Using the BNC to produce dialectic cryptic crossword clues

David Hardcastle, Birkbeck College, University of London

1. Overview

This paper describes an attempt to generate seemingly meaningful cryptic crossword clues without trying to analyse meaning but relying solely on word occurrence statistics. It is a continuation of a project in which I developed an application toolkit for cryptic crossword clue compilers. The software described here assembles simple cryptic clues using the resources developed in the earlier project combined with the British National Corpus (BNC) Sampler.

Some pieces of the process remain problematic making it tempting to look for recourse in grammatical and syntactic data or investigations with a high processing overhead. However, my aim is to try to extract as much mileage as possible from data derived from the BNC that can be processed with a limited overhead.

All of the clues are of a particular type, which I term 'dialectic' clues using the taxonomy of D St P Barnard (Barnard 1963). A dialectic clue is a pair of synonyms for the answer word, appositely combined as a single short phrase. For example, "Delicate but dainty" and "Pretty light"[1] are acceptable dialectic clues for the word "fair". Ideally the apparent syntax of the clue should mislead the person solving the clue by strongly suggesting a different sort of answer. Clues for "fair" such as "Market average" and "Sound common" would fall into this category.

To assemble such clues, the software must first evaluate all possible synonym pairs[2] for the clue word, and decide which pairs are more apposite than the others. Once a list of suitable pairs has been found, the second task is to attempt to link the pairs together in a manner which ideally is meaningful, or failing that at least not too jarring. The principal focus of this paper is on the first of these two tasks, namely identifying apposite pairings and ranking lists of pairings for shared meaning. I shall then briefly address possible ways of linking the pairs together.

2. Finding and ranking apposite pairings

2.1 Retrieving the list of synonyms

This part of the process is very straightforward. I required a list of synonyms for my MSc thesis (Hardcastle 1999), and downloaded a machine-readable version of Roget's 1912 thesaurus from the Gutenberg Project. I then assembled the data from it into a synonym dictionary. Unfortunately many of the synonyms are out-of-date and many form lists of co-hyponyms or loosely associated terms. As a result, the clues are often marred by poor synonyms. The fix for this is clearly to locate a more up-to-date, machine-readable thesaurus, but for the moment the focus of my work and of this paper, is on combining the resulting pairs rather than their quality as synonyms for the clue word.

2.2 Language and meaning in cryptic crossword clues

The aim of this project is the generation of cryptic crossword clues and not sentences. Although these may be seen as an analogue for natural language, there are key differences. Cryptic crossword clues usually have a simple minimal syntax which at best determines the rubric[3] of the clue, and frequently merely acts as filler between the key elements. This is particularly the case for dialectic clues, since they have the simplest syntax of any of the groupings of clues in Barnard's taxonomy.

The key difference between cryptic clues and sentences is that of meaning. Although cryptic clues do not have a reference in the real world, they potentially offer two other levels of meaning. The first is the rubric of the clue, a list of instructions which can be used to solve a clue. The second gives the reader a vague sense that the clue could refer to some situation and as such that it resembles a meaningful English sentence.

---

[1] All example clues given are from output generated by the software, unless stated otherwise.
[2] Lists of synonyms are derived from a synonym dictionary that I constructed for an earlier project
[3] By 'rubric' I mean a simple set of instructions through which the clue may be solved

For example, the clue “effeminate English to embrace the church”[4], a clue for the word “epicene”, is both a rubric and a sort of sentence. The answer means “effeminate”, and can be formed with the letter ‘E’ for English, then the word “pine” around the abbreviation “CE” for church, as such the clue is a set of instructions. Of the many alternatives to communicate one item around another, such as “cover”, “ensnare”, “surround”, the compiler chose the word “embrace”. Similarly the letter ‘E’ could have been clued differently, or the whole shape of the clue could have been different. I suggest that from a wide variety of possibilities, the human compiler settled on a particular set of words with a certain feel to them, since they had just enough in common to suggest a reference. Although the clue is not a proper sentence, it seems to have some sort of meaning, in that it appears to refer to something real. It is this sense of appropriate feel that I am aiming to capture in the ranking of the pairs of synonyms for the simpler dialectic clues.

