Semantic ID tags : Corpus Evidence for Dictionary Senses

Beryl T. Atkins

Collins Publishers Ltd.
London, England W1X 3LA

ABSTRACT[1]

Large general language corpora offer lexicographers a systematic approach to establishing the senses of a given word for a dictionary entry. This paper describes an analysis of a set of citations for one word, grouped on the dictionary senses; in each citation objective (syntactic and lexical) evidence is sought which identifies the appropriate dictionary sense. Some 70% of the citations are shown to contain such evidence, and a further 20% offer some indication of the particular sense they exemplify. Such an in-depth analysis of the 2,500 most frequently used lemmas of English would it is suggested result in a database relevant to some 85% of language in general use.

1 In darkness, and with dangers compass'd round, and solitude[2]

Corpus lexicography depends on a wealth of data which allows lexicographers to record relevant facts and to check their intuitions about the way any

---

1. All citations in this paper come from the COBUILD corpus, part of the Birmingham Collection of English Text, held at the University of Birmingham, England, and jointly owned by the University of Birmingham and Wm. Collins Sons & Co Ltd. This corpus contains 7.3 million words (tokens), and includes approximately 3.1 million words of non-fiction, 2 million words of fiction, 1 million words of journalism, and 1.2 million words of spoken English (conversations, unscripted broadcasts, lectures). 5 million words are British English and 2 million American English, and 0.3 million words come from other regions. With very few exceptions, the material in this corpus dates from 1970 onwards.

I should like to thank Richard Thomas, Michael Lesk and Roy Byrd for their interest in and encouragement of this work, and particularly Mary Neff, without whose patient help the collocate analysis program would not exist.

2. This description of the lexicographer's habitual state comes from Milton's Paradise Lost

given word may be analysed into senses[3] for the purpose of a dictionary entry. A systematic study of large numbers of citations not only alerts one to facts that might have been missed about the word in question, but also allows one to check out the sense divisions of the entry: if more than a small proportion of citations could be assigned to two or more dictionary senses, or cannot reasonably be assigned to any of the dictionary senses, then the lexicographer's account of that particular word is unlikely to be an accurate one. The process allows the compiler to see a little more clearly the way towards a simple, elegant and helpful description of the word, often highlighting deficiencies not only in one's own work but in existing dictionaries.

Such was the case when the COBUILD entry for the word danger was compiled: the monolingual dictionaries available for reference offered only two current senses, illustrated by the following entries from the Collins English Dictionary (1987) and the American Heritage Dictionary (1985), although any of the fifteen others would do just as well:

[1]

danger ('deindʒə) *n.* 1. the state of being vulnerable to injury, loss, or evil; risk. 2. a person or thing that may cause injury, pain, etc. 3. *Obsolete* power. 4. **in danger of.** liable to. 5. **on the danger list.** critically ill in hospital. [C13 *daunger* power, hence power to inflict injury, from Old French *dongier* (from Latin *dominium* ownership) blended with Old French *dam* injury, from Latin *damnum*] —'**dangerless** *adj*
danger money *n* extra money paid to compensate for the risks involved in certain dangerous jobs.

Collins English Dictionary

dan·ger (dān'jər) *n.* 1. Exposure or vulnerability to harm or risk. 2. A source or instance of risk or peril. 3. *Obs.* Power, esp. power to harm. [ME *daunger*, power, dominion, peril < OFr. *dangier* < Lat. *dominium*, sovereignty < *dominus*, lord, master.]

American Heritage Dictionary

The large body of evidence in the corpus which these two senses could not account for is exemplified by the following citations:

[2] *the danger for an actor-director is that he is tempted ...*
there is the danger that the class think that this is all ...
there's now a danger that we're going to the other extreme
there's no danger that there's somebody working away in Rumania ...

The sense of danger in these examples is clearly neither (1) 'the state of being vulnerable to injury, loss or evil', or 'exposure or vulnerability to harm or risk', nor (2) 'a person or thing that may cause injury, pain etc' or 'a source or instance of risk or peril' A third sense must be added, that of the possibility of something unpleasant happening, giving the following set of senses for danger:

---

3. For the purposes of this discussion, 'word' means a string of characters delimited by spaces, and 'sense' or 'dictionary sense' means any numbered or lettered section of a dictionary entry which supports its own definition or requires separate treatment from the surrounding material.

[3] (1) unsafeness, riskiness, state of being (often physically) threatened or at risk, eg *The matador accepts danger in its most immediate form*; this includes the lexically variable phrase in (...) danger, in the sense of 'at risk', where the (....) slot may be occupied by a variety of adjectives and quantifiers, eg *I never felt in any danger among them*; also noun modifier uses, including items which may be considered as compounds, eg *danger money*.

(2) someone or something that constitutes a threat, eg *A weak person is a danger to his family, to his village.*

(3) possibility that something undesirable might occur eg *There's now a danger that we are going to the other extreme*; this includes the phrase in danger of ... in the sense of 'running the risk of' (the occurrence of something undesirable but not necessarily physically threatening) eg *you may be in danger of not doing your job properly as a governor.*

Before leaving the details of the dictionary entry for danger, brief mention must be made of two particular problems that this word poses for the lexicographer. The first is that the singular form may be substituted for the plural in many contexts, leaving the meaning virtually unchanged, eg

[4] *We have pointed to the dangers inherent in this kind of political system*
Shopkeepers ready to complain at the dangers the soap combine held for them... .

This phenomenon (known in the trade as 'X or instance of X') is a common source of trouble to lexicographers, and occurs with a large number of nouns, not only semantic relatives of danger like threat, peril, risk , but others like attraction, charm or pressure.

A trickier problem is that posed by the ambiguity of the word 'of', in the construction 'the danger of + noun group'. Consider the following :

[5] the danger of heavily congested roads
[6] the danger represented by heavily congested roads
[7] the danger that the roads will become heavily congested.

It is clear that [5] is ambiguous, and may be paraphrased either by [6], giving danger the sense of 'peril', or by [7], giving it the sense of 'unpleasant possibility'. Disambiguation of the word danger in such constructions may depend on the constituents of the noun group after 'of'; sometimes however a longer context makes it clear which sense is intended, as does the actual example in the corpus:

[8] *for an elderly person - now involves the added danger and discomfort of heavily congested roads, fast-moving traffic ...*

In practice, the human brain finds real ambiguity in suprisingly few of these instances, though in many cases of this construction it is certainly possible to force oneself to interpret the citation in a second way. I shall return to this point later. These problems complicate the lexicography of danger, but are essentially peripheral to the task in hand, which simply takes as a starting point the dictionary entry already compiled for this word. Indeed, danger makes an excellent case study, in that it has not so many senses as to make any discussion difficult to follow, but the senses are not so clear cut that disambiguation is simple and instant.

