Unfortunately, however, I believe that (i) his report of the test leaves open so many questions and (ii), from what we are told, Marsden's test is flawed in so many respects that, on methodological grounds alone, he has not contributed to the issue. For a start, we do not know how many subjects participated in the test, making it difficult to generalise from the results. Second, the description of the test quoted above suggests that the subjects were teachers and, thus, were not linguistically naïve and/or possessed some knowledge of prescriptive grammar. Therefore, the experimental results are probably biased by this knowledge, especially since Marsden does not mention any measures taken to rule out such effects. That is to say, it is highly unlikely that the experiment does indeed tap into 'contemporary usage', as he claims that it does. In a similar vein, the experimental design (forced-choice selection) and the lack of (mention of?) filler items and randomisation lead me to expect that the subjects could immediately guess what the test was about, making it even more likely that they access conscious prescriptive knowledge rather than truly usage-

based information. Finally, the results obtained are not subjected to any of the standard statistical tests, making it impossible to take any of the results at face value.[4]

Thus, if one intends to examine the contemporary usage of particular adjectives, the use of corpus data is a much more reliable way to pursue. In this respect, Kaunisto's (1999, 2001) work is both methodologically and conceptually superior to all preceding analyses: it relies on corpus data, thereby including many more examples than can normally be analysed and ruling out any (unconscious) bias on the part of the investigator. One minor shortcoming, which Kaunisto is always aware of, is that his corpus does not include data from different registers; another drawback is that Kaunisto does not resort to contemporary corpus-linguistic techniques which might add to the clarity and generalisability of his results. This latter point will be addressed in more detail below.

As to the question of a predictable component of meaning consistently added by a suffix, on the basis of the available data few conclusions appear warranted. Before we turn to the individual distinctions introduced, let me briefly mention two ways in which previous analyses can be shown to be inadequate or incomplete. First, the proposed distinction may be found to work only for a limited number of cases, rendering it useless for the majority of cases. Second, the proposed correlation between the suffixes and their meaning contribution may be found to work in another or even the opposite direction. It has already been mentioned in earlier studies that, given the source of data (mainly dictionaries and literature), previous conclusions are not always borne out by data from authentic usage (cf eg Ross 1998:43).

Let us start with the frequently purported tendency of *-ic* adjectives having a more direct relation to the meaning denoted by the root than *-ical* adjectives. Note that this idea is extremely difficult to operationalise objectively in the first place. It receives prima facie support by pairs such as *historic(al)* and *electric(al)* as discussed by, for instance, Ross: *historic* ('not only generally related to history, but also important') can be argued to be specific/direct, whereas *historical* ('generally related to history') is more general and less direct. Similarly, *electric* is used with basic level terms and subordinate terms (eg *kettles, sunroof, toothbrush,* etc), whereas *electrical* is used with general superordinate terms (eg *appliances, equipment,* etc).

However, this distinction is a paradigm case where the two possible inadequacies mentioned above can be observed. First, we can easily see that the 'direct-less direct relation' distinction is far from applying to all adjectives: eg with adjective pairs which are unanimously considered synonymous (eg *geometric(al)* or *problematic(al)*). This is probably why the OED itself expresses

doubts as to the validity of this distinction. Note in passing a certain degree of terminological fuzziness that is not explicitly commented on: the examples are concerned with the 'specific-general' dimension on two different levels – with *historic(al)*, the 'specific-general' distinction is applied on the semantic plane of linguistic description – with *electric(al)*, the same dimension is applied to the level of collocates.[5] Second, there is a variety of problems where the distinction does not hold up to scrutiny. For instance, we find that, as Ross (1998:42, quoting Crystal 1984), and Kaunisto (1999:345) point out, *economic* also seems to be used recently in the meaning of 'money-saving', undermining the proposed distinction. Similarly problematic is *electric(al):* with one exception only, it is generally acknowledged that, contrary to the prediction, it is *electric* rather than *electrical* that also has a 'less literal' meaning, namely that of 'excited'. In this connection, it is also interesting to consider Ross's treatment of *economic(al)*. He states that '[b]oth *economic* and *economical* relate to finance, but *economic* is strictly related to the world of economics […], while *economical* is used in the wider sense of not wasting money' (1998:42). However, if one followed Ross's treatment of *historic(al),* one could also argue exactly the other way round (cf also Marsden 1985:28): *economic* would then be basic (simply meaning 'related to economics'), whereas *economical* is more specific ('not only generally related to economics, but also in a particular way, namely money-saving'). True, the argument as such does not falsify the distinction as a whole, but it indicates that more precise formulations or operationalisations are required to decide on its validity. Finally, as to Marchand's above-mentioned treatment of *metallic,* I must admit I simply fail to see the purported 'direct' connection between sound and metal.

Let us now look at the preference for scientific adjectives to end in *-ic* mentioned by Marchand (1969), namely the general tendency for recently coined adjectives to end in *-ic* rather than in *-ical*. I do not wish to argue against the tendency as such; I only doubt that it can function as support for Marchand's claim, because this pattern can be explained more simply/more parsimoniously. Since the development of science and technology is a relatively recent phenomenon, the correlation between *-ic* adjectives and scientific terms might as well be explainable in terms of a parallelism of linguistic and technological development (cf also Marsden 1985:29). Also, Marchand does not seem to be convinced of the purported tendency of *-ical* adjectives to be in more common use, since (i) he himself points to various counterexamples and (ii) this claim could only be supported by frequency data anyway, to which neither he nor any other analyst has referred (cf, however, below section 5).

Similar problems are encountered with the distinction of 'neutral-value judgement'. Basically, the same two problems arise. On the one hand, many (if not most) adjective pairs do not exhibit a difference along this dimension (eg *egoistic(al)*, *electric(al)*, *geometric(al)*, *magic(al),* etc). On the other hand, the distinction is not uniformly valid. With *economic(al)*, it is generally argued that *economic* is neutral, simply meaning that something belongs to the domain of economics, whereas *economical* is generally taken to imply a positive value judgement ('money-saving'). Unfortunately, however, Marsden himself points out that, with *historic(al)*, it is, if anything, the other way round: *historical* refers to something as being related to history whereas *historic* communicates, as it were, a positive judgement ('important (enough to be remembered)').

What about 'genuine-resembling adjective' senses? There are at least no examples directly contradicting the proposed tendency. Still though, apart from the few examples, such as *comic(al)*, *lyric(al)* and *magic(al)*, that support the proposed distinction, there are many cases to which the distinction does not seem to apply at all (eg *geometric(al)*, *historic(al)*, *symmetric(al)*, etc), although it could be applied in principle and would make sense (eg *psychic(al)*, *cyclic(al)*, etc).

If we look at all of the proposed generalisations, we must conclude that:

- they often apply only to a limited set of adjectives (while not applying to other adjectives where the distinction would also make sense);
- they are in some cases contradicted by the data;
- they are in some other cases accompanied by caveats, counterexamples or doubts by the researchers proposing them in the first place.

Thus, *-ic*/*-ical* adjectives leave open a variety of questions, most of which can probably not be answered by the traditionally prevailing methods of literature and dictionary research. Elaborating upon the first laudable steps by Kaunisto, who also advocated further collocational studies for other adjectives (1999:345), I would like to make the point that more recent corpus-based techniques going beyond the simple examination of all collocates should be brought to bear on the questions listed above. The following section outlines one such proposal.[6]

3 A corpus-linguistic approach

3.1 A model of similarity: Tversky (1977)

An area that triggered a lot of psychological research in the 1970s is that of theories of categorisation, similarity and prototypes. One particular subpart of this research focussed on the description and development of models of how to measure similarity and how to embed the notion of similarity into, for instance, prototype-based theories of concepts. A particularly influential model is that of Tversky (1977), which I will illustrate briefly in what follows.