2.3 Finding apposite pairings

This is the central focus of this paper. Given a list of pairs of words, the software should put the list into rank order according to the extent to which each pair has something in common, rather like the common game of word association. Tightly linked pairs such as “bus” and “stop” should come at the top, then pairs with a tight scope of common context such as “spanner” and “mechanic”, then words with broad shared scope such as “write” and “work” and finally words which show scant promise of association, such as “bookcase” and “goaded”.

One way of achieving this might be to consider the definitions of the words, or to map them according to their membership of certain thematic subsets. However I chose to consider only the extent to which they co-occur in the BNC in order to investigate whether this data alone would be sufficient to rank the list.

2.3.1 Getting the raw data

Since the mark-up language of the BNC specifies a clear hierarchical structure, it is possible to break it down into many different sized chunks, from the level of a whole chapter of a book to a single sentence or even word boundary. Pairs that share a word boundary should be good candidates for the top-scoring set, those which share sentences or paragraphs should be good candidates for the second, and so on down to pairs which are not even much in evidence co-habiting the largest chunks.

To examine pairs and determine the extent of their shared space in the corpus I constructed an index of the corpus sampler. In this index, every instance of the four hierarchical section boundaries (“<div1>” to “<div4>”) is sequentially numbered, as is each paragraph within them, each sentence in each paragraph and each word within each sentence. The resulting index is a dictionary file which provides a list of coded keys for each word in the corpus.

bridge / 313|-1|-1|-1|8|1|39 / 327|102|60|-1|13|3|4 / 350|164|98|-1|6|4|18

Figure 1: A sample entry from the index to the BNC Sampler

Figure 1 shows part of the entry for the word “bridge” and shows the keys for the first three occurrences of “bridge” in the BNC Sampler. The final key (350 164 98 –1 6 4 18) states that the word “bridge” can be found in <div1> number 350, <div2> number 164, <div3> number 98, in an area with no <div4> context and in the 18^{th} word of the 4^{th} sentence of the 6^{th} paragraph of that section. Were the word “suspension” to have a key for example 350 164 98 –1 6 4 17 this would tell us that the phrase “suspension bridge” is evidenced once in the BNC Sampler. Were the word “bidding” to have a key 313 –1 –1 –1 8 1 10, thus sharing the same paragraph as the first key for “bridge”, this would tell us that the words “bidding” and “bridge” shared the same paragraph once in the BNC Sampler.

[4] “Phi”, The Independent Saturday Crossword, No. 4461, 3^{rd} February 2001

Using this index, the software can rapidly return the number of occurrences of any word in the BNC Sampler and also the number of co-occurrences within each level of scope[5] for any pair of words. At present the coding does not take advantage of non-hierarchical scope data such as page breaks or clause boundaries, nor does it differentiate between different sources. Both of these enhancements could improve the quality of the raw data. The former would increase the depth of the data returned, and the latter may prove a useful enhancement to the "<div1>" data, since it could give an indication of size. While a few co-occurrences at the top hierarchical level of a novel, namely at chapter level, may not mean a lot, in the context of the spoken corpus or of ephemera where the top hierarchical level is likely to be much smaller it would be more significant.

2.3.2 Interpreting the raw data

The raw numbers of co-occurrences returned at the different scope levels do give some rough indication of the level of shared context between the two input words. However, such data requires further refinement since common words tend toward the top of the list, and rarer words toward the bottom, regardless of how apposite the pairings.

I examined the feasibility of using the statistical test chi-squared to determine the significance of the co-occurrences for each pair against a random distribution. Unfortunately chi-squared, and other tests for statistical significance, lend additional weight to larger samples. While this makes perfect sense for data such as polls, it only deepened the rift between common and rare words rather than promoting the more apposite pairs. For example, the pairing "work" and "go" scored 17,649 in a chi-squared test, whereas the pairing "ski" and "salopettes" scored only 4.