The threefold sense division in [3] above proved adequate to handle the corpus evidence for the purpose of a general dictionary such as those that we (and our computers) are using in natural language processing. In the course of compiling the entry, the lexicographer can rapidly allocate most of the corpus citations to one or other of the senses, without more than a score or so (say 5%) remaining as impenetrably ambiguous, and experience teaches that most of these yield to reason when a longer context is available. In this study I set out to see whether the intuitions and mental processes of the lexicographer are based on objective linguistic evidence which a computer might be able to identify, if it had access to a very detailed and sophisticated lexical database. Is it possible to say, for instance, that this or that fact of syntax or of lexical collocation points unequivocally to this or that dictionary sense? To this end, I attempted to identify and record as many as possible of the linguistic facts relating to the way this word is used, and to its syntactic and lexical contexts, in its different senses.

The task then is to find out how the lexicographer decides that in citation A the word danger is used in sense 1 while in citation B it is used in sense 3? How far does this decision rest on real-world knowledge, and how far on objective linguistic facts? Do most citations carry an ID tag, which if identified points unequivocally to one sense of the word? If so, what form do such tags take? Syntax is simpler for computers than semantics; lexical facts also are easier to identify. (I do not propose to digress into a discussion of what the relationship is between semantics and the other aspects of language, or indeed what semantics actually is.) How far do syntax and lexis take us? How successfully could a syntactically sophisticated computer identify a specific dictionary sense of a word in context? Lexicographers, in the immortal words of the car sticker, do it all the time.

I shall now describe the various steps I took in my search for objective ID tags in the 441 citations for danger in the corpus. In this I am, as it were, simply externalising the processes of lexicography. This work is not a sophisticated operation, for that would require more resources, both intellectual and computational, than I possess; it is simply an effort to look at the objective evidence for sense disambiguation of one single word of the language, albeit a complex one.

2 Types of evidence to be found in the corpus citations

The first step involved identifying and logging up in a simple program on an IBM PC certain types of information for each of the citations which in the normal process of lexicography, had been allocated as follows (the sense numbers relate to [3] above):

[9]
(sense 1)	131 citations	
(sense 2)	136 citations	
(sense 3)	155 citations	
+	16 citations	impossible to allocate to one particular sense without longer context

The notational code used to record the facts in a systematic way is one familiar to many lexicographers, using mnemonic abbreviations for grammar labels such as ADJective, OBJect, and so on. A number of symbols allow various relationships to be noted, such as for example:

[10]
-	(hyphen) joins items into one unit
+	(plus) joins units into code for one structure
[]	means 'is the' eg NG[OBJ-D] = 'noun group is the direct object'
$	means 'within a ..', eg $NG[OBJ-D] means that in this citation danger is found within a noun group which is functioning as the direct object of a verb
))	means 'here is the internal syntax of a phrase'
<<	means 'the following items precede the headword'
>>	means 'the following items follow the headword'

The types of information recorded were:

(a) The word class and number of the headword in the citation, or, if the citation contained an idiom[4], then a description of the lexical unit involved.

Under word class were noted only distinctions of the type used in the Oxford Advanced Learner's Dictionary or other learners' dictionaries, that is whether danger was an uncountable (N-U) or a count noun (N-C) in the citation; singular was taken as the default number, if the citation contained dangers, then the plural (N-PL) was recorded. If the citation contained a phrase, its internal syntax was noted. Examples are:

[11] *a way of life that offers adventure, danger, quick riches or fame*
Recorded as: N-U

... as though he had been in danger of breaking his promise
Recorded as: PHRASE))PREP/in+NOUN+PREP/of

---

4. I use this term as defined by Rosamund Moon: 'a blanket term (referring) to a sequence of two or more words that together function as a unit'; there follows a list of criteria for considering a string of words as a unit. R. Moon, Time and Idioms, in the ZüriLEX '86 Proceedings (forthcoming), ed. M. Snell-Hornby.

(b) The structure(s) appearing in the citation and essential to a full and correct use of the word

These also form part of the regular descriptive arsenal of the learners' dictionary, for example the Longman Dictionary of Contemporary English, where the first sense of danger begins : n 1 [U (of,to)]. Examples are:

[12] *... the power and potential danger of the armed forces*
Recorded as: +OF SB^STH (= of somebody or something)

... danger to health which this entails
Recorded as: +TO SB^STH

(c) The syntactic function of the headword in the citation, as for example:

[13] *... when "Pan" nationalism was the major danger*
Recorded as: $NG[COMPL-COP] (= in noun group which is complement of copula)
Occasionally, this danger can be reversed
Recorded as: $NG[SUBJ]

(d) The syntactic context of the headword in the citation, as for example:

[14] eg *Another danger is that a by-product of nuclear reactors is ..*
Recorded as: <<DET/another : >> COP/be+CL-THAT

(e) The lexical context of the headword in the citations of the corpus.

To record this, a program counted the number of times any word form appeared in a window of seven words on either side of the node word danger or dangers; this was done separately for three subsets of citations, grouped on the basis of the dictionary sense each represented. The positions were named 'minus' if before the node word and 'plus' if after it, so that for the citation *... complained about noise, and mothers worried about the danger of stairs. There was a strong link between ...* the word type mothers stands at minus 4 (or M4), worried at M3, of at plus 1 (or P1), stairs at P2, and so on. The output was in this form:

[15]

WORD	M7	M6	M5	M4	M3	M2	M1	P1	P2	P3	P4	P5	P6	P7
an		1		1	1				1			1	1	1
ancestor											1			
and	5	3	5	5	4	4	4	11		4	1	1	2	1

The object of this operation was to discover whether the occurrence of a specific word type in a specific position vis-a-vis the node word indicated the probability that the node word belonged to one sense of danger rather than either of the others.

3 Objective indications of semantic identity (ID tags)

The four types of evidence described in (a) to (d) above were each logged up in a separate key (sort) field in the citation record, and the records themselves were sorted into four files: three containing the citations for one of the three senses (see [3] above), and a fourth which held the ambiguous citations. The records were then sorted on each of the four key fields and the contents of these fields printed out in tabular form, with separate printouts for each sense-based subset of citations. The syntactic contexts recorded did not seem to furnish many indications of which dictionary sense the citation should be assigned to, and were not in fact used in this operation. However, the other three proved very productive, both individually and in combination; the following extracts show the patterning of 'FOR SB^STH' structure from the resultant listings:

[16] for sense 1 citations:

Headword	Structure	Function
N-U	FOR SB^STH	$PREPG$ADJG
N-U	FOR SB^STH	$NG[MOD-POST]

[17] for sense 2 citations:

Headword	Structure	Function
N-PL	FOR SB^STH	$PREPG[ADJCT]
N-PL	FOR SB^STH	$NG[OBJ-D]
N-PL	FOR SB^STH	$PREPG[COMPL]
N-PL	FOR SB^STH	$NG[OBJ-D]
N-C	FOR SB^STH	$NG[COMPL~COP]