Tversky's starting point was a critique of so-called geometric models of similarity, ie models where the similarity of two entities E_1 and E_2 was typically represented by the metric distance between these entities in an n-dimensional space. Among the points of critique mentioned by Tversky, one is particularly relevant to our present purposes, namely that these models presuppose a symmetric approach towards the similarity of concepts. More precisely, from the fact that similarity between E_1 and E_2 is represented as a metric distance, it follows that E_1 should be as similar to E_2 as E_2 is to E_1. This, however, is contradicted by empirical findings (cf eg Tversky 1977:334), which is why a satisfactory model of similarity must be able to handle such asymmetries. In Tversky's contrast model, the asymmetric similarity of E_1 to E_2 is a function of:

- the number of features that are common to both E_1 and E_2 ($E_1 \cap E_2$);
- the number of features that belong to E_1 but not to E_2 (E_1-E_2);
- the number of features that belong to E_2 but not to E_1 (E_2-E_1) (cf Tversky 1977:330).

Similarity increases with the addition of (possible differentially weighted) common features and/or deletion of distinctive features. Consider the block letters E, F and I as an example. Tversky (1977:330) argues: 'E should be more similar to F than to I because E and F have more common features than E and I. Furthermore, I should be more similar to F than to E because I and F have fewer distinctive features than I and E'. On this basis, Tversky (1977:332f) defined the similarity scales S in terms of which the similarity of E_1 and E_2 is measured as follows:

(6) $S(a, b) = \theta f(E_1 \cap E_2) - \alpha f(E_1 - E_2) - \beta f(E_2 - E_1)$, for some $\theta, \alpha, \beta \geq 0$ where f is an interval scale reflecting the contribution of a feature to the similarity[7] and θ, α, β are parameters that can be used to express the direction of contrast and weight of the kinds of features involved.

The latter point is of special importance. It means that if $\alpha=\beta$, ie if the focus of the similarity assessment is equally on E_1 and E_2 (which could be paraphrased as 'Assess the degree to which E_1 and E_2 are similar to each other'), then the similarity between two entities E_1 and E_2 is symmetric. On the other hand, if $\alpha>\beta$, ie if the focus of the similarity assessment is more on E_1 (which could be paraphrased as 'Assess the degree to which E_1 is similar to E_2'), then the similarity between E_1 and E_2 is asymmetric.

This approach to similarity lends itself very well to an investigation of similarity of word usage – however, before I demonstrate how, let us turn to the second theoretical basis of my analysis, ie Biber (1993).

3.2 The identification of word meanings: Biber (1993)

Biber (1993) has introduced a by now classic technique to identify the different senses of polysemous words, such as *right* or *certain*. This technique works as follows.

In early corpus linguistics, it has already been recognised that words differ with respect to the company they keep, ie their collocates. On the basis of a few more recent influential studies, it could be shown that pairs of functionally synonymous words (or words that at first sight appear to be exchangeable in a variety of contexts) can be distinguished on the basis of their significant collocates, even if the patterns observed cannot be easily characterised.[8] Church et al (1991:119ff), for instance, have demonstrated how the semantically similar adjectives *strong* and *powerful* differ markedly with respect to their significant collocates, where the significance of each collocational pattern was determined by the statistic of mutual information, an information-theoretical measure of collocational strength and similarity. That is to say, the meaning of words is definable and distinguishable in terms of their (significant) collocates.

On the basis of the idea of (significant) collocates, Biber (1993) has determined the different meanings of eg *right*. To that end, he first determined R1 collocates of *right* occurring more than 30 times and their absolute frequencies in many reasonably large corpus files.[9] The results were entered into a spreadsheet, each cell of which listed the frequency of a significant collocate of *right* in a particular corpus data file. This data set was then entered into a principal component analysis (PCA), and, as a result, the PCA has established groups of significant collocates, such that each group can be interpreted as reflecting basic semantic properties of one meaning of *right*. As to his findings for *right,* consider Table 3:

Table 3: Senses of *right* as established by Biber (1993)

Sense 1:	Sense 2:	Sense 3:	Sense: 4
opposite of *left*	'immediately', 'directly', 'exactly'	'ok', 'correct'	stylistically marked sense (clause-final)

As Biber (1993:537) points out, the 'analyses produced unanticipated but systematic results, indicating that this approach can provide a useful complementary perspective to traditional lexicographic methods'. As possible extensions, he proposes (i) to use measures of collocational strength (eg mutual information) to identify the set of possibly relevant collocations and (ii) to tag the corpus to use grammatical category information.

3.3 Synthesis: Estimation of Significant Collocate Overlap (ESCO)

By now it should have become clear what I am about to do, namely combine the lessons of sections 3.1 and 3.2. If (i) word meanings can be differentiated on the basis of significant collocates and if (ii) we, thus, interpret a significant collocate of a word as one of its features (namely one indicating the presence or absence of the collocate),[10] then we can determine the degree of semantic similarity of one word to another one on the basis of:

- the number of significant collocates (ie features) that both $word_1$ and $word_2$ exhibit;
- the number of significant collocates exhibited by $word_1$, but not $word_2$
- the number of significant collocates exhibited by $word_2$, but not $word_1$.

That is to say, the semantic similarity of $word_1$ to $word_2$ increases with the number of significant collocates they share and decreases with the number of significant collocates they do not share. Still though, there is more to be done since, even if we have these numbers of significant collocates shared and not shared – what do we do with them? It is important to avoid the mistake of simply throwing them together into, say, a single multiplicative index because, once we do that, our measure of similarity is again symmetric. The solution to this problem and the resulting way of analysis will be explained in the following section together with other technical particulars.

Before applying this analysis, I would like to anticipate an objection that might be raised by sceptical readers. The objection is: while it is possible to use

significant collocates for the differentiation of the meanings of polysemous individual words such as *right* or *certain*, it is not possible to use the same technique for comparing two (or more) words. This is so because the fact that both *blue* and *expensive* might have *car* as a significant R1 collocate does not render their meaning similar at all: *blue* and *expensive* mean something completely different, and, therefore, the whole approach is bound to fail.

Admittedly, this objection has some intuitive appeal – at a second glance, however, it does not pose too much of a problem for two reasons. Firstly, note that the present approach (just like Biber's technique, on which it is based) does not attempt to formulate a definition of the meaning of a word on the basis of its significant collocates; it uses significant collocate overlap as a measure of similarity of linguistic usage. Secondly, and much more importantly, the objection misses an important point of the technique, namely the inclusion of significant collocates *not* shared by the two words to be compared. Even if *blue* and *expensive* share a significant collocate such as *car* (or even a few more), the number of collocates they do not share is even larger. Thus, given the inclusion of the non-shared significant collocates as following from Tversky's approach, it is impossible that two words so different in meaning as *blue* and *expensive* accidentally result in being synonymous just because they happen to share a few collocates.

In this connection, one might raise the question of whether meaning/semantics and collocational behaviour are in fact two different aspects of a word's behaviour, a question also brought up by Kaunisto (1999:349). Given his way of analysis, however, he implicitly seems to assume that there is a close enough relation between the two to investigate the former in terms of the latter. I will adopt the same opinion, following Firth's notion of collocational meaning, common word sense disambiguation methods (cf Kilgariff 1997 and the references cited therein) and Church et al's (1994) *sub*-test.

