# Summary of Cryptic Crosswords

## Cryptic Crosswords In general

## Cryptic Crosswords in the Literature

While not a topic well covered in scientific literature in general, what few analytical studies around cryptic crosswords there are tend to be classifiable into three main groups:

The largest body of work that exists is centred around the generation of cryptic clues, focused largely around analysis of how string literals from a pre-determined answer can be transformed by set clueing patterns, as well as some work around measures of the quality of generated clues.

The next set are the select few who have done prior, similar investigations into interpreting cryptic clues, with some work put into formalising definitions and notation for the sorts of clue types that appear in the majority of cryptic crosswords, and some attempts at solving based on these interpretations. There has also been some work done towards solving non-cryptic crosswords probabilistically, working on whole-grid solutions rather than individual clues.

There are also some more left-of-field studies done: statistical studies into errors made during manual solving, and psychological studies into solving.

## Complexity

A variety of factors make solving cryptic crosswords a difficult problem:

* **Ambiguity**
  + Cryptic crosswords are deliberately ambiguous. Instruction indicators are indistinguishable from string literals, which are identical to words’ semantic meanings. Often, the setter will deliberately chose words to give rise to further ambiguities.
* **State Space**
  + Even with only a few different clue types, the number off different readings of one clue based on those grows exponentially with the length of the clue
* **Lack of Standardization**
  + Although all cryptic crossword sahre some common conventions, there are no fixed rules shared between publications for what can and can’t be a clue, indicator etc.. Although most publications have internal guidelines, these are not accessible to the solver, and some publications (such as the Guardian) have named setters whose styles and self-imposed rulesets differ, even between one publication.
* **Knowledge Base**
  + As well as being made up of encrypted and hidden meanings, cryptic crosswords also draws on a diverse knowledge base of synonyms, abbreviations, facts etc.

## Programming Language Analogues

Much of the current work on interpreting crosswords draws on work by Backus, Naur and Chomsky

# Parsing Frameworks and Notation

Some different notations for denoting parsing of cryptic clues have come out of previous work – in order to properly provide a rigorous analysis of the structures and conventions of cryptic crosswords, it is necessary to analyze and choose a framework in which to do it.

## LACROSS

William and Woodhead produced language called LACROSS, which forms a sort of calculus for describing crossword clues. They also provide a BNF definition of this grammar.

Their clues are of the general form

Clue := Δ = G | G = Δ

the orientation of which corresponds to the order in which we find the definition (Δ) and the wordplay (G) in the clue. The wordplay may be further expanded out – the wordplay section of the clue is expressed as a sequential annotation for the constituent parts, either as ‘text’ (t), ‘shortening’ (S) (etc.) or as placeholders for the operators (\*), which are detailed afterwards, including a reference to the substituted indicator. So for instance:

Get in odd bit of colour (5) [= tinge]

t\* = Δ, a (odd, a)

There are several issues with this grammar. Firstly, all unitary operators are treated the same, as are all binary operators, and there is some issue with binding and precedence which they address with an underlining notation, in addition to brackets. Secondly, the grammar attempts to include both the structure of the parsed cluing and how that structure relates to the original sequence of words at the same time. As a result, we end up with complex grammar which does not aid human parsing of the solution well, nor does it lend itself easily to computer or mathematical manipulation

## Simple Clue Markup Language

Proposed by Hall and Rapanotti, Simple Clue Markup Language (SCML) attempts to notate the structure of the solution directly onto the clue.

Double underlining (bold underlining, here) is used to denote the definition, underlining denotes an operator, with its class as an optional subscript, with scope provided by brackets and concatenation (and definition/wordplay separation) given by a semi-colon.

Thus in their given example:

Note the shuddering domestic appliance Bill regularly installed, noisy thing (6,7)

Note; (the) shudderinga; (domestic appliance, (Bill)regularlyt)installede;**noisy thing**

* ‘Note’ often indicates a musical note, resolving to one of ‘a’ to ‘g’, ‘do’, ‘re’, ‘mi’, etc;
* ‘the shuddering’ may be an anagram indicator applied to ‘the’;
* the ‘regularly’ of ‘Bill regularly’ may indicate alternate letters (‘t’); i.e., ‘bl’ or ‘il’; and
* ‘installed’ suggests the embedding (‘e’) of those letters within something meaning ‘domestic appliance’.

In this, we have no markup differentiation for literal strings (‘Bill’) against words with their semantic context (‘domestic appliance’), and we also take certain words that reduce to abbreviations (‘Note’) to be non-deterministic nullary operators.

With some changes and additions (tagging of string vs. semantic word, for example), this markup serves as a good way to represent a parsing of a clue in a human readable way. It even has the advantage that a printed clue could be annotated (carefully) by hand, as a teaching aid, for example.

Unfortunately, the language as it stands is not expressible as a BNF grammar, nor is it a particularly good format for representing the clue and its parsings internally in a program (as it would need to be re-parsed to use!)