The hypothesis behind the process of comparison was that the closer the association between a pair of words, the more likely it would be to find co-occurrences of the pair within small chunks of the BNC such as sentences or paragraphs. Therefore the scoring algorithm needed to lend extra weight to co-occurrences found within a small scope, such as a sentence. In order to determine whether the number of co-occurrences found at each level of scope counted as likely or unlikely, I constructed a control set of over 20,000 randomly selected pairs from a machine-readable dictionary (Mitton 1986) to measure the average number of co-occurrences at each level of scope for each pair. To counter the advantage toward words with a high frequency in the earlier scoring systems, I recorded the ratio of co-occurrences to total occurrences of the pair rather than just the raw total of co-occurrences.

The scoring algorithm scores the factor of difference between the average ratio of the baseline pairs and the recorded ratio for the pair under examination at each level of scope. The score is the total of these factors of difference. The factoring process lends extra weight to co-occurrences at small scope level as the baseline ratios for sentences and paragraphs are extremely small. Word boundary co-occurrences are scored in raw quantity, and pairs which evidence such juxtapositions are promoted to the very top of the list regardless of their overall score. To return to our earlier example, "work" and "go" which scored 17,649 with a chi-squared test now scored 1,415 while "ski" and "salopettes" was promoted from a chi-squared score of 4 to a score of 3,785.

There are some drawbacks to this scoring system. Firstly, the use of ratios discriminates heavily against unbalanced pairings where one word is very common and the other very rare. It is also left open to some bizarre results where a word shows up just a few times in the BNC Sampler and all of the occurrences are in a particular atypical context. Finally words that do not appear in the BNC Sampler score zero by default, thus potentially missing good pairings. These problems can all be addressed by using a larger portion of the BNC.

In spite of these potential problems, the scoring system seems to function relatively well, although testing the system proved a difficult task. Since the shared context, or lack of one, between any two words is not a given it is difficult to test the scores generated against some other property of the pairings. At some future stage I would like to compare the rankings of a set of input pairings with the results of a survey where people rank the pairs according to how well they go together. For the time-being I have run two simple tests on the scoring system.

[5] By 'scope' I mean the size of the context within which a co-occurrence has been found, such as a sentence, a paragraph, or a chapter

The first involved some consideration of the distribution of test scores for a large set of randomly generated pairings derived from the dictionary. The resulting distribution is represented in Figure 2 and appears to be relatively promising. Firstly, it seems to be discriminatory, in that the vast majority of pairs failed to score, and that the frequency of scores follows quite a steep curve from frequent mediocrity to the rare sublime pairing at the top of the scale. A cursory inspection of some examples of the ranked pairings also suggested a relatively successful scoring mechanism (see Figure 2).

Score band	%age	Examples
1,001 to 10,000	0.1	goodwill, contract (3,089) factory, working (2.436) jazz, waistband (1,753)
201 to 1,000	0.4	Rain, outlook (907) raid, escapee (440) fluid, stomach (330)
51 to 200	0.8	Club, sixty (187) holy, building (157) guest, tribute (101)
11 to 50	2.7	Waistcoat, guilt (26) blank, game (22) torch, wind (35)
1 to 10	3.1	Timing, accession (4) cardboard, scamper (4)
0	92.9	Haircut, garlic (0) radish, thump (0) Lutheran, shaker (0)

Figure 2: some examples of pairs in the different score bands following a test on the scoring system of a large sample of word pairs. The percentages represent the total number of pairs in each score band as a percentage of the total pairs under examination

Finally I tested the scoring mechanism with a set of chosen pairs using the word "market" that I felt to have more or less in common along a relatively clear scale. The resulting rankings do at least seem to group the pairings roughly into the correct halves, with most of the top ten being those which I picked because I felt that they shared a common context, and most of the bottom half of the table being those which I chose to represent a poor association (Figure 3). Four of the expected top five pairs received the highest scores, with "fruit" unexpectedly coming in 7th place. However, the words which received low scores did not do so just because of poor frequency; as "fruit" scored 177 with "tree" and 871 with "veg", while "church" and "Catholic" scored 1,703 as a pairing.