4 Practical application: **-ic** *vs* **-ical**

4.1 Estimating the degree of semantic differentiation

In order to apply ESCO to the question of how synonymous adjectives ending in *-ic* and *-ical* are, we first need to obtain a representative and register-diverse sample of such adjectives. To that end, I performed a search in all files from the written part of the British National Corpus (version 1) amounting to 3,209 files with about 90m words (oral data contain too few examples of these adjectives). I then determined the adjectives that occur most often with *-ic* and *-ical*; these are listed in Table 4:

Table 4: Absolute and relative frequencies of the most frequent *-ic*/*-ical* adjectives in the written part of the BNC

Stem	**analyt-**	**class-**	**com-**	**econom-**	**electr-**	**geometr-**	**graph-**	**histor-**	**logist-**	**lyr-**	**mag-**	**numer-**	**polit-**	**problemat-**	**symmetr-**	**total**
-ic	219	1,613	493	22,937	690	583	630	2,022	142	104	912	215	31	734	352	31,184
-ical	781	3,174	128	472	457	190	629	5,329	93	258	835	684	29,630	126	323	42,981

Of each of these 30 adjectives, all those R1 collocates were identified that occurred at least two times.[11] Then, the significance of the co-occurrence had to be determined.[12] For such purposes, a variety of measures of collocational strength is available: the *t*-test (Church et al 1991), the *z*-score (cf Berry-Rogghe 1974), mutual information (cf Church and Hanks 1990), the Chi-square test (cf Manning and Schütze 2000), Fisher's exact test (cf Weeber, Vos and Baayen 2000), etc. However, a study of the relevant literature indicates that most of these are problematic to some degree. I believe that Dunning's (1993) log-likelihood ratio (-2logλ) suits our purposes best since it (i) does not rest on any particular distributional assumptions (eg normality) and (ii) can handle sparse data very well. Thus, I calculated -2logλ and Chi-square for each bigram adjective and its R1 collocate and sorted them according to the size of -2logλ. Going down from the highest score of -2logλ, I counted all collocations as significant until the first Chi-square value not exceeding 6.63 (the threshold value for p=.01 with df=1) anymore.[13] The reason for this is that -2logλ does not have an inbuilt standard threshold value for significance, and, in order to test conservatively (ie to make sure H_0 is not rejected too early), this procedure excludes collocations from the first item where the Chi-square test, which itself overestimates significance of infrequent collocations easily, begins to produce the first non-significant result.[14] Having obtained all significant collocates of each adjective, I determined the number of significant collocates that the two adjectives of a pair have in common. Now, however, we face the problem mentioned above, namely how to avoid a symmetric measure of the similarity between the adjectives (which is why traditional measures of similarity, such as Jaccard coefficient, Dice coefficient, etc cannot be used).

I suggest to use a two-dimensional diagram, which I will call $ESCO_2$. In this kind of diagram (Figure 1), each axis represents the percentage of significant collocates of one word that are also shared by the other. The result is a two-dimensional coordinate plane in which a dot's location indicates the degree to which each adjective of a pair is similar to the other pair member in terms of collocational behaviour. This diagram can accommodate cases where the relation of similarity between two words is not symmetric. Note that this is not only a purely academic distinction following from Tversky's model, since such asymmetries are at times even reported in standard dictionaries: we have seen several cases where the meaning of one adjective is defined by reference to the other adjective, but not vice versa.[15] The design of the diagram follows from both general psychological considerations and linguistic behaviour and is thus an adequate technique to represent our findings:

- dots representing adjectives exhibiting little or no overlap are in the lower left part of the diagram;
- dots for adjectives exhibiting much symmetric overlap are in the upper right part of the diagram;
- the degree to which the relation between the adjectives' collocational behaviour is asymmetric will be reflected in the values' distance from the main diagonal.

Consider now Figure 1, representing the results of the corpus analysis described above; for reasons of exposition, the axes are logarithmically scaled. Before turning turn to the results, let me explain how the dots in this diagram came into existence on the basis of one example, namely *symmetric(al)*. According to the corpus analysis of *symmetric(al)*, *symmetric* has 36 significant R1 collocates, five of which we also find to be significant collocates of *symmetrical*. *Symmetrical*, by contrast, has 18 significant collocates, of which it shares the already identified five collocates with *symmetric*. That is, 13.89 percent (5 out of 36) of the significant collocates of *symmetric* are also significant collocates of *symmetrical,* while 27.78 percent (5 out of 18) of the significant collocates of *symmetrical* are also significant collocates of *symmetric*, resulting in the dot at (13.89; 27.78).

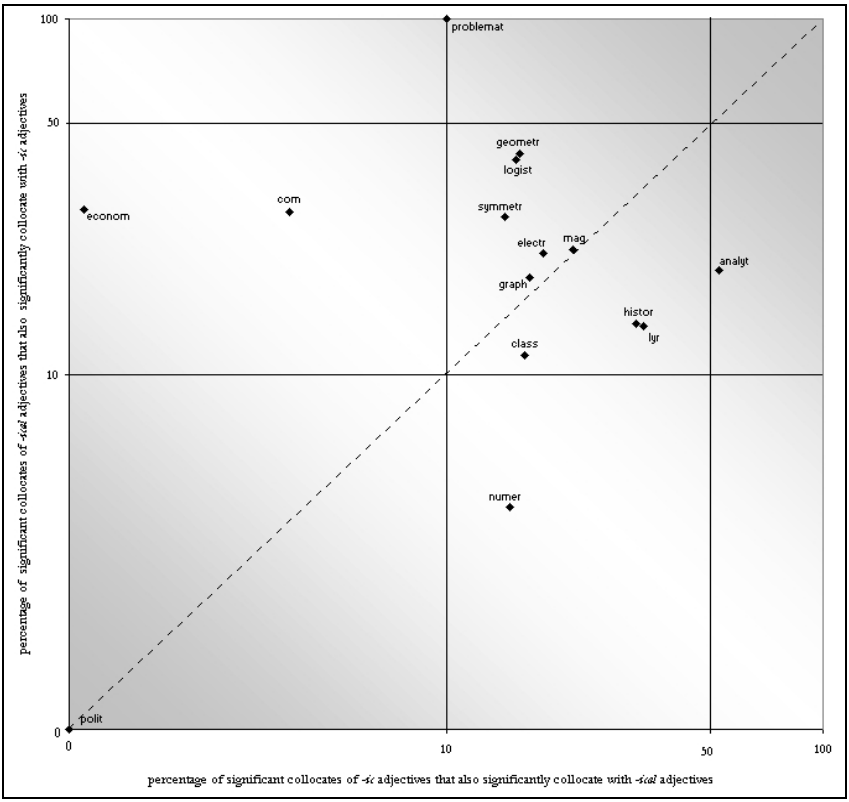

Figure 1: ESCO$_2$ for frequent adjectives ending in -ic *and* -ical *(excluding function words)*[16]

Several observations can be made as a result of this analysis (while at the same time explaining the technique's results more comprehensively). On a very general level, we learn that the adjective pairs behave very heterogeneously: one cluster of nine adjective pairs (those in the section delimited by the lines for 10% and 50% on each axis) with moderate values can be distinguished from the six remaining, more extreme cases. These remaining cases also make up two groups: on the one hand, we have extreme cases such as *politic(al)*, where ESCO$_2$ supports the result of previous studies (namely complete differentiation); on the other hand, we have cases like *problematic(al)*, *analytic(al)* and

economic(al), where the analysis shows, again in conformity with some previous results, that the two adjectives of a pair exhibit considerable, if not complete, overlap.

But let us look at some of the results concerning individual adjective pairs in more detail. In order to assess the validity of this technique, I will first discuss the $ESCO_2$ results for some adjectives in relation to previous findings; then, I will discuss $ESCO_2$ findings going beyond previous analyses.

As to the first point, $ESCO_2$ is supported in several ways. Take, for instance, *politic(al)*; this adjective pair is one where all authors agree that the two adjectives are as clearly differentiated as possible. This is also reflected in the $ESCO_2$ analysis since the two adjectives do not share a single collocate, thereby exhibiting maximal differentiation. In fact, it is interesting to point out that, while *political* is, like most *-ic*/*-ical* adjectives most often used attributively, the most significant collocates of *politic* are function words, namely *to*, *for* and *not*.[17] Thus, the two adjectives differ with respect to both meaning and syntactic distribution, and the analysis, although restricted to R1 collocates, could identify this difference straightforwardly (cf note 11).