[18] for sense 3 citations:

Headword	Structure	Function
N-C	FOR SB^STH : +THAT	$NG[SUBJ~COP]

The fifth area of comparison, that of the lexical contexts, resulted in very long printouts for each of the three senses. The three sets of sense-grouped citations each produced approximately 800 word types occurring in the specified context of seven words on either side of danger or dangers. I did not do a full statistical analysis of the results of this program, as the frequencies seemed too low to be of much value to this particular study. It was clear that for any such analysis to provide useful data, the corpus would have to be much larger than the 7.3 million COBUILD corpus. With the exception of the function words, most of these types occurred no more than once in any specific position, and the majority occurred no more than once overall, although there were some interesting exceptions, such as grave, great, real and risk. For some of the function words, however, the figures were collated in the following way (the numbers in brackets after the word indicate the relevant sense of danger):

[19]	WORD		M7	M6	M5	M4	M3	M2	M1	P1	P2	P3	P4	P5	P6	P7
	for	(1)		2		1	2	2	3	3	2	1	3	2	2	
	for	(2)	1							7	1		2	3		
	for	(3)	1							1		1				1

Patternings such as these reinforced the idea that certain lexical features might be used to differentiate dictionary senses; for instance, the for statistics suggest that a context in which danger is immediately followed by for is less likely to be sense 3 than sense 1, and more likely to be sense 2 than either. This type of fact is obviously very easy for a computer to identify, and there were a number of such interesting patternings amongst the function words (see Appendix 1). This is an aspect of context which of course must be exploited when very large corpora are available.

4 Types of ID tags for danger

The next step was to isolate individual 'ID tags' for danger - syntactic or lexical markers in the citations which point to a particular dictionary sense of the word. The evidence described above furnished a multiplicity of clues: for example, the structure 'for somebody or something' appeared to be a strong pointer to sense 2: not only is this indicated by the material exemplified in [16], [17] and [18] above, but it is reinforced by the evidence of the collocates computation, in [19] above. (The apparent discrepancy in the occurrences of the FOR SB^STH structure and the occurrences of for in P1 position is explained by citations like *I've been in danger for a long time, Klaus.*)

Evidence of this nature, presented in this way, not only revealed the presence of sense clues in many citations, but also highlighted the existence of some apparently indisputable pointers to specific senses of the word: for example, the occurrence of danger as a count noun (either N-C or N-PL), together with the 'for somebody or something' structure without any other structures, indicates sense 2 and only sense 2. A count noun cannot be sense 1, which by definition (as in learners' dictionaries) is the uncountable noun, while sense 3 usages require some dependent structure such as an appositive THAT-clause (*the danger that a conventional conflict might escalate*) or OF SB DOING (*the danger of the teacher talking too much*). The citation which produced the evidence in [18] is in fact *they understand the technical problems, but the danger for an actor-director is that he is tempted ...* .

It is important to remember at this point that the original allocation of citations to dictionary senses had been done during the normal process of lexicography. I do not claim that the three senses we have here are the 'accurate' - or even an accurate - analysis of danger[5], nor for the

---

5. Indeed, one could argue that this word belongs to the class described by Moon as 'quasi-monosemous'. (Moon, op. cit.)

citation-to-sense allocation being fact rather than opinion (for example, the line between sense 1 citations and those assessed as sense 2 is often not sharp, especially in the case of the singular noun danger with the definite article). These facts are relevant for a lexicographer, but peripheral to the present exercise, which takes as given this particular analysis of the word and the particular way in which the citations were allocated to senses by the lexicographer, and sets out to discover whether it is possible to identify data which explain the compiler's decision that a particular citation exemplifies a particular sense of the word. In other words, are there objective (syntactic and lexical) clues to dictionary senses in real language usages?

On the basis of facts revealed by the recording and sorting operations described in sections 2 and 3 above, I drew up a list of the 'ID tags', that is of syntactic and lexical pointers to specific dictionary senses of danger, which were to be found in the context of this word in corpus citations. There were twenty-six of these altogether: nine for sense 1, six for sense 2 and eleven for sense 3. Here is one example of what I considered to be an ID tag.

[20] (for sense 1) 'uncountable noun with no support'

ie the fact that the usage of danger in the citation was clearly as an uncountable noun (it will be remembered that in learners' dictionaries such as LDOCE, COBUILD and OALD this in fact is stated explicitly for this sense), together with the fact that there was in the citation no structure that was part of the subcategorisation frame for danger, no structure that as it were 'depended on' or 'supported' danger. Sense 2 was excluded by the uncountability of the headword in the citation, and sense 3 by the fact that there was no dependent structure, thus this combination of facts in the context of danger indicated that this example of the word in use could be allocated to sense 1.

An example of a citation which shows these features, which contains certain 'ID clues', is:

[21] *... without mishap. They are, however, fraught with danger. Many an innocent party has come unstuck ...*

In this example, danger is clearly an uncountable noun, as it is used in the singular without a determiner; there are no structures depending on it, as it is followed by a full stop. (There are of course instances, particularly in subordinate clauses, where the fact of being the last word in a sentence or clause does not indicate the absence of dependent structures; however none of these occurred in a significant form in the corpus citations.)

Another example of an ID tag is:

[22] (for sense 2) count noun + in, without that-clause

ie the fact that danger as used in the citation is clearly a count noun (part of the description of this sense in the learners' dictionaries), also

that it has as a dependent structure 'in something' but contains no that-clause. Sense 1 is excluded by the countability of the headword in the citation, and sense 3 by the structure 'in ...' without a that-clause, cf 'there are real dangers in this situation' (sense 2) compared with 'the dangers in this situation are that we may become hardened to ...' (sense 3, although in fact this combination of structures never appears in the corpus); thus a citation showing these factors in the context of danger could be allocated only to sense 2.

An example of a citation carrying ID clues to this particular ID tag is:

[23] *There are, he agrees, real dangers in a partisan Civil Service.*

In this example, the plural form shows that danger is being used as a count noun (following the fairly crude countable / uncountable distinction made by the learners' dictionaries), and the 'in something' construction is clearly seen.

Another example of an ID tag is:

[24] (for sense 3) + appositive that-clause

ie the fact that danger is followed by an appositive that-clause makes it impossible to give this anything but a sense 3 reading ('the undesirable probability that ...'), rather than sense 2 or sense 1.

An example of a citation showing this feature is:

[25] .. *the more necessary because of the evident danger that the Corporate State can ossify into a ...*

The clause beginning 'that ...' is clearly an appositive and not a relative clause, since danger is not the antecedent of the subject or the object of the verb of the clause.