Expected top 5 scoring set			Expected bottom 5 scoring set		
	Score	Ranking		Score	Ranking
Average	1542	1st	Church	145	5th
Sell	1000	2nd	Plate	67	6th
Money	580	3rd	Catholic	15	8th
Europe	560	4th	Ski	15	8th
Fruit	47	7th	Headlights	5	10th

Figure 3: Scores for a set of words paired with "market"

2.3.3 Refining the scoring process – 'third party co-referents'

Although this system of scoring provided a reasonable system of ranking pairings, it did not perform well for words with very low frequencies in the BNC Sampler. For example the words "headlights" and "wipers" which clearly have a strong association receive an unimpressive score (9), with only 2 matches between them and those at the widest scope. Despite the strong word association, it is unlikely that we will find many co-occurrences within the same sentence or the same paragraph. Indeed, even with a much larger corpus we may not find a significant number of co-occurrences within small scopes.

To deal with pairs of words where one or both occurred relatively infrequently, I decided to examine which other words occurred in close proximity. Reversing the index shown in Figure 1, I generated a look-up table of all the words which co-occur in the same paragraph as the key. I will refer to the list of words which co-occur at paragraph level with both input words as ‘third party co-referents’. Initially the lists of co-referents were dominated by function words, pronouns and common verbs. To determine the culprits of this ‘noise’, I generated lists for a large sample of input words and identified words which appeared in more than 80% of the outputs. These words are filtered out of all lists of co-referents which the software produces.

Figure 4 shows the most commonly co-occurring words within paragraph scope for “headlights”, “brake” and “donkey” filtered for noise. The words “headlights” and “brake” share two of the top seven entries, and indeed many more of their respective full lists, while “donkey” shares none in the top seven and few in the full list with either of the other words. Such ‘third-party co-referents’ provide a means to promote such a pair up the rankings. An important point is that the list of paragraph-level co-referents does not necessitate that “headlights” and “brake” co-occur in any paragraphs themselves, instead the co-referents represent an intermediate measure of shared context to make up for the lack of direct evidence.

Headlights	Brake	Donkey
Fog *	Caravan	Ballot
Grandma	Solenoid *	Votes
Bumper	Portable	Rifle
Vibrating	Toot	Exhaustive
Spotlights	Slate	Candidate
Bonnet	LTD	Battalion
Balancing	Fog *	Voters
Flash	Fluid	Camel
Solenoid *	Daimler	Ape
Flashing	Starter	Spanish

Figure 4: co-referent lists, “fog” and “solenoid” act as third party co-referents between “headlights” and “brake”

Returning to the example of “headlights” and “wipers”, the software was now able to score a list of third party co-referents, words which co-occurred at paragraph level with “wipers” and also with “headlights”. Using this intermediary the pair now scored 62 rather than 9, a modest but notable improvement. The system also provided some measure for pairs which had previously scored zero, such as “wipers” and “brake” which rose from 0 to 40. The system occasionally threw up some rather unusual pairing suggestions, such as “Catholic” and “sheep” which scored 610, although further investigation always uncovered a rational explanation and an overlap of reference.

At present this system of third party co-referents remains experimental, although the results so far have been very promising. I selected paragraph level scope since I felt that it was wide enough to provide sufficient information to work with, while being sufficiently small that the co-reference could reasonably said to mean something. It may be that a larger scope such as page level, or a reduction to sentence level may produce cleaner or more informative result sets.

It is notable that at present this system has been tested predominately with nouns. Nouns, and count nouns in particular, seem intuitively to be more apt to this type of processing as they can be more readily ascribed to thematic subsets. Although many verbs and adjectives share this property, it is arguable that the majority have too general a function for this process to bear fruit. Whether or not this distinction holds is something I hope to explore as I develop this co-referencing process.

2.3.4 Third party co-referents and meaning lists

As I explored the first few lists of third party co-referents I observed that for many words the lists contained many key thematic elements that to some extent defined the context of the input word. Although an unedited list of these words would prove a crude measure of meaning, I felt that the cross-reference between this list and lists of co-hyponym sets might provide some measure of meaning.

To clarify, the idea would be to compare an input word to a set of co-hyponym lists. The cross-reference would involve neither the input word, nor the co-hyponym key, but a cross-referencing between the list of 'third party co-referents' for the input word and the list of co-hyponyms under the keyword. A large overlap might indicate that the co-hyponym keyword represents some core aspect of the meaning of the input word.