As to *analytic(al)*, the results of $ESCO_2$ correspond to previous findings in two important respects: first, they show that previous studies were right in claiming that the two adjectives are fairly similar to each other (since we find considerable collocational overlap); second, they corroborate the practice of the CoCD and CCED, where *analytical* serves as the base of the comparison defining *analytic* (since $ESCO_2$ shows that *analytic* can frequently be subsumed under *analytical*). How exactly do I arrive at this judgement? *Analytic* is more similar to *analytical* than vice versa, because the ratio of shared significant collocates to all significant collocates of *analytic* (53.13%) is much larger than the ratio of shared significant collocates to all significant collocates of *analytical* (19.54%).[18]

Let us now turn to *economic(al)*. While it is one of the most widely quoted cases of an adjective pair with clear semantic differences, recall the observations by Crystal (1984) (quoted in Ross 1998) and the OED quoted above that *economic* nowadays also tends to be used in the sense of 'money-saving', which has traditionally only been associated with *economical*. Again, these observations seem to be confirmed by the data. In terms of the above analysis, we would expect that the dot representing *economic(al)* rises vertically in the course of time (irrespective of its position on the horizontal axis) since that would represent that collocates of *economical* are also used with *economic*. If we look at the data, we indeed find that:

- *economic* and *economical* are not completely different (which would have resulted in a dot at 0, 0), but share some significant collocates (*processes, reform, repair* etc) which can be interpreted in either way;
- there is a moderate ratio of significant collocates of *economical* that are also significant collocates of *economic* (as opposed to a small ratio of significant collocates of *economic* that are also used with *economical*).

Thus, one might speculate that, since *economical* is much less frequent than *economic* anyway, *economic* seems to take over this sense, which might in the long run result in, as Fowler (1926:250) put it, the 'clearing away' of *economical*. This would also tie in with the generally observed tendency of the recent superiority of the *-ic* forms.

Finally, with *problematic(al)*, the result obtained is a particularly extreme one, but one that is also supportive of $ESCO_2$: according to CEDT and CoCD, both adjectives are virtually synonymous, and *problematical* is defined via *problematic*. This is exactly the result obtained by $ESCO_2$: (i) the high degree of collocational overlap in the data reflects the high degree of semantic overlap as postulated in the dictionaries, and (ii) the fact that all significant collocates of *problematical* are also used with *problematic* but not vice versa reflects the fact that *problematical* is defined via *problematic*. In sum, the results of the $ESCO_2$ analysis strikingly correspond to those of previous analyses, and they do so for both cases with no or relatively little differentiation (*problematic(al)* and *analytic(al)*) and cases with clear or extreme differentiation (*economic(al)* and *politic(al))*.

Now that $ESCO_2$ has been validated with reference to fairly clear-cut cases, let us briefly turn to some other cases. Reasons of space do not permit comprehensive analyses of all adjective pairs included here, but I will point to how the present analysis and its extensions to be introduced below in section 4.2 enable insightful observations. Let us start with a case that can be accounted for straightforwardly, namely *logistic(al)*. Interestingly, not all dictionaries have entries for both adjectives and, among those that do, some dictionaries list only one meaning for this adjective pair. The only laudable exception in my list of references in this respect is the OED, which, however, treats the words as completely synonymous. $ESCO_2$, on the other hand, reveals that these works are somewhat mistaken: contrary to CEDT, *logistic* exists (actually, it is more frequent than *logistical*!) and there is a moderate degree of overlap, but (i) the percentages of overlap do not exceed 40 per cent and (ii) the overlap is not as symmetric as the OED's entry led us to expect. Consider the set-theoretic diagram (following Tversky 1977:330) in Figure 2, where collocates are sorted according to their collocational strength ($-2\log\lambda$):

Figure 2: Significant collocate sets for logistic(al)[19]

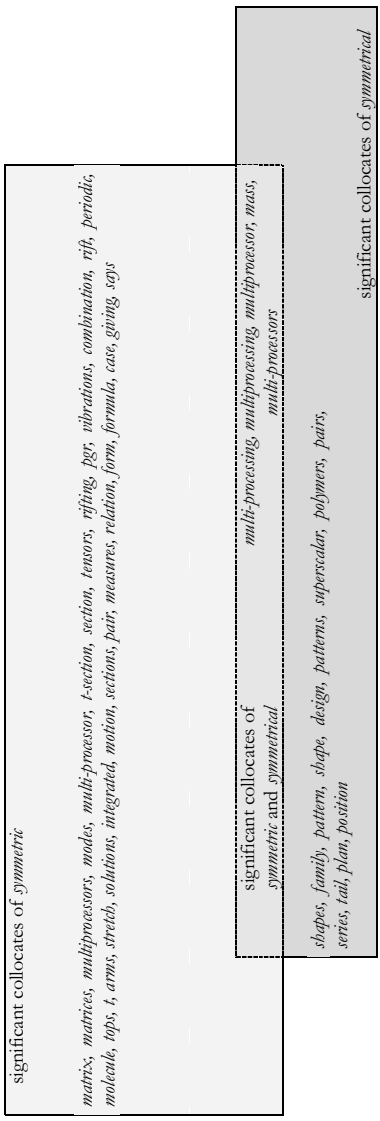

Figure 3: Significant collocate sets for symmetric(al)

Here, the number of significant collocates of each adjective and the similarity of *logistic* to *logistical* (and vice versa) are indicated by the proportionally corresponding sizes of the differently-shaded areas. For instance, the larger the overlap, the more the usage of one word can be subsumed under that of the other. In this example, one would conclude that, while the two adjectives were claimed to be virtually completely synonymous, they do exhibit a measurable and significant difference in usage. Interestingly, the results for *logistic(al)* are surprisingly clear: both meanings listed by the OED are present. The 'mathematical' reading is reflected in R1 collocates such as, eg, *regression*, *regressions*, *model*, *curve,* etc, whereas the 'organisation/transportation' reading manifests itself in R1 collocates such as *corps*, *units*, *support*, *problem,* etc. However, the authentic usage data show that *logistic* is indeed used in both senses (as claimed by the OED) – but *logistical* is exclusively used in the 'organisation/ transportation' sense. This is a case where $ESCO_2$ reveals usage patterns and ways of differentiation that contemporary dictionaries hitherto seem to have missed.

Let us now turn to a more challenging case by returning to *symmetric(al)*. *Symmetric(al)* is yet another case of an adjective pair where some dictionaries do not even have an entry for *symmetric* and, if both forms were listed, their meanings are considered to be virtually identical. The present analysis, however, demonstrates that this traditional treatment does again not do justice to the data: *symmetric* exists (in fact it also is more frequent than *symmetrical*; cf *logistic(al)* above), and the degree of significant collocational overlap is moderate. In other words, there must again be some differences that have hitherto not been discovered. Consider, therefore, Figure 3:

In this example, one would also conclude that there is a significant difference in usage. More importantly, however, *symmetrical* is much more similar to *symmetric* than vice versa. These findings clearly contradict some dictionaries' practice of either lacking an entry for *symmetric* (cf CoCD, CCED and CEDT) or defining *symmetric* by referring to *symmetrical*, where it apparently should be the other way round (cf the OED).

However, there is a somewhat problematic aspect of this result which threatens the validity of the proposed observations or the semantic analysis that might follow. When we try to distinguish between *symmetric* and *symmetrical* on the basis of their significant collocates, we can run into a problem. In this case, the five shared significant collocates in Figure 3 are among the eight most significant collocates of *symmetric*. Also, there are some word forms that are not shared significant collocates proper (since collocates were not lemmatised), but are nevertheless shared instances of (i) a significant lexeme or (ii) a spelling variant, namely *pair/pairs* and *multi-processor/multiprocessor*, respectively.

That is, paradoxically, it is exactly those (highly significant) collocates which can serve least to elucidate the differences we are interested in, since, in this case, they happen to be shared collocates (which do not differentiate by definition). For a more detailed semantic/lexicographic analysis, this problem needs to be addressed, which is why I will return to it below in section 4.2.

Finally, let us briefly look at *numeric(al)* and *magic(al)*. All dictionaries claim that both adjectives are synonymous (if both have entries in the first place – some do not have an entry for *numeric* as an adjective). On the one hand, the dictionaries' tendency to define (if at all) *numeric* with reference to *numerical* is clearly supported by the observed asymmetry of collocational overlap. On the other hand, the overlap is fairly modest, so a more thorough inspection of the significant collocates is necessary. In this case, however, this analysis would have to include 96 significant collocates for *numerical*, and this huge number makes it quite difficult to detect patterns. We find a similar situation for *magic(al)*. While we have seen that traditional analyses have resulted in a perplexing variety of accounts, the present analysis can shed at least some light on these two adjectives. First, the overlap is moderate and, thus, supports previous claims that the adjectives are not completely synonymous. However, the question of how their difference(s) can be explained seems very difficult since *magic* and *magical* have 92 and 89 significant collocates respectively, ie numbers of collocates that do not lend themselves to manual analysis easily. In the light of the last two adjective pairs, it would, therefore, be desirable to be able to filter out the relevant collocates even more rigorously. The following section will introduce a technique to achieve these two objectives, namely addressing the problems of shared significant collocates and large numbers of significant collocates.