Some ID tags are not as clearcut and indisputable as the three described above, yet they encapsulate features common to the contexts of danger in all the citations of one sense-grouping and in few or none of those allocated during the lexicography to the other senses. They must therefore represent some factor in the lexicographer's decision to allocate these citations to one sense rather than another. The following ID tag is an example of this type:

[26] (for sense 2) 'the danger to', without that-clause

ie the fact of being followed by 'to someone or something' without a that-clause excludes a sense 3 interpretation (see explanation after [22]). However, objective evidence for the distinction between sense 1 (uncountable) and sense 2 (countable) is less clear; in 'the danger to ...', the word danger might be countable or uncountable. Here we leave the realm of certainty and enter that of probability. Checking through the three groups of citations, allocated during the lexicography to one or other

sense, we find that most (9 out of 13 : 70%) of the citations containing 'the danger to' construction had been allocated to sense 2, and only 30% to sense 1. Obviously, the numbers involved are far too small to allow any statement of relative probabilities, but this is not the purpose of this operation. All that may be done, perhaps, is to identify syntactic and lexical features in the context of danger which may have contributed to the lexicographer's decisions, which in their turn must be implicit in the dictionary entry for the word.

An example of a citation carrying the ID tag "'the danger to', without that-clause" is:

[27] *...about the legality of Temple weddings and the danger to public morals of such permissive co-habitation*

Another example of an ID tag which shows probability rather than certainty is:

[28] (for sense 2) singular noun in preposition group functioning as an adjunct, without support

ie the fact that danger is in the singular, and that it stands within a preposition group, which itself is functioning as an adjunct in the clause; and that there is no structure dependent on danger of the sense 3 type, thus excluding sense 3. Why sense 2 rather than sense 1, however? This again is not a statement of certainty, but rather an attempt to reflect corpus data that support the lexicographer's decision to allocate certain citations to specific senses. In this case, examination of the various syntactic functions of the headword (see [13]) revealed that 34 of the 49 citations with $PREPG[ADJCT], or 70%, had been allocated by the lexicographer to sense 2; many of these also carried a more reliable ID tag for this sense, as for example the citation *the professor harangued the silent room on the dangers to civilization of random productive coitus*, where danger is clearly a count noun, thus forcing a sense 2 allocation. In the following citation

[29] *"We will not tell him of the danger," Mohmoud said.*

the ID tag given at [28] is present, with no other clue to indicate whether this citation should be allocated to sense 1 or to sense 2; on the basis of this ID tag, therefore, it was allocated to sense 2.

Mention must be made here of the varying reliability of the ID tags. It is clear that while some of these groups of features (ID tags) could be described as *rules*, that is to say if they exist in a citation there is no option about what sense of the word danger is involved, others are rather *preferences*, since while their presence indicates a likelihood of one sense rather than another this is not shown beyond any doubt. Thus it is possible to classify each ID tag as (a) certain : definite marker leaving no option but specific sense, ie excluding both of the other two; eg [20], [22] and [24] or (b) probable : one sense definitely excluded, the other contra-indicated by 66% or more of the available corpus evidence; eg [26] and [28]. This distinction was taken into account when each corpus citation was

assigned a 'probability rating', depending on the category of the ID tag which it carried, and the ease or difficulty with which this tag could be identified, that is to say the quality of its 'ID clues'.

5 Types of ID clues in danger citations in corpus

If one accepts the principle of the existence of these ID tags, that is of objective (syntactic or lexical) features explaining a lexicographer's decision that in one citation the headword danger belongs to sense 1 while in another it belongs to sense 2, then one factor must be mentioned before any attempt is made to discover how many of the corpus citations carry ID tags. The presence of any particular ID tag is revealed by what in the previous section I call 'ID clues'. Many of these clues are clear and unambiguous; for example there is no doubt about the presence of the ID tag for sense 3 "danger(s) of (somebody or something) ('s) doing" (where do stands for any verb) in these citations:

[30] *... or indeed from simply making known the dangers of relying on Christine Keeler's evidence*
... could play a part in the centre, without the danger of feeling intimidated by professional teachers

while the ID tag for sense 1 "in (+ determiner / quantifier / adjective) danger" is immediately apparent in these citations:

[31] *... you didn't want to worry us. Were you in any danger? If things get any worse you'll just have to ...*
Law and order seems in greater danger than at any time in the last twenty years.

In some cases, however, the presence of an ID tag is assumed on the basis of secondary syntactic or lexical evidence: that is, the ID clue is not immediately clear from the citation, but depends on features of a collocate of danger. An example of this is furnished by the way in which the ID tag for sense 2 "count noun with no support" is identified in the following citation:

[32] *Aside from the radiation risks, the main danger stems from the notion that even one nuclear ...*

The ID clue to the countability of danger in this citation is two-tiered:

(1) the fact that it is modified by adjective main

(2) the fact that main in its 990 (approximate) instances in the corpus modifies over 300 different lemmas, of which 96% are count nouns and only 4% uncountables (see Appendix 2); hence the much higher probability that a noun (in this case danger) modified by main is a count noun rather than an uncountable.

Another example of a non-transparent ID clue is that indicating the presence of the ID tag for sense 1 "uncountable noun with no support" in the following citation:

[33] *... if you know what you're doing, you can reduce the danger to almost nil." "You must be the world's greatest ...*

In this instance there is a three-tiered ID clue, this time to the uncountability of danger in the citation:

(1) the fact that it is the direct object of the verb reduce

(2) reduce has several dictionary senses and only one of these is relevant here

(3) in its sense of 'decrease in amount or intensity', there are 50-odd instances in the corpus of reduce with an abstract noun in direct object slot: 86% of these nouns are uncountables, only 14% are count nouns (see Appendix 2).

6 Probability rating of correct citation-to-sense assignment

It is clear therefore that, just as some ID tags are 'certain' while others are only 'probable', some ID clues are easier to identify and rest on more solid evidence than others: for example, some are clear and incontrovertible indications of the presence of an ID tag; some are incontrovertible indications but obscured by something in the citation; some are clearly identifiable (eg modified by main) but not incontrovertible because they are based on a probability factor (of countability in nouns modified by main), and the probability factors themselves vary from strong (80-100% in corpus evidence) to moderate (above 66%). In an attempt at a realistic assessment, though of course one which is no more than impressionistic and has no claim to statistical validity, the probability of the syntactic and lexical context resulting in a citation's being assigned to the appropriate sense of the headword danger was rated on a scale of 3 to 0, as follows:

[34] 3 certain ID tag(s) with clear and incontrovertible ID clues

2 certain ID tag(s) with clear and very strongly probable ID clue(s), or incontrovertible clues which are in some way obscured or

strongly probable ID tag(s) with clear and incontrovertible ID clue(s).