The thesaurus I have been using contained many lists of co-hyponyms, indeed many of the clues I have generated still suffer from a poor rubric as they contain co-hyponyms rather than synonyms. In order to reduce the extent of this problem I removed the tagged lists of co-hyponyms from the thesaurus. Figure 5 shows the results of cross-referencing the third party co-referents from the input words with the co-hyponym lists from Roget. The percentages represent the percentage of cross-references between a co-referent and a co-hyponym which occurred in each category. Although the lists are crude and short, numbering less than 30 in total, the results quite clearly favour some reasonable and interesting keywords.

Church:	Social (23%)	Religion (17%)	Military (17%)	Politics (12%)
Horse:	Animals (30%)	Industry (9%)	Military (9%)	
Doctor:	Politics (18%)	Social (15%)	Science (9%)	Medicine (7%)

Figure 5: Examples of the results of cross-referencing the co-referent list for the input word with lists of co-hyponyms.

With the help of the full version of the BNC and a fuller set of co-hyponym lists, I hope to be able to return a list of meaning keywords for most noun input words. As above, it is highly likely that while this system will work well for nouns and for count nouns in particular, it will not work so well for verbs, adjectives and adverbs since they are not so readily classified into sets.

3. Linking the pairs together

Having ranked the pairs of synonyms for the clue solution word according to their shared context using the processes described above, the next step is to link the pairs together in as appropriate a fashion as possible.

3.1 Juxtaposed pairs

Some of the pairs may have been found in juxtaposition within the corpus. Given that there is evidence of how to combine them already, the best chance for an idiomatic feel is to present them as they were found. For the most part, this approach seems to have been relatively successful producing clues such as "Still soft" for "gentle", "Standing order" for "condition" and "Tax band" for "press". Where the approach is less successful is where the pairing forms a part of a longer idiomatic phrase. While "Pretty big" sits happily as a unit on its own, the pairing "Great big" scores the same but is clearly inferior, as a third word is expected.

It is possible that the difference in quality between "pretty big" and "great big" could be measured with closer examination of the surrounding tags and clause boundary markers. However, I feel that to set out in such a direction would detract too much from the core of this project, which is to produce dialectic cryptic clues at low processing cost, and using descriptive data from the BNC.

Thus pairs evidenced in juxtaposition are left as they are found, and the software takes its chances with the results. Provided that the words are not function words or pronouns, unlikely in the context of synonyms for a clue word, the result should at the very least be acceptable and seems quite frequently to produce a nice pun. I describe the scoring of puns below (see below Section 4).

3.2 Pairs requiring a link word

The remaining pairs of synonyms were not found in juxtaposition, and so they cannot be presented as a clue until a suitable way has been found to link them together. One way of achieving this is to examine what part of speech the words could be and linking them in appropriate ways. For example, singular noun plus "to" plus infinitive intransitive verb, infinitive transitive verb plus singular noun, adjective plus singular noun, noun plus "or" plus noun, and so on.

While this approach generates some passable clues they do not read very idiomatically and, since such an algorithm requires a decision on what part of speech each component is, homograph puns are likely to be missed. Furthermore, the system does not allow for idiomatic links between pairs, since any departure from the simplest grammatical structure is likely to generate unsightly exceptions. For example, given the words "able" and "resolve" we could choose to interpret them as an adjective and noun pair and produce "able resolve". However, they could also be interpreted as an adjective and verb pair, but the phrase "able to resolve" would be a significant risk as this structure is not guaranteed to work with the majority of adjective-verb pairings and may produce some awful phrases such as "red to drill" or "peaceful to dig".