4.2 Differentiating senses: a few brief case studies

The preceding section showed how the analysis of significant collocates contributes to detecting differences that have gone unnoticed in many, if not all, previous analyses. However, we have seen that, in some cases, the numbers of significant collocates as such are very high, which is why a more economical and elegant technique is desirable. More problematical, though, was the observation that significant collocates might not even be able to distinguish between the adjectives properly (recall the above example of *symmetric(al)*).

Addressing a similar problem, Church et al (1991:124f.) have argued in a by now classic study that MI and related measures of collocational strength are measures of similarity between words; what we need, however, is a measure of dissimilarity between words. More precisely, if we want to distinguish between,

say, *symmetric* and *symmetrical*, we should not look at those words which simply co-occur significantly with the two adjectives (the significant collocates as listed above); rather, we need to look at those words which significantly discriminate between *symmetric* and *symmetrical* (what I will call discriminating collocates). That is, we need those words which occur significantly more often with *symmetric* than with *symmetrical* (and vice versa). Obviously, the sets of significant collocates and discriminating collocates need not coincide.[20] As a more adequate measure for the dissimilarity between words, Church et al (1991) propose a variant of the *t*-test. While the application of the t-test to our question promises to be very useful in general, there are nevertheless three particularly noteworthy areas of application, namely:

- adjective pairs with great differences between significant collocates and separating collocates;
- adjectives where the number of shared significant collocates is high; or
- adjective pairs where a usage difference has not been realised so far such as *symmetric(al)* and *numeric(al)*.

Given constraints of space, I will demonstrate the application of the *t*-test to our problem only cursorily. Let me start by applying the *t*-test (with the Expected Likelihood Estimator and a threshold value of .05) to the data on *symmetric(al)*. This procedure has yielded the discriminating R1 collocates listed in Table 5.:

Table 5: Discriminating collocates of *symmetric(al)*

symmetric		***symmetrical***	
R1 collocate	**t; p**	**R1 collocate**	**t; p**
multi-processing	t=4.536; p<.001	*shapes*	t=–3.033; p=.001
multiprocessing	t=4.290; p<.001	*family*	t=–2.864; p=.002
matrix	t=3.157; p=.001	*about*	t=–2.761; p=.003
matrices	t=2.646; p=.004	*pattern*	t=–2.289; p=.011
modes	t=2.245; p=.012	*shape*	t=–2.064; p=.020
section	t=2.245; p=.012	*design*	t=–2.064; p=.020
multiprocessors	t=2.017; p=.022	*on*	t=–2.064; p=.020

mass	t=1.946; p=.026	*patterns*	t=–1.813; p=.035
multiprocessor	t=1.733; p=.042		

The first implication worth mentioning is that the result supports the criticism voiced above against simply using collocates or even significant collocates of *symmetric(al)*. Although *multiprocessor* (and its morphological and orthographic variants) co-occurred significantly with both adjectives, the *t*-test shows that these forms discriminate between the adjectives such that they are in fact significantly more typical of *symmetric* than of *symmetrical*.[21]

As to the difference(s) between the two adjectives, observe that, given the many discriminating collocates of *symmetrical* making reference to visual arrangements (*shape(s)*, *pattern(s)* and, perhaps, *design*), its meaning seems to be captured well in the CoCD and the OED (cf Table 1 above). However, the data show that *symmetric* and *symmetrical* are not completely synonymous in two respects. First, according to the discriminating collocates, not all of *symmetric*'s discriminating collocates exhibit the 'visual arrangement' part of *symmetrical*'s meaning: the collocate *matrix/matrices*, for instance, does,[22] but *symmetric multiprocessing* (and its variants) does not, since the latter refers to a computer architecture where (omitting the details):

- multiple CPUs of a single computer work in parallel (in peer-to-peer relationships) on individual processes;
- there is no master processor;
- all processors can equally access resources (eg memory, peripherals, graphics, other controllers, etc).[23]

Of course, this meaning of symmetric multiprocessing has some semantic commonality with the OED's definition of the purportedly synonymous *symmetrical*, but it lacks the 'visual arrangement' meaning component that both the OED and CoCD have considered central.

Second, there is another more subtle distinction. The discriminating collocates of *symmetric* refer to (mostly concrete) things that are symmetric themselves, eg a matrix (cf (7)). The discriminating collocates of *symmetrical* can also refer to the things themselves (cf (8)), but they can also refer to perceivable properties of these things (cf (9)), a distribution that might instantiate the 'direct-less direct' distinction proposed above.

(7) the symmetric matrix (ie the matrix is symmetric)

(8) the symmetrical shape (ie the shape is symmetrical)

(9) the matrix has a symmetrical shape

In this respect, note also that, whereas all discriminating collocates of *symmetric* are nouns (ie *symmetric* is used attributively), *symmetrical* has two discriminating collocates that are function words and, thus, hint at predicative usage. Finally, the discriminating collocates of *symmetric* seem to be infrequent technical terms, while those of *symmetrical* strike one as being much more frequent and of a less technical nature. This intuition is of course reminiscent of Marchand's proposal mentioned above that *-ical* forms are in wider common use. A *U*-test shows that the discriminating collocates of *symmetrical* are indeed significantly more frequent: U=8; z_{corr}=-2.69; p=.007.[24] We will return to this result below in section 5.

As pointed out above, the technique of discriminating collocates is also useful in cases where the number of significant collocates is high. Let us, therefore, have a brief look at the purportedly synonymous adjective pairs *magic(al)* and *numeric(al)*. (10) to (13) list the discriminating collocates (at the significance level of .05) for *magic*, *magical*, *numeric* and *numerical* respectively:

(10) *items, item, kingdom, wand, flute, word, words, box, standard, roundabout, show, carpet, music, formula, mushrooms, lantern, armour, circle, bullets, ring, sponge, sword, phase, night, potions, cards, standards, number, weapons, potion, bullet, fairy, lamp, circles, bus, art, age, for*

(11) *and, effect, powers, properties, mystery, as, field, practices, the, atmosphere, experience, quality, rites, illustrations, healing, philosophy, arts, knowledge, force, means, it, place*

(12) *up, variables, keypad, format, value, character, coprocessor, parameters, operators, quantity, but*

(13) *order, terms, identity, superiority, example, methods, strength, modelling, analysis, diversity, experiments, score, form, flexibility, solutions, sequence, ability, results*

As to *magic(al)*, the collocates reveal a clear tendency. The majority of the discriminating collocates in (10) denote concrete and perceivable/manipulable objects, whereas those in (11) overwhelmingly have abstract denotata. Note: the 'concrete-abstract' dimension observed for *magic(al)* is quite similar to the 'specific-general' dimension observed for *electric(al)*, but not completely identical. As mentioned above, *electric* has many specific and basic-level terms as collocates (which are, thus, mostly concrete things like the collocates of *magic*), while *electrical* has more general and superordinate terms as discriminating collocates, but these collocates are both abstract and concrete (unlike the collocates of *magical*). This is symbolised in Figure 4:

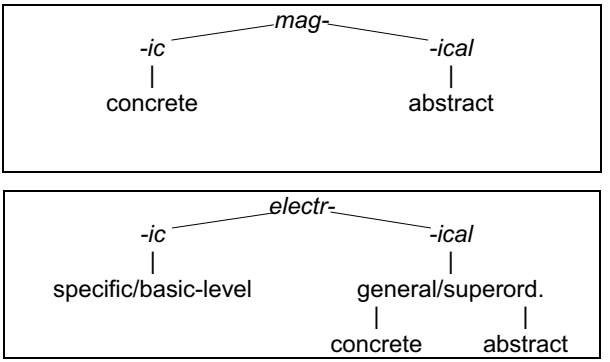

Figure 4: Properties of typical discriminating collocates of magic(al) *and* electric(al)

In addition, the results show that, with a single exception, R1 collocates of *magic* are nouns, pointing to attributive usage; on the other hand, *magical* has a few function words as discriminating R1 collocates, which (i) again points to predicative usage, supporting Fowler's claim about *magical*, and (ii) demonstrates that, for some reason, *magical* is often used in coordination with other adjectives.