1 probable ID tag(s) with probable clues

0 inadequate or non-existent ID tag: the citation could equally well be assigned to two, or in some cases three, senses

The purpose of such a rating is to reflect the quality of the objective (syntactic and lexical) evidence on the basis of which the citation is assigned to one sense of danger rather than another. Each citation was therefore assessed individually, and its probability rating takes into account the certainty or probability of the ID tag in question, and the quality and clarity of the ID clue which indicates the presence of that tag. Citations carrying the same ID tag were not necessarily given the same probability rating, as may be seen from the following examples which both carry the "count noun with no support" ID tag for sense 2 :

[35] *... advertisement in which someone runs through terrible dangers and risks simply to put a box of chocolates ...*

[36] *Aside from the radiation risks, the main danger stems from the notion that even one nuclear ...*

The citation at [35] was allocated a probability rating of 3, on the grounds that there were clear and incontrovertible ID clues pointing to the "count noun with no support" tag: the plural form of danger makes it clear that this is the count noun, and the material following the headword, namely the conjoined 'and' phrase followed by an infinitive clause of purpose, excludes the possibility of any sense-3-type supporting structure. The citation at [36], on the other hand, was allocated a probability rating of 2: while the absence of supporting structures is clearly indicated by the fact that danger is immediately followed by the verb of which it is the subject, determination of count noun status depends on the corpus statistics of countable and uncountable nouns modified by main, and this does not constitute the clear and incontrovertible ID clue necessary to afford a probability rating of 3. However the statistics for the probability of being a count noun if modified by main are high (96% of all corpus instances), the ID clue in itself is clear in that main immediately precedes danger and may be assumed to modify it; the "no support" part of the tag is certain, and clearly visible in the syntax of the citation - for these reasons, the citation was allocated a rating of 2, not 1.

On the basis of such reasoning, a rating on a scale of 3 to 0 was assigned to each of the 441 citations for danger in the COBUILD corpus, in an attempt to assess how many of these contained objective (syntactic or lexical) features explaining the lexicographer's decision to assign it to one sense of the headword rather than another, and how incontrovertible and easily detectable this objective evidence was.

7 Syntactic and lexical markers of dictionary senses

The results of this necessarily subjective rating process suggest that a large proportion of the citations for danger in the COBUILD corpus incorporate syntactic and/or lexical evidence of the particular dictionary sense that they exemplify:

72.3% (319 with rating 3) had clear and reliable markers of dictionary sense
10.9% (48 with rating 2) had reasonably clear and reasonably reliable markers of dictionary sense
8.4% (37 with rating 1) had some markers of dictionary sense and
8.4% (37 with rating 0) had no markers of these senses.[6]

The figures for the various dictionary senses were:

For sense 1 (126 citations):	83.3% (105)	rated as 3
	8.7% (11)	rated as 2
	8 % (10)	rated as 1
For sense 2 (136 citations):	73.6% (100)	rated as 3
	15.4% (21)	rated as 2
	11 % (15)	rated as 1
For sense 1 (142 citations):	80.3% (114)	rated as 3
	11.3% (16)	rated as 2
	8.4 (12)	rated as 1

8 Out of this nettle, danger, we pluck ...[7]

Nothing of course can be assumed from a sample of one word (lemma) in a corpus of 7.3 million words (tokens) of written and spoken language[8]. However, if this assessment of danger has any value at all, it is I believe to encourage further work to be done along these lines, more scientifically and systematically, and informed by a coherent and comprehensive linguistic theory. Danger is not an easy word from the point of view of sense assignment in the lexicography: the fact that its full semantic range has not been identified and explained in any major dictionary until the COBUILD corpus evidence was available indicates the pitfalls it holds for

---

6. The discrepancies between this set of figures and those given in section 2 are explained by the fact that (predictably) the human brain had assigned to a particular dictionary sense citations which apparently carried no objective evidence on which to base this decision. The lexicographer, for instance, had allocated to sense 3 the citation *... churning out tracts and posters warning of the dangers of bureaucracy, collectivization and ...*, whereas there seems to be no evidence here for sense 3 ('these nasty things might come about') rather than sense 2 ('these things are dangerous').

7. Shakespeare, 1 Henry IV

8. A brief clarifaction of terminology: take, danger are *lemmas*; take, takes, taking, took, taken; danger and dangers are *types*; and each occurrence in the corpus of any of these types is a *token*.

lexicographers. Many words with a larger number of distinct senses, having higher frequency rating than danger in general corpora and meriting much longer entries in general dictionaries, present fewer and less grave problems of sense differentiation. It is fair to assume that if this type of assessment were done on such words, the syntactic and lexical evidence for sense assignment of citations would be even more pronounced.

A lexical database holding for each of the major words of the language the type of information used in this assessment of danger, culled from a large corpus of general language, would be of interest and value to everyone working in natural language processing, and particularly I believe to those working on machine (assisted) translation. It is interesting to ponder a little on what this would involve.

The COBUILD corpus is part of a larger corpus of general English, the Birmingham Collection of English Text. In the following statistics, figures have been rounded up to the nearest thousand, and percentages to the nearest integer.[9]

The BCET contains 17.78 million words (tokens) including approximately 1.3 million words of spoken material.

In this corpus there are 246,000 word types, of which 130,000 (53%) occur only once.

The 5,000 most frequent word types (2% of total word types) account for 15,469,000 tokens (87% of total tokens). 5000 word types represent approximately 2,200 lemmas, counting average of 2.3 types to each lemma.

If these figures are a correct indication of the way natural language works, it means that an in-depth analysis of the 2,200 most frequent lemmas would provide data on 80 - 90% of the language used every day.

Danger comes in at 1318 tokens, and dangers at 372 : a total of 1690 tokens. Danger is approximately number 1,100 in the word type frequency stakes, and dangers is about number 4,000. This means that they both fall within the 'core' vocabulary listing of 5,000 types, ranging from the (1,082,000), of (535,400) and and (511,300) to militant, mistress and mortgage, all checking in at 293 occurrences. The 5,000th word type ranks around 280 occurrences in this corpus.

The remaining 13% of the material - 2.3 million tokens, 241,000 word types, say 105,000 lemmas - are for the most part words which have only a few senses, or perhaps only one, and which belong to only one or two word classes, in many cases only one. The lexicography of such words is not nearly as complex, difficult or time-consuming as that of the first 2200 lemmas. If a very large corpus of general language were available (I believe that the one-American-billion - 1,000,000,000 - words mentioned informally

---

9. See Typicality and Meaning Potential by Patrick Hanks, in the Zürilex '86 Proceedings (forthcoming), ed. M. Snell-Hornby.

by Roy Byrd[10] in this connection is realistic) an in-depth analysis of the most frequent 2,500 lemmas would furnish a language resource of inestimable value to all those working in natural language processing. Such an analysis should, I believe, be made by lexicographers, with theoretical guidance and computational assistance from linguists and computer scientists experienced in this field, and is not an impossible task. The work already done by the COBUILD team is a considerable step along this road. Predictions made on the basis of the BCET figures are of course very approximate and a much larger corpus is needed before any really accurate estimate can be made. But they do suggest that an analysis of 2% of the language potential would give us data on 80 - 90% of the language in use. If machine analysis of the remaining 98% of the vocabulary were to start with a reference resource of the type I have described, the task should be considerably eased. I should like to urge all those developing very large databases for natural language processing to consider whether some cooperative venture might be undertaken to construct a jointly owned but individually exploited core database for English.