The phrase "able to resolve" is acceptable since one can be "able to" do many things. Rather than set about listing the adjectives that could fit this pattern and build a prescriptive grammar, I decided to list the conjunctive phrases found immediately before and immediately after all of the words in my dictionary in the BNC. For want of a better source of conjunctive phrases I opted to use a list of function words (Mitton 1996) and to count a phrase as one, two or three function words in a group. Some examples of the resulting statistics are listed in Figure 6 and Figure 7. The entries for "Red" and "Grounds" in Figure 7 exemplify problems with adjective position and framing phrases, these issues are discussed in greater detail below

Contact:	with (63), between (8), by (5), from (3)
Popped:	in (44), along (11), down (11), on (11) round (11), up (11)
Ready:	to (45), for (30), and (6)

Figure 6 Examples of statistics for conjunctive phrases following dictionary entries, the numbers in brackets are percentages representing the frequency of each phrase in the total sample found

Austria:	in (25), of (25), and (12), from (12), to (12), with (12)
Perspective:	in (35), of (14), on (14), and (7), from the (7), of the (7), that (7)
Red:	a (19), of (12), in the (6), into the (5), that (5), with (4), and the (2) …
Grounds:	on the (65), in the (10), of (7), on (2), outside the (2), within the (2) …

Figure 7 Examples of statistics for conjunctive phrases preceding dictionary entries, the numbers in brackets are percentages representing the frequency of each phrase in the total sample found

Rather than attempt to combine the pairings by determining their part of speech and looking for a grammatical structure to accommodate them, the software looks at the function words which are evidenced preceding and following each word in the pair. If a fit is found between the preceding conjunctive phrases of one word and the following conjunctive phrases of another it is scored according to the combined relative frequency of both. In this way the phrase "able to resolve" becomes a safe bet with a score of 186 out of a maximum 200 rather than a gamble. Other outputs from this process are listed in Figure 8.

left a pot (46), pot on the left (34), left to pot (20), pot to the left (17), left in the pot (15), left on the pot(13) …
likely to go (158), likely that go (13), likely a go (4) …
pile of grass (82), pile in the grass (25), pile on the grass (16) …

Figure 8 Some pairs generate several similar scores, such as "left" and "pot", others just a single high score

Where a fit cannot be found the pair does not receive a score for linking and may be demoted to the benefit of other better fitting pairs. If none of the pairs with good context scores has a good fit, then a rough fit must be made by determining the what part of speech each member of the pair could be and assembling them according to a very rudimentary grammar.

Although this approach produces successful results, there are still occasions when we are left with a clue which is one word short of an idiomatic phrase. The key problems which remain in the assembly of the pairs are:
Intervening words. Most commonly adjectives are the culprit. Not only does the intervening adjective deprive the software of data about the conjunctive phrase and the noun, it also results in phrases such as "in the yellow" and "under the long" seeming acceptable. Cleaning out such data without losing idioms such as "in the red" and "in the main" will be difficult, although it may be possible to find the idioms through testing the whole combination for frequency.
Frames. As with the previous problem, the result is phrases that feel like they are missing something, such as "for the reason" and "on the assumption", both of which require the word "that" and a clause to follow. It is difficult to see how to remove these without a considerable processing overhead.
Clause boundaries. This problem arises from the indexing of the spoken section of the BNC Sampler. Since the clause boundaries are coded differently, they are not accounted for in my index, and as a result pairs of words crossing word boundaries are recorded as contiguous. For example the word "hand" records "yes" as a following phrase, probably from the text "would you like a hand? Yes … ". This could also cause unusual results in the return of supposedly juxtaposed pairings. Fortunately the solution is not complex, only time-consuming, and will involve refreshing the index to take account of the difference in mark-up.

4. Scoring puns

At present the resulting pairs pass through a very straightforward filter to score the puns which examines the thesaurus entry and checks to see if the pair is a combination of words from single or multiple entries of the thesaurus.

Homograph puns feature fairly heavily in the resulting clues, in particular in the clues derived from juxtaposed pairs. The present system, although a crude measure, picks up the majority. A precise measure would require a substantial overhead, especially since homograph puns are encouraged by the software not differentiating between homographs in the processing of the synonym pairs. Indeed, for the bulk of completed clues, the software will at no point have investigated what part of speech the composite parts might be.

Pairs scoring as puns for the word "fair" included "market average" and "sound common". Non-scoring pairs included "pearly white" and "white light". The latter of these is in fact a pun, "light" being suggested as a noun. However to determine this the software would have to return to the reference in the corpus and establish that the pairing "white light" which evidenced the juxtaposition was an adjective noun pair, and this requires a substantial overhead.