## Clue-answer notation

There are several emergent solutions within online cryptic crossword communities for notation to explain solutions derived from clues. From cryptics.wikia.com:

“Consider the down clue

A message from the setter, hauled up with broken arm after heroin withdrawal (8)

yielding the answer TELEGRAM. The corresponding wordplay, having the prolix and possibly ambiguous explanation

*THE next to LEG reversed next to an anagram of ARM, all with H (heroin) removed*

could be concisely represented in clue-answer notation simply as T[h]E,GEL<=,(ARM)\*.”

These meanings are not fixed, but some definitions are given here:

|  |  |
| --- | --- |
| **Notation** | **Usual meaning** |
| ABC<= or ABC (rev.) | ABC reversed. The (rev.)notation is most commonly used when the wordplay consists of a single reversal. |
| [abc] or -abc or (abc) | Letters abc removed, as in**[c]OUNT**to represent '**count' with c removed**; the convention is to use lower case for the removed letters. |
| (ABC) | Letters placed inside others, as in**C(AND)ID** to mean '**and' inside 'cid'**. |
| "ABC" | Homophone of ABC. |
| (ABC)\* | Anagram of ABC. |
| A+B or A,B | A concatenated with B. Sometimes both notations are used together where ambiguities may arise. |
| aBcDeF | Alternate letters of ABCDEF (shorthand for[ a]B[c]D[e]F). |

## PICCUP

Hart and Davies define what is currently the most satisfying proposal for a formal syntactical definition of cryptic crossword syntax, in a loosely BNF grammar. Theirs is the only current definition that closely resembles a usable formally defined language.

## Syntactic and Metasyntactic Conventions

Here we apply a similar convention to Hart, in using a modified Backus Naur Form (BNF). We will later see that a context-free grammar may not be sufficient to model a cryptic crossword, and may have further deficiencies as a basis for finding a solution. Nevertheless, we will adopt a similar notation:

→ = is composed of

, = followed by

| = or

(x) = x is optional

x\* = 1 or more occurrences of x

(x)\* = 0 or more occurrences of x

We also take the BNF conventions

word = non-terminal symbol

“word” = string literal

[x, y, z] = list containing x y and z

(x, y) = pair x and y

For clarity, we additionally define:

{words} = English word including semantic meaning

For this last one, we really mean that we can still use the word in a synonym equivalence relationship.

# The Cryptic Crossword Clue

## Structure of a cryptic clue

A cryptic crossword differs from a normal crossword in that the clue for each answer consists of two parts.

The first is the definition, which performs the same function as a clue in a 'regular' crossword. The answer to the clue is usually a synonym for the definition ('circular' and 'round') or may be an example of the definition ('farm animal' and 'pig'). Other forms that the definition may take will be discussed later on.

The second part of the clue is the wordplay. This is an encoded and often ambiguous second method of deriving the answer, using techniques such as anagram, substitution and concatenation.

The clue as a whole is presented as a concatenation of the two parts, sometimes with a subsidiary word indicating that one can be derived from the other (for example, 'from' or 'is').

We can present this breakdown as:

Clue → Definition, (Indicator)\*, Wordplay   
 | Wordplay, (Indicator)\*, Definition

The final clue will often resemble a valid English utterance, although this 'surface reading' (i.e. {clue} ) very rarely has any relation to the answer.

Later on we will consider other information and context within the definition of a clue.

## Definition

The definition of the clue consists of one or more English words. The answer to the clue will be a word or phrase that fits an appropriate equivalence function. Usually, the words are synonyms:

{allies} = “Friends”

although sometimes the relation is an ‘example of’ relation.

{country} = “France”

The definition carries a variety of linguistic features with it that the overall answer, and so the answer as derived by the wordplay, must match. These include aspect (noun, verb, adjective), plurality (tree, trees), tense (go, going, gone). These features may also be considered as context to the clue itself.

## Formally

We can define the definition as

Definition → {Word}

## Wordplay

The wordplay section of a clue is a set of deliberately ambiguous instructions that allows the solver to arrive at the eventual answer. As the instructions are ambiguous, multiple possible parsings of the instructions are possible. Some of these parsing will not lead to a valid English word:

Imbecile, bonkers, in a cult (7)

⇒ Wordplay ‘Imbecile, bonkers = definition ‘in a cult’

⇒ Anagram ‘imbecile’ [indicator = bonkers] = definition ‘in a cult’

⇒ ??? (no anagrams of imbecile)

(correct reading was anagram

Others will lead to a valid English word, but one that is not equivalent to the definition:

Minder shredded corset (6)

⇒ Wordplay ‘minder shredded’ = Definition ‘corset’

⇒ Anagram ‘minder’ [indicator = shredded] = Definition ‘corset’

⇒ ‘remind’ = Definition ‘corset’ X

(correct reading was anagram ‘corset’ = escort = minder)

The solver must find the correct parsing of the wordplay that yields the correct definition: even though they may not know which part is wordplay and which is definition.