A look at *numeric(al)* also provides some interesting information although the first conclusion we must draw is a sobering one: according to the list in (12) the most discriminating collocate of numeric is *up*. Unfortunately, closer inspection reveals two peculiarities of this finding. First, *numeric up* occurs only in one file, so this collocation might therefore be characteristic only of one author rather than some property of *numeric*.[25] Second, all these instances of *numeric up* are unfortunately not instances of *numeric* as an adjective. In fact, all 43 cases instantiate the pattern [The N identifier is a numeric up to six digits long],

so we are forced to conclude that numeric is a noun and, thus, incorrectly tagged.

But let us now turn to the other discriminating collocates. These also undermine the purported synonymy of *numeric* and *numerical* and provide further, though admittedly less direct, evidence for the 'concrete-abstract' distinction just proposed. For instance, the discriminating collocates of *numeric* in (12) include concrete and abstract but countable collocates, whereas those of numerical in (13) do not include a single concrete denotatum. While more detailed analysis of the collocates' patterns are beyond the scope of the present analysis, we have nevertheless seen that the technique of discriminating collocates is a useful way of detecting hitherto unnoticed regularities. Also, we have seen that the abstraction away from individual collocates to discriminating features (ie semantic and syntactic/distributional features) involves some complex features different from the indicator noun attributes Justeson and Katz (1995) observe for the five high-frequency adjectives they analyse. More comprehensive analyses are necessary to develop (i) more detailed accounts of *-ic*/*-ical* adjectives and (ii) more comprehensive lists of discriminating features (indicator attributes).

5 Additional results

The previous two sections have shown how contemporary corpus-linguistic techniques going beyond unfiltered collocate listing can help us to improve upon previous treatments of the degree of synonymy and usage overlap as well as the differentiation between forms. However, while these issues were at the heart of the current paper, it is worthwhile showing that other questions benefit from such a corpus-based approach as well. Recall two quantitative claims mentioned in the review of the literature that have not been tested empirically so far: Marchand's claim that *-ical* forms are in wider common use and Kaunisto's claim that prefixed adjective forms exhibit a preference for *-ic* suffixes. This section will investigate these claims.

Let us start with Marchand's claim. In a way, it is surprising to see that his claim has never been tested, since (i) with contemporary resources it can be tested fairly easily and (ii), as we have seen above, preliminary results support Marchand's intuition. For a more general test, I used the about 119,000 tokens mentioned in note 1. On the basis of the concordance listing all rightmost adjectives ending in *-ic* or *-ical*, I computed the arithmetic means of the frequencies of all adjectives with these suffixes. The average frequency of *-ic* and *-ical* forms in this corpus are 32.78 and 72.67 respectively; according to a *t*-test, this difference is significant: t_{Welch} (680)=2.02; p=.043.[26] Thus, although unsubstantiated at the time it was made, Marchand's claim is fully borne out.

Finally, as was mentioned above in passing, Kaunisto has argued that, given a tendency to favour economy, prefixed adjective forms should exhibit a preference for the *-ic* suffix, probably in order to yield shorter words. However, he did not investigate this issue. I decided to do so for two reasons: First, I wanted to see whether his claim is actually borne out by the data. Second, if it is, it would be interesting to notice how the tendency to have short words interacts with the non-prefixed words' meanings. For instance, Kaunisto (1999:347) has noted in passing (that is, without reporting the actual frequencies he obtained) that the prevalence of *prehistoric* over *prehistorical* seems to imply that 'the effect of economy as a tendency would override the semantic differentiation usually present in the unextended form'. To that end, I investigated the 15 adjective pairs analysed above. The procedure I adopted on the basis of *analytic(al)* was then carried out for all other adjectives, too.

First, I determined all the adjective forms where (i) *analytic(al)* was preceded by some linguistic material and (ii) followed by nothing else. Second, all these cases were distinguished in terms of whether whatever preceded the adjective had been added with or without a hyphen (eg *hydro-analytic* and *hydroanalytic*) in order to avoid any bias this distinction might introduce. Finally, I noted the frequencies of the types and tokens of *-ic* and *-ical* suffixes in non-hyphenated and hyphenated forms and tested their distribution for significance with an exact binomial test; the prior probabilities used in the binomial tests are the ratios of the two suffixes with each bare adjective in the corpus. For the result, consider Table 6, where the results column contains plusses/minuses (depending on whether the observed frequency for *-ic* forms is higher/lower than the expected one); the numbers of plusses/minuses indicate the significance level of the difference between the observed *-ic* frequencies and the expected frequencies. For example, the second row shows that 232 non-hyphenated forms ending in *-ic* (as opposed to 37 such forms in *-ical*) are highly significantly more frequent than one would expect, given the frequencies of *analytic* and *analytical* (219 and 781 respectively) in the corpus.

On the whole, Kaunisto's claim is clearly borne out by the data: when *-ic*/*-ical* adjectives are derived by preceding linguistic material, then *-ic* forms are with very few exceptions either significantly more frequent than would be expected or there is no significant distributional difference. In general, this holds for type and token frequencies as well as for hyphenated forms and non-hyphenated forms alike. The only exceptions to this pattern are *classic(al)* and *economic*, which is interesting since these are adjectives where the two forms are fairly clearly differentiated in meaning. That is, in these cases, the semantic differentiation between the adjectives cannot be overruled by the economy princi-

ple suggested by Kaunisto, which is what we would expect: if the two adjectives do not make much of a difference in terms of semantics, then there is no reason not to have economical derivational processes – if, on the other hand, the two adjectives are semantically very different, then using only the *-ic* suffix for reasons of economy might result in losing the original semantic distinction and, thus, jeopardising communication.

Table 6: The distribution of suffixes in complex *-ic*/*-ical* adjectives

root	frequency	non-hyphenated forms		result	hyphenated forms		result
		-ic	*-ical*		*-ic*	*-ical*	
analyt-	type ~	2	5	ns	8	8	+
	token ~	232	37	+++	13	11	+++
class-	type ~	3	4	ns	1	16	--
	token ~	3	108	---	1	258	---
com-	type ~	3	1	ns	3	1	ns
	token ~	8	1	ns	4	1	ns
econom-	type ~	9	1	ns	16	1	ns
	token ~	638	48	---	779	1	+++
electr-	type ~	12	1	+	20	4	+
	token ~	200	7	+++	160	6	+++
geometr-	type ~	2	-	ns	2	1	ns
	token ~	4	-	ns	2	1	ns
graph-	type ~	100	36	+++	30	14	+
	token ~	3,423	2,697	+++	74	27	+++
histor-	type ~	3	9	ns	4	23	ns
	token ~	365	57	+++	27	94	ns
logist-	type ~	6	1	ns	-	-	
	token ~	29	1	+++	-	-	

lyr-	type ~	-	1	ns	1	2	ns
	token ~	-	1	ns	1	2	ns
mag-	type ~	8	2	ns	-	4	ns
	token ~	9	3	ns	-	4	ns
numer-	type ~	1	1	ns	2	2	+++
	token ~	87	6	+++	14	2	+++
polit-	type ~	1	12	+	1	40	+
	token ~	8	139	+++	1	321	ns
prob-lemat-	type ~	1	1	ns	2	-	ns
	token ~	77	10	ns	7	-	ns
symmetr-	type ~	5	3	ns	6	2	ns
	token ~	205	112	+++	23	3	+++