5. The production of a dialectic clue

The following description provides an overview of the whole process of the generation of a dialectic clue with examples of each processing stage, using "reign" as the input word.

The input word, "reign", is used as a key against the thesaurus to generate a list of all the possible pairs of synonyms for the word. Each pair is scored using the index of the BNC Sampler against the baseline ratios for co-occurrence at various levels of scope. This process is refined by comparing the extent to which lists of all co-occurring words at paragraph level scope, third party co-referents, overlap for each pair. The pairs are sorted into ranked order on the co-occurrence and third party co-referent scores. Any pairs found in direct juxtaposition are separated off.

Lead weight	(direct juxtaposition “lead weight”)
Lead capability	(direct juxtaposition “capability lead”)
Authority pressure	
Importance authority	
Authority control	
Effect authority	

Figure 9 The top pairings for “reign”. Many of the synonyms are not particularly good, but the pairings share some common context.

Data from the BNC Sampler on conjunctive phrases occurring before and after each member of each pair is examined to find a match. The top-scoring match is selected for each pair, and the score recorded. By combining the linking score with the context score, pairs which have less context but link together more idiomatically are promoted.

Pressure of authority (23), authority under pressure (22), authority of pressure (20) …
Importance of authority (64), importance of the authority (51), authority and importance (28) …
Control of authority (57), control of the authority (44), authority to control (40) …

Figure 10 The results of combining the top-scoring pairs using link words. The first two pairs (lead weight and capability lead) are not linked as they were found in juxtaposition

A filter checks each pair to determine if it could be a homograph pun. Puns are promoted to the top of the list. The software selects the top-scoring punning clue, or the user can select their favourite from the head of a list of formatted clue suggestions.

Lead weight (5)
Capability lead (5)
Authority under pressure (5)
Importance of authority (5)
Authority to control (5)

Figure 11 Sample formatted clues for “reign” using varied link words

6. Evaluation

I evaluated an early version of the software by generating dialectic clues for a set of six clue words which had been clued dialectically in broad-sheet crosswords. I selected the top five scoring clues from each clue word, and presented them to a group of crossword enthusiasts in groups with the broadsheet clues mixed in. The enthusiasts then scored the clues for readability without knowing the answer word. I then averaged the scores of the broadsheet clues. For readability the clues from the broadsheets scored 3.4, meaning that on average they were placed between 3rd and 4th place against 5 clues generated by the software, indicating stiff competition from the computer-generated clues on this front.

The software-generated clues fared less well when the answers were revealed as they frequently contained good puns made with very poor synonyms. Some computer-generated clues were deemed to be of good quality, even when the answer was revealed. These included “market average”, “sound common” and “common market” for “fair”, “dictatorship of the regime” for “reign”, “tax band” for “press” and “separate part” for “isolate”. However, many clues were deemed to be insoluble as the synonyms were too difficult or inappropriate. For example, the clue “Lead weight (5)” was a neat pun, but could not be solved as “weight” is too remote from “reign”. Similarly “Degree of distinction (9)” again read well, but “distinction” for “condition” was felt to be unfair.

7. Summary

The result of the evaluation suggested that the project has so far been largely successful and has generated some appealing dialectic cryptic crossword clues without human intervention. However, an improved thesaurus is clearly much-needed if clue quality is to be raised.

The software identifies apposite pairings without recourse to definitions or tables of meaning, but using the levels of co-occurrence in the BNC Sampler and through indirect co-occurrence by means of lists of 'third party co-referents'.

It identifies appropriate bridging phrases to link pairs together where they are not evidenced in juxtaposition without recourse to syntactic or grammatical guides, but purely through frequency data gleaned from the BNC Sampler.

References cited

Barnard D. St. P 1963 *The Anatomy of the Crossword* Bell

Hardcastle, D 1999 *SphinX Crossword Compiler's Toolkit* Unpublished MSc Thesis, Birkbeck College, University of London

Mitton R, 1986 "A partial dictionary of English in computer-usable form" *Literary and Linguistic Computing* 1(4): 214-5

Mitton, R, 1996, *English Spelling and the Computer*, Longman, Appendix 2