References

American Heritage Dictionary Second College Edition (1985), Houghton Mifflin Company, USA.

Collins COBUILD English Language Dictionary (1987), Collins Publishers, UK.

Collins English Dictionary, Second Edition (1986), Collins Publishers, UK.

Longman Dictionary of Contemporary English, Second Edition (1987), Longman Group, UK.

Oxford Advanced Learner's Dictionary of Current English Oxford University Press, UK.

---

10. During a seminar at the Lexicon Workshop, Linguistic Institute 1987, Stanford University.

Appendix 1

COLLOCATE FREQUENCIES FOR 'DANGER' (dictionary senses in square brackets)

WORD		M7	M6	M5	M4	M3	M2	M1	P1	P2	P3	P4	P5	P6	P7
a	[1]	7	2	1	2	4	1	1	2	3	4	4	1	3	2
	[2]	3	1	1		2	11	11		4		4	7	2	1
	[3]	3	3	2	3	3	4	16	1	6	7	3	2	3	2
and	[1]	5	3	5	5	4	4	4	11		4	2	1	2	1
	[2]	2	5	2	2	6	7	3	11	1	3	4	4	2	5
	[3]	3	2	3	7	5	3		1		2	3	3	3	5
be	[1]	1		1	3					2	3	1			
	[2]	1			2	2	2			1	1			1	1
	[3]				1	1	6					1	2		1
being	[1]				1						1				
(none)	[2]														
	[3]									8	3				
but	[1]	1			1	1			7					1	1
	[2]			2	2		1		3					2	1
	[3]	1		3	3	1	1				2		1		1
from	[1]		1	1		2		4	1		2				
	[2]		1		1		1		3	1		2	2		
	[3]	1		2							1	1			
in	[1]	3	2	3		2	5	14	10	3	1	1	2	2	3
	[2]	2	4	2	3	1			9	6	3	6	6	4	3
	[3]	5	1	3	3		6	29	2		3		3	3	2
is	[1]	2	1	2	2	2	7	5	2	2	2		3	2	1
	[2]	2	2	5	3	4	4		3		2	1		1	1
	[3]	3	3		5	5	16	2	6	4	2	2	1	4	3
no	[1]		2				1	9		1				1	
	[2]	1		2		1								1	
	[3]	1	1		1		1	2				2			
of	[1]	5	8	5	2	3	7	14	2	2	1	4	10	2	4
	[2]	6	7	7	1	6	11		20	2	5	5	5	3	6
	[3]	4	2	3	2	3	5		92		3	6	4	6	4
that	[1]	2	1	5	5	4		1	1	2			3		5
	[2]	2	4	1	2	2	2		6	2	1	4	3	4	1
	[3]	1	2	4	4	3	2	1	19	7	2	2	2	1	3

WORD		M7	M6	M5	M4	M3	M2	M1	P1	P2	P3	P4	P5	P6	P7
the	[1]	3	6	8	6	4	1	17	3	13	16	5	2	13	9
	[2]	8	9	10	7	2	15	51	5	26	6	3	14	9	6
	[3]	8	10	6	6	5	8	50		12	7	8	14	7	11
to	[1]	5	2	3	8	3	5	1	7	2	1	2	3	3	1
	[2]	8	4	10	5	7	8	1	15	2	1	4	5	3	4
	[3]	5	3	2	4	4	3				1	4	3		2
there	[1]				1	7	5		1	1					
	[2]	2		1	2	5	4	1			1				1
	[3]	1	5	2	10	25	2		2			1			
was	[1]		1	1	1	2	3			2	2		1		
	[2]	2	1	1	2	4	1				1	2	1	2	
	[3]		1	1	3	3	9		1				1	3	
which	[1]		1		1						1				
	[2]			2	1				5	3	1	1		1	
	[3]	1									1		1	3	

Appendix 2

The words main and reduce are taken as examples of the way in which the typical behaviour of collocates may serve to disambiguate the headword.

(1) MAIN appears in the corpus both as a noun and and - much more frequently - as an adjective. In the concordances for the adjective, the list of lemmas modified by main contains approximately 312 different nouns (many appear many times in the 990 citations for this word): of these 312, only eleven (4%) are uncountable nouns, and 301 (96%) are count nouns. Many of these nouns have both uncountable and countable usages, but the list which follows includes only those instances where there was no doubt about which usage was involved:

COUNT NOUNS : road, force, building, city, event, hero, characteristic, classification, feature, gallery, terminal, kind, headquarters, library, limb, axe, palace, line, command, structure, department, part, profit-earner, stem, street, subject, theme, thing, area, accuser, achievement, activity, actor, address, advantage, agenda, aim, base, antagonist, application, architect, argument, arsenal, artery, assault, attack, attempt, attraction, avenue, axis, barrier, base, facility, battle, tank, bearing, bedroom, benefit, beneficiary, blade, boulevard, brain, branch, break, bulk, burden, business, cadre, carburettor, card, career, catalogue, category, town, cause, cell, centre, service, challenge, chamber, chance, change, channel, character, charge, group, churchyard, clause, clubhouse, column, computer, comb, committee, complaint, complex, element, concentration, concern, conclusion, concourse, consideration, constituent, contender,

contribution, contributor, market, course, courtyard, crime, criteria, criticism, crossroads, culprit, customer, meal, danger, deck, position, destination, device, difference, difficulty, disadvantage, dish, display, distributor, door, drag, draw, drawback, drift, track, loss, highway, effect, effort, speech, element, enemy, entrance, stumbling-block, essential, event, exporter, mechanism, factor, factory, family, breadwinner, fascinator, field, figure, response, fleet, floor, focus, force, forest, form, spearhead, function, grouping, garden, gate, gateway, goal, group, controller, habitat, hall, hallway, hobby, hope, horizon, hotel, house, hurdle, idea, impact, handle, impact, country, ingredient, instrument, interest, intersection, interval, investment, island, issue, item, jet, job, joint, journal, justification, link, locus, machine, man, market, means, menu, mission, moment, motion, motive, collection, battleground, object, objective, objection, obstacle, one, operation, opportunity, opposition, organisation, orientation, paper, part, passage, people, person, phenomenon, picture, pillar, place, point, party, pollutant, pool, port, contribution, office, preoccupation, prerequisite, principle, priority, problem, product, weapon, medium, pull, purpose, pylon, qualification, question, reason, book, representative, result, ringleader, river, role, room, treatment, route, district, signal, skill, source, spacecraft, speaker, species, sport, square, stage, staple, statement, station, stimulus, story, stream, strength, subsidiary, substance, centre, supply route, sweetener, target, task, association, theme, thesis, thing, thoroughfare, threat, thrust, topic, track, trait, trend, tub, tutor, two, type, unit, user, value, vein, vice, wall, wavelength, way, window, worry, yardstick.