6 Conclusion

Let me summarise what I believe to be the most important issues of this paper. First, I provided a comprehensive summary and critical discussion of, as far as I can see, all factors that were hitherto proposed; hopefully, this study can thus also serve as a starting point for future studies. On the basis of this review, three different methodological steps preceding the present study can be distinguished. First, in most previous (traditional) treatments, the proposed differentiation of senses ultimately rested on the inspection of usage contexts of the adjective forms. While the notion of frequency was included in the analysis, its importance was assessed only on an intuitive basis. Then, Marsden attempted to elucidate contemporary usage on the basis of a questionnaire study. Finally, a more advanced way of analysis became possible with the sort of corpus-based analysis of, say, Kaunisto, where frequencies could be determined objectively. However, I hope to have shown that the analysis of *-ic*/*-ical* adjectives requires a somewhat more advanced methodology in terms of data and their analysis. The corpus-based methodology pursued here has the advantage that its data are:

- more natural than those obtained from dictionaries and literature data as well as Marsden's survey since they were taken from language used authentically;
- more representative in the sense that (i) they do not only stem from 'native speakers of English at English institutions of higher education' (Marsden 1985:31), and (ii) they come from a carefully balanced corpus and are, unlike Kaunisto's data (cf above), not register-specific;
- gathered and evaluated objectively, namely to a large degree on a statistical basis;
- less likely to be distorted by prescriptive attitudes and experimental effects that might result from the forced-choice paradigm administered to the subjects (who, thus, must have been aware of the purpose of the experiment immediately).

What is more, two techniques were introduced to assess the degrees of similarity and the kinds of differentiation of the adjectives of a pair. More precisely, on the basis of Tversky's contrast model, Biber's recommendations concerning useful extensions of his work on *certain* and *right* and contemporary statistical corpus-linguistic techniques I demonstrated that:

- the significant collocates show that many pairs hitherto claimed to be synonymous do pattern very differently (eg *logistic(al)* and *symmetric(al)*) and require changes in contemporary dictionaries;
- the traditional collection of examples or the manual inspection of corpus data should be supported by the identification of discriminating collocates;
- the discriminating collocates techniques introduced make it possible to establish new distinctions that differentiate between adjectives (recall *magic(al)* and *numeric(al)*) which defy easy characterisation on terms of significant collocates.

In this respect, the logic behind the sub-test of Church et al (1994) was explicated and related to psychological findings, and it was shown (i) how the results concerning many word pairs can be summarised in a single diagram and (ii) in what way a semantic analysis can build upon the distributional results. Finally, the present study provided the first empirical tests of Marchand's 'frequency claim' and Kaunisto's 'prefixation claim'.

Given the methodological suggestions and findings, I hope this paper stimulates further research on this topic. For instance, I could, unfortunately, not address very many adjectives here, so I have to leave others for further and more

comprehensive study. Cases like *geometric(al)*, which all sources I have looked at consider 'totally interchangeable' (Ross 1998:43) but which exhibit only moderate overlap, promise to be interesting cases, for whose study I believe to have suggested a rewarding way of analysis. Also, there are some distributional curiosities awaiting explanation. For instance, what is the motivation, if any, for the unexpected frequency distribution in Table 7?

Table 7: Frequencies of *egoistic(al)* and *egotistic(al)* in the written part of the BNC

egoistic	*egoistical*
41	1
5	55
egotistic	*egotistical*

Finally, Kaunisto (2001) has brought some diachronic work to bear on *-ic*/*-ical* adjectives, which might be supplemented by, for instance, tracing the historical development of these adjectives by seeing how significant/discriminating collocates change over time. All these phenomena and more, which are relevant to linguists, teachers and dictionary makers alike, are truly worth further investigation.

Notes

* I thank Stefanie Wulff from the University of Hamburg for her help and several useful comments on an earlier draft of this paper. Naturally, I alone am responsible for any remaining shortcomings.

1. To give an idea of the productivity of this pattern and the distribution of *-ic*/*-ical* adjectives, I conducted a search for such adjectives in 779 files of the British National Corpus (version 1), ie all files from the corpus parts labelled H and K containing written English amounting to about 24.4m words. The search results yielded nearly 119,000 tokens, most of which consisted of simple adjectives such as *economic* – in others, the *-ic*/*-ical* adjective was the rightmost part of a more complex adjective (eg *socio-economic* or *financial-cum-economic*). Of all tokens, there were 1,879 rightmost adjectives ending in *-ic* only, 331 rightmost adjectives ending in *-ical*

only, and 268 rightmost adjectives (amounting to 38,678 tokens, yielding an average of 1,583 pm) ending in *-ic* or *-ical*.

2. For reasons of space, I abbreviated authors' names; the abbreviations are listed in the reference section. Also, not all works I refer to mention all adjective pairs provided in Table 1 or, if they mention them, they do not always address the issues with which we are concerned about here.
3. This is only indirectly relevant to the distinction between *-ic* and *-ical* adjectives, but indicative of problems lexicographers/linguists face in this context, cf eg Fillmore and Atkins (1994) or Sandra and Rice (1996).
4. For instance, the frequency of subjects choosing the prescriptively correct adjective in a frame such as '___ languages' could easily have been tested against the number of (randomly distributed) expected choices with a Chi-square test. Alternatively, since Marsden's claims crucially rely on the unequivocality of the subjects' answers, it would be helpful to report a coefficient of concordance (cf eg Carletta 1996) or a measure of consistency (cf eg Shipstone 1960), allowing the reader to assess to which degree the conformity of the subjects' responses exceeds chance expectations.
5. This difference of levels to which the dimension is applied is not problematic as such, but needs to be embedded in one's overall account properly. I will return to this issue (cf section 3.3).
6. The reader will notice some similarity between my analysis and the *sub*-test of Church et al (1994), of which I unfortunately only learned after my study had been completed. I will, however, point out the relevant differences and commonalities between the *sub*-test and my technique where appropriate and necessary. Also, the present analysis extends work by Justeson and Katz (1995) according to some of their proposals (1995:22).
7. A feature's contribution depends on its intensity, frequency, familiarity, good form and informational content (Tversky 1977:332, 342f.).
8. For example, Halliday (1966) remarked that, while we drink strong tea (rather than powerful tea), it is difficult to motivate the lexical patterns/idiosyncrasies we find.
9. The expression 'R1 collocates' refers to collocates at the first position of the right of the search word.
10. For a similar approach, cf Justeson and Katz's (1995) definition of 'word sense indicators'.
11. The question might be posed why only R1 collocates were used. The main reason is that, as most researchers have pointed out correctly, *ic*/*-ical* adjectives are nearly always used attributively. Thus, position R1 is most likely to be occupied by a modified noun which is, therefore, a good candidate for

the desired analysis (cf Justeson and Katz 1995:1f, 8, 21). Note also that the limit on the utility of a 'mere' R1 collocate analysis is less severe than one might expect since it is still possible to, for instance, distinguish attributive and predicative uses, etc.

12. Since the collocates were not lemmatised, different inflectional forms or spelling variants were counted individually in order to be able to differentiate between the adjectives more precisely. In some cases at least, this turned out to be the right decision since there were R1 collocates whose singular and plural forms were significantly associated with different adjective forms.
13. My analysis was restricted to significant collocates that occurred *more* often than expected. I left out those cases where -2logλ and Chi-square indicated significance, but where MI showed that the collocation is significant in occurring *less* often than expected.
14. This procedure may seem overly complicated and arbitrary. However, given the methodological criticism voiced above, I, of course, need to be especially careful to avoid similar shortcomings. Also, in some of the cases, the procedure adopted functions perfectly in the sense that, after a long list of steadily decreasing values of -2logλ for content words, it identified exactly those cases (and started to exclude them) where function words had intruded in the list of collocates. Thus, the combined measure of -2logλ and Chi-square seems to be an adequate technique for the identification of significant collocates.
15. Less technically, other researchers also commented on this phenomenon. For instance, Marsden (1985:29) showed that '[t]he OALD makes a distinction along Fowlerian lines, but with some overlap allowing the use of *lyrical* in the sense of *lyric* though not vice versa'. Also, Church et al's (1994) *sub*-test shows that the degree of substitutability is generally asymmetric. The argumentative difference between their account and mine is that their justification of asymmetry seems to be founded on their empirical results alone, whereas mine is an a priori commitment founded on general psychological considerations of similarity and categorisation *and* supported by the empirical results.
16. Figure 1 is based on significant collocates other than function words only, but the results are virtually identical to those including function words; both product-moment correlations between the percentages exceed .97.
17. Typical expressions were characterised by the following two patterns: [NP_{SUBJ} THINK/FIND it politic (for NP) (not) to V] and [it BE politic (for NP) (not) to V], where capitalisation means 'any form of that lemma' and parenthesised elements are optional.