UNCOUNTABLE NOUNS : electricity, energy, fabric, food, furniture, population, accommodation, pressure, spare, support, work.

(2) REDUCE : 49 abstract noun lemmas appear in the direct object slot of of this verb in the corpus citations; of these forty-two (86%) are uncountable nouns, and only seven (14%) are count nouns, as follows:

UNCOUNTABLES : taxation, activity, speed, immigration, expenditure, demand, supply, effectiveness, dependence, demand, inflation, ironing, ability, support, frequency, noise, strength, investment, output, employment, rainfall, stress, suffering, threat, capacity, movement, temperature, efficiency, intensity, volume, cost, weight, wealth, dependence, trickery, content, length, amount, possibility, need, vulnerability, unemployment, number.

COUNT NOUNS problem, tax, loss, business, price, gas bill, feeling.

Many of these nouns have both uncountable and countable usages, but the above list includes only those instances where there was no doubt about which usage was involved.'

Appendix 3 List of ID tags for danger

In the list that follows, the probability rating of the ID tag is marked at 'Tag'; that of the citation is given in brace brackets at 'Example', and takes into account both the probability of the ID tag and the clarity and reliability of the ID clue(s) in the actual citation. Some shorthand terms require elucidation:

'unique to sense X' means that only if danger is interpreted as sense X does the citation read intelligibly. An example of this is the ID tag 'the danger is if' which is unique to sense 3, eg ... *and the danger then is if life ... anyone is trapped ...*

'if Z, not sense X' means that if this ID tag is present, danger can never be interpreted as sense X. An example of this is the ID tag 'uncountable noun', which cannot be sense 2, or the ID tag 'count noun', which excludes sense 1.

'support' refers only to sense 3 usages: danger is this sense (= possibility that something unpleasant will happen) never occurs without one of a small number of 'supporting' constructions, eg ... *that strains the capacity to adapt and creates the danger of future shock*, or *there would be little danger that lateral thinking habits would interfere* ... ; if the word is used without one of these support constructions, as it is for example in *he is a danger to any child who comes in contact with him*, it cannot be given a sense 3 interpretation.

'initial sort' refers to the sorting process during the compiling, when the lexicographer rapidly allocated each citation to one dictionary sense.

SENSE 1

Tag **	:	uncountable noun + no support
Explanation	:	if uncountable, not sense 2; if no support, not sense 3.
Example {3}	:	... *plains, so full of food, were beset with danger. Some sought safety in burrows.*
{2}		*I know all their habits. There's no danger. Reluctantly, I followed him back into the ...*
Note	:	22 instances of danger as complement of 'there is'; of these initial sort gave 18 (82%) to sense 1 and 4 to sense 2; cannot be sense 3 because no support.
Tag **	:	uncountable noun + for/from/in/to/with, without that-clause
Explanation	:	if uncountable, not sense 2; if + for/from/in/to/with, without that-clause, not sense 3.
Example {3}	:	*sufficiently swiftly to be of effect, to prevent danger to hostages*
Tag **	:	single word as sentence
Explanation	:	unique to sense 1
Example {3}	:	*I made a secret signal to him. It meant 'Danger. Go away' because I wasn't sure that he knew ...*

Tag ** : + following noun
Explanation : all noun modifier uses and possible noun + noun compounds
Example (3) : *... construction of the Panama Canal and exposed the danger spots in the Pacific naval defences ...*

Tag ** : in (+ determiner / quantifier / adjective) danger
Explanation : all variants of this lexically variable phrase
Example (3) : *I never felt in any danger among them; I could walk anywhere in the area ...*

Tag ** : out of danger
Explanation : all variants of this lexically variable phrase
Example (3) : *"That's right. Keep Piggy out of danger."*

Tag ** : into (determiner / quantifier / adjective) danger
Explanation : all variants of this lexically variable phrase
Example (3) : *... helpless, new-born baby being brought into this danger, in every knife and pistol fight on the ...*

Tag * : singular noun in preposition group functioning as an adjunct, with no support
Explanation : in initial sort, 70% occurred in sense 1 citations and 30% in sense 2; if no support, not sense 3.
Example (1) : *... had dragged me away that I thought about the danger. But then it didn't matter: the fight had ...*

Tag * : modified by possessive, with no support
Explanation : of the few cases of singular danger modified by possessive, all occurred in sense 1 citations according to initial sort
Example (1) : *if you could smell VC or their danger the way the hunting guides smelled the coming ...*

SENSE 2

Tag ** : count noun with no support
Explanation : if count noun, not sense 1; if no support, not sense 3.
Example (3) : *The male, having survived every danger, thrusts his pedipalp into the female's ...*
Note : every modifies count nouns only not uncountables

Tag ** : 'the danger(s) of' +concrete or +specific; unmodified except by classifying adjective; without gerund.
Explanation : classifying adjective = no comparative or superlative, never modified by 'very', 'less' etc. See Note at end of list.
Example (3) : *... complained about the noise, and mothers worried about the danger of stairs. There was a strong link between ...*
(1) : *... to bring home to everyone the dangers of any other course than that proposed by ...*
Note : noun type restricted to something specific, ie NOT such-and-such a course, but difficult to identify, so low rating

Tag ** : count noun + for/from/in/to/with, without that-clause
Explanation : if count not sense 1; if + for/from/in/to/with, without that-clause not sense 3.
Example (3) : *...and if society is sick, how could it fail t be a danger to peace?*

Tag * : 'the danger to', with no that-clause
Explanation : if + 'to' and no that-clause, not sense 3; of the 13 instances, nine (70%) allocated in initial sort to sense 2 and four to sense 1.
Example (2) : *"If I intervened in the harmony of nature, the danger to you ..." his voice was shaking. "I have ...*

Tag * : count noun + and + noun + of
Explanation : if count noun, not sense 1; unlikely sense 3, as this construction never occurs in sense 3 in corpus.
Example (2) : *These factors have to be weighed against the dangers and anxieties of pregnancy.*

Tag * : singular noun + from, without that-clause
Explanation : if + 'from' without that-clause, not sense 3; unlikely sense 1 because all instances of singular with 'from' allocated in initial sort to sense 2.
Example (1) : *... some time have to be a reckoning with her. But the danger from a re-armed, industrially powerful West ...*

SENSE 3

Tag ** : 'danger(s) of (somebody or something) ('s) doing'
Explanation : unique to sense 3 in initial sort
Example (3) : *... threatened now more than at any other time with the danger of being cast into the pit of pauperism.*

Tag ** : 'the danger is (somebody) ('s) doing'
Explanation : unique to sense 3 (linked to above tag)
Example (3) : *The danger, then, is the therapist's not knowing or ...*

Tag ** : + appositive that-clause
Explanation : unique to sense 3.
Example (3) : *Second, there was a danger that firms might abuse the enormous concentration ...*

Tag ** : 'the danger is that ...'
Explanation : unique to sense 3 (linked to appositive that-clause)
Example (3) : *... and I think the danger with students is that they feel that unless ...*

Tag ** : in (+ determiner / quantifier / intensifier) danger of
Explanation : all variants of this lexically variable phrase
Example (3) : *Art tradition has not prospered at all. It is in danger of collapsing altogether.*

Tag ** : '(a / no / some / little / less etc) danger of'
Explanation : unique to sense 3 in initial sort
Example (3) : *... proclaiming in several languages, "The danger of a new world war still exists, and the people ...*

Tag ** : 'the danger is of'
Explanation : unique to sense 3 in initial sort
Example (3) : *But the main danger, ultimately is not so much of U-turns in ...*

Tag ** : 'danger is if'
Explanation : unique to sense 3.
Example (3) : *... and the danger is if life ... anyone is trapped, or ...*

Tag ** : anaphoric determiner with sentence referent
Explanation : unique to sense 3.
Example (3) : *And it becomes very much a one-man show. How is that danger overcome?*

Tag ** : anaphoric 'this / that' to preceding sense 3 danger
Explanation : unique to sense 3.
Example (2) : *... is here that the danger of future shock lies. This danger, as we shall now see, is intensified by the ...*
Note : danger of future shock identified independently as sense 3.

Tag * : '(the) dangers of' + restrictions on noun group
Explanation : the reasoning here is similar to that explained in Note 1; for sense 3 interpretation, noun following 'of' must be nominalisation, or non-specific, or hyponym of EVENT or ACTION or DISEASE (&c), or contain hypothetical element (eg future or modal verb in clause), and often including -GOOD element.
Example (2) : *Foreseeing the dangers of disintegration in the Warsaw Pact and ...*
(1) : *... some of our friends may be over-conscious of the dangers of dictatorship.*
Note : difference in probability rating for two citations above is because the first one has nominalisation (disintegration) and the second has less strong sense 3 indicator (drought).

NOTE

This complex set of conditions represents an attempt to identify some of the factors that make the human brain opt for a sense 3 rather than a sense 2 reading of constructions of the form 'the danger of + noun', referred to at [5]-[8] in the first section.

(a) Concrete noun or specific noun (ie not abstract + non-specific):

Compare the following examples, in which motorcycles / a road / Joe's motorcycle / our road stand for the class of concrete nouns:

[37] He talked a lot about the danger of motorcycles / a road / Joe's motorcycle / our road.
[38] He talked a lot about how dangerous motorcycles / a road / Joe's motorcycle / our road could be.
[39] * He talked a lot about the undesirable possibility that there could be motorcycles / a road / Joe's motorcycle / our road.

I believe that the fact that [37] may be interpreted as [38] but not as [39] excludes a sense 3 interpretation for [37], and I suggest that this is because the nouns after 'the danger of' are concrete nouns.

Compare the following examples, in which the Second World War / the Ruritanian army / Greenham Common represent specific entities (including of course all proper names) :

[40] He talked a lot about the dangers of the Second World War / the Ruritanian army / Greenham Common.
[41] He talked a lot about how dangerous the Second World War / the Ruritanian army / Greenham Common could be.
[42] * He talked a lot about the undesirable possibility that there could be the Second World War / the Ruritanian army / Greenham Common.

I believe that the fact that [40] may be interpreted as [41] but not as the unacceptable [242 excludes a sense 3 interpretation for [40], and that this is because the nouns after 'the danger of' represent specific entities.

Compare the following examples, in which her irrationality / these prejudices represent nouns which, although abstract, are still specific:

[43] He talked a lot about the dangers of her irrationality / these prejudices.
[44] He talked a lot about how dangerous her irrationality / these prejudices could be.
[45] He talked a lot about the undesirable possibility that there could be * her irrationality / ? these prejudices.

I believe that the fact that [43] may be interpreted as [44] but not as the (possibly) unacceptable [45] excludes a sense 3 interpretation for [43], and that this is because the nouns after 'the danger of' are specific ('her' and 'these') and not general.

Consider now the following examples, where the nouns represented by war / disease are neither concrete, nor specific.

[46] He talked a lot about the dangers of war / disease.
[47] He talked a lot about how dangerous war / disease can be.
[48] He talked a lot about the undesirable possibility that there could be war / disease.

I believe that the fact that [46] may be interpreted as both [47] and [48] indicates that in [46] danger is ambiguous, allowing both a sense 2 and a sense 3 reading, and I suggest that this is because the nouns after 'the

danger of' are not concrete nouns, nor do they represent anything specific.

(b) Unmodified except by classifying adjective :

Compare the following examples (not from the corpus), in which Italian stands for the class of classifying adjectives, while congested represents adjectives which do not belong to this class:

[49] He talked a lot about the danger of Italian roads.
[50] Italian roads are a danger to everyone.
[51] * There is the danger that the roads might become Italian.

[52] He talked a lot about the danger of congested roads.
[53] Congested roads are a danger to everyone.
[54] There is the danger that the roads might become congested.

I believe that the possibility of [50] and impossibility of [51] prove the correctness of the single sense 2 interpretation of [49], while the possibility of [53] and of [54] prove that [52] is essentially ambiguous between sense 2 and sense 3. I suggest that it is the presence or absence of a qualifying adjective in the adjective slot that excludes or permits the sense 3 interpretation.

[55] He talked a lot about the danger of a motorway through the town.
[56] A motorway through the town is a danger to everyone.
[57] There is the danger that there might be a motorway through the town

Similarly, a postmodifier such as the prepositional group above opens the base sentence to both interpretations; [55] above may be read as either [56] or [57].

(c) Without gerund:

Compare the following pairs of examples. In each pair, the first example has already been proved to be capable of a sense 2 interpretation only; however, the presence of the following clause with a gerund transforms the interpretation of danger, making a sense 3 interpretation the only possible one.

[58] He talked a lot about the danger of motorcycles.
[59] He talked a lot about the danger of motorcycles being noisy.

[60] He talked a lot about the dangers of the Second World War.
[61] He talked a lot about the dangers of the Second World War being repeated

[62] He talked a lot about the dangers of her irrationality.
[63] He talked a lot about the dangers of her irrationality infecting the others.

[64] He talked a lot about the danger of Italian roads.
[65] He talked a lot about the danger of Italian roads being blocked by landslides.