18. The same result can be obtained more technically on the basis of Tversky's account summarised in (6) by comparing the amounts of non-shared collocates of *analytic* and *analytical* (what was above referred to as (E_1-E_2) and (E_2-E_1)). Comparing these, we find that *analytic* has fewer non-shared collocates than *analytical*, ie less of 'an identity on its own'; cf also Church et al (1994:171f).
19. I do not analyse all adjectives in this way, because it would be quite space-consuming to compare many adjectives' collocate overlap on the basis of such diagrams. Also, the sizes of these diagrams vary according to the absolute number of significant collocates, which makes it difficult to compare the diagrams of different adjective pairs to one another.
20. For example, Church et al (1991:128) demonstrate that, while *strong thunderstorms* is a significant collocation, their data do not support the statement that *strong thunderstorms* is more likely than *powerful thunderstorms*. Therefore, the significant collocation *strong thunderstorms* is not useful for the distinction between *strong* and *powerful*, or, alternatively, *strong thunderstorms* is not a discriminating collocation.
21. Similar results are obtained for other adjective pairs. Eg, *shape* is a significant collocate of both *geometric* and *geometrical*, but the *t*-test, for instance, indicates that *shape* is in fact a discriminating collocate for *geometric*.
22. A symmetric matrix is a special kind of square matrices; the most straightforward example is a so-called symmetric diagonal matrix, where every element other than those of the principal diagonal are zeroes.
23. The criteria for symmetric multi-processing were obtained from the following websites (last access: 1 September 2001): http://www.nswc.navy.mil/cosip/nov97/osa1197-1.shtml, http://sunsite.uakom.sk/sunworld online/swol-09-1999/swol-09-insidesolaris.html and http://www.intel.com/eBusiness/products/workstation/ processor/tools.htm.
24. I use the *U*-test because the compared frequencies are limited in number and not normally distributed. Note also that the significant collocates are not only more frequent, they are also found in significantly more corpus files ($U=9$; $z_{corr}=-2.6$; $p=.009$).
25. Even though the present corpus is the largest ever analysed with respect to *-ic*/*-ical* adjectives, there are some other examples of this kind where authors' idiosyncratic preferences distort the picture. Take, for instance, *electric(al)*. Of the many significant collocates of *electric*, a few occur only in one corpus file, eg *tramways* (eleven times at R1 of *electric*) or *media*

(seven times at R1 of *electric*). Similarly, though less extreme, *electrical breakdowns* occurs six times, which is often enough to be a significant collocation, but again these occurrences are found in only one file. On the whole, in view of the corpus size of the present analysis, this does not pose too much of a problem (since there are only few similar cases), but analyses based on significantly smaller corpora must take great care to identify such cases and either exclude them or avoid placing too much emphasis on them.

26. Given the deviation of these frequency data from a normal frequency data, I again tested the significance of the difference with the already familiar non-parametric alternative, the *U*-test. While the median frequencies were of course different from the above averages ($median_{-ic}$=2; $median_{-ical}$=3), the difference is still in the suspected direction and (even more) significant: U=604,859; z_{corr}=3.85; p<.001.

References

Berry-Rogghe, Godelieve L. M. 1974. Automatic Identification of Phrasal Verbs. In J. L. Mitchell (ed). *Computers in the Humanities*, 16–26. Edinburgh: Edinburgh University Press.

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

Bortz, Jürgen. 1999. *Statistik für Sozialwissenschaftler.* 5th ed. Berlin, Heidelberg, New York: Springer.

Carletta, Jean. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* 22: 249–254.

Church, Kenneth Ward and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16: 22–29.

Church, Kenneth W., William Gale, Patrick Hanks, and Donald Hindle. 1991. Using Statistics in Lexical Analysis. In U. Zernik (ed). *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*. Hillsdale, NJ: Lawrence Erlbaum, 115–164.

Church, Kenneth W., William Gale, Patrick Hanks, Donald Hindle, and Rosamund Moon. 1994. Lexical Substitutability. In B. T. S. Atkins and A. Zampolli (eds). *Computational Approaches to the Lexicon*. Oxford, New York: Oxford University Press, 153–177.

Cobuild on CD-ROM. 1995. Glasgow: Harper Collins Publ. [CoCD]

Collins Cobuild E-Dict. 1998. Glasgow: Harper Collins Publ. [CCED]

Collins English Dictionary and Thesaurus. 1993. Glasgow: Harper Collins Publ. [CEDT]

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19: 61–74.

Fillmore, Charles, J. and Beryl T. Sue Atkins. Starting Where the Dictionaries Stop: The Challenge of Corpus Lexicography. In B. T. S. Atkins and A. Zampolli (eds). *Computational Approaches to the Lexicon*. Oxford, New York: Oxford University Press, 349–393.

Fournier, Jean-Michel. 1993. Motivation savante et prononciation des adjectifs en *-ic* en anglais contemporain. *Faits de Langues* 1: 235–240.

Fowler, Henry W. 1968. *A Dictionary of Modern English Usage.* 2nd ed. London: Oxford University Press. [HWF]

Halliday, Michael Alexander Kirkwood. 1966. Lexis as a Linguistic Level. In C. E. Bazell et al (eds). *In Memory of J. R. Firth*. London: Longman, 148–162.

Hawkes, Harry. 1976. *-ic* and *-ical*: The Terrible Twins. *Lenguaje y Ciencias* 16: 91–102. [HH]

Jespersen, Otto. 1942. *A Modern English Grammar. On Historical Principles – Part VI (Morphology)*. London: George Allen & Unwin Ltd. [OJ]

Justeson, John S. and Slava M. Katz. 1995. Principled Disambiguation: Discriminating Adjective Senses with Modified Nouns. *Computational Linguistics* 21: 1–27.

Kaunisto, Mark. 1999. *Electric/Electrical* and *Classic/Classical*: Variation between the Suffixes *-ic* and *-ical*. *English Studies* 4: 343–370. [MK]

Kaunisto, Mark. 2001. Nobility in the History of Adjectives Ending in *-ic* and *-ical*. In R. Brend, A. K. Melby and A. Lommel (eds). *LACUS Forum XXVII: Speaking and Comprehending*. Fullerton, CA: LACUS, 35–46.

Kilgariff, Adam. 1997. I Don't Believe in Word Senses. *Computers and the Humanities* 31: 91–113.

Manning, Christopher D. and Hinrich Schütze. 2000. *Foundations of Statistical Natural Language Processing*. 4th printing with corrections. Cambridge, MA: The M.I.T. Press.

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

Marsden, Peter H. 1985. Adjective Pairs in -*ic* and -*ical*: Towards a Systematic Description of Current Usage. *Lebende Sprachen* 30: 26–33. [PHM]

Oakes, Michael P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Oxford English Dictionary on CD-ROM. 1994. Version 1.15, 2nd ed. Oxford: Oxford University Press. [OED]

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language.* London and New York: Longman. [QGLS]

Ross, Nigel J. 1998. The -*ic* and -*ical* Pickle. *English Today* 14: 40–44. [NJR]

Sandra, Dominiek and Sally Rice. 1995. Network Analyses of Prepositional Meaning: Mirroring Whose Mind-The Linguist's or the Language User's? *Cognitive Linguistics* 6: 89–130.

Shipstone, E. I. 1960. Some Variables Affecting Pattern Conception. *Psychological Monographs: General and Applied* 74: 1–41.

Snell, Mary. 1972. *German-English Prose Translation*. München: Hueber.

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

Weeber, Marc, Rein Vos, and Harald Baayen. 2000. Extracting the Lowest-Frequency Words: Pitfalls and Possibilities. *Computational Linguistics* 26: 301–317.