## Wordplay Types

Some sort of leadup

Wordplay → Anagram | Concatenation […]

## Unitary Operators

**Anagram**

A very commonly used operator in crossword clues is an anagram. These take the form of an indicator word that denotes that the anagram function is being used (called an ‘anagrind’ within cruciverbalist circles), along with the candidate letters to be anagrammed. The simplest form of this gets the candidate letters verbatim from the clue:

Anagram → Anagrind, “candidate words” | “candidate words”, Anagrind

Sometimes, however sometimes there is some sort of operation applied to the letters before the anagram is applied. For example:

Comic bare for short comedy play (7,5)

⇒ Wordplay ‘Comic bare for short comedy’ = Definition ‘play’  
⇒ Anagram ‘bare for short comedy’ [anagrind = ‘comic’]   
⇒ Anagram (“bare for” + Shorten ‘comedy’)  
⇒ Anagram (“bare fore” + “comed”)   
⇒ Anagram (“bare fore” + “comed”)   
⇒ Anagram (“bareforecomed”)   
⇒ “Bedroom Farce”

In which case we find the more general case:

Anagram → Anagrind, Wordplay | Wordplay, Anagrind

Wherein we know that the repeated evaluation of the Wordplay will eventually result in a string literal that can be anagrammed. Ximenes argued against this form of indirect anagram:

Secondly – and here, for once, I differ from Afrit – I hate what I call an indirect anagram. By that I mean "Tough form of monster" for HARDY (anagram of HYDRA). There may not be many monsters in five letters; but all the same I think the clue-writer is being mean and withholding information which the solver can reasonably demand. Why should he have to solve something before he can begin to use part of a clue? He has first to find "hydra" – and why shouldn't it be "giant"? – and then use the anagrammatic information to help him think of "hardy". ... My real point is that the secondary part of the clue – other than the definition – is meant to help the solver. The indirect anagram, unless there are virtually no alternatives, hardly ever does. He only sees it after he has got his answer by other means.

Even so, most setters that claim to be Ximenean more details pls will allow small abbreviations and contractions (as above) to be included in their clues.

Anagrind → “free” | “novel” | “comic” [...]

**Reversion**

Clues can also be reversed. While this is functionally a subset of anagrams, there are some crucial differences. Firstly the ‘directionality’ of the clue (i.e. whether it is a ‘down’ or an ‘across’) comes into effect, in determining the sorts of indicators that can form it: “turned back” may only apply to ‘across’ clues, where “taken up” may only apply to ‘down’ clues.

Further, these clues are usually taken to be ‘fairer’ game for subsequent operations to be applied to the target of the reversion. Therefore, a clue with nested wordplay such as “Fruit, odd-one-out turned back (5)” would be acceptable, where an equivalent clue “Fruit, odd-one-out messed around” would often not be seen as Ximinean.

Anagram → ReversionIndictator, Wordplay | Wordplay, ReversionIndictator

## Binary Operators

As with the unitary operator, each of the arguments of binary operators can be one or more words.

**Subtraction**

In a subtraction clue, a number of letters are removed from the target. Usually, the target is some wordplay itself, although sometimes just a string literal. The letters to be subtracted are also often the product of some sort of cluing, although this is usually fairly limited in scope (abbreviations, contractions, first letters of string literals). There are two constraints on this: all the letters from the subtraction set must be in the target, and the length of the subtraction set must be less than the length of the target.

Anagram → SubtractionPreIndictator RemovalSet (SubsidiaryIndicator) Target | RemovalSet SubtractionMidIndictator Target | RemovalSet (SubsidiaryIndicator) Target SubtractionPostIndictator

The letters in the set are thought to be removed in the order in which they’re found in order to be a properly clued wordplay. Thus “standing” with “tan” removed, gives “sanding”, whereas “ant” cannot be appropriately removed. Note though that the order in which nested clues are applied can change what the set is applied too. If we also had an anagram indicator, as in “Boy muddled standing missing trap” we can apply the muddled to standing to get “dansting” before removing “sting” to get the answer “Dan”.

**Insertion**

Insertion clues work similarly to subtraction, except in inserting the sequence of letters into

**Hidden word**

The hidden-word clue, while not particularly common, is still worthy of attention, chiefly because of the way in which it uses the surface meaning of the sentence to distract the solver's attention away from the fact that the solution is lying before his or her eyes. This type of clue involves the concealment of the solution word (or less commonly, words) within a string of other words. The fact that this has been done is usually communicated indirectly to the solver by the hidden-word pointer, which can take one of two forms: it can be a word similar in meaning to the inclusion pointers described above (such as in, within, inside, etc.), which further compli- cates solution, or it can be a phrase such as we see or appears in. The solution is, of course, a synonym of the synonym part of the clue. An example of this clue is:

'Smack which appears in East Anglian ports.(4)'

The solution to this example is 'tang', a word meaning 'smack' (in the sense of 'taste'), and which is concealed in 'easT ANGlian ports'.

HiddenWord → Indictator RemovalSet (SubsidiaryIndicator) Target | RemovalSet SubtractionMidIndictator Target | RemovalSet (SubsidiaryIndicator

Compound clue

Although all the above-mentioned clues are commonly found as entire clues in their own right, on occasion two or more may be combined to form a compound clue. Where this happens, each clue operation will produce only a part of the ultimate solution. For example:

'Fodder for the right variety of grey donkey.(3-5)'

This gives 'rye-grass' as an answer. The operations that take place in the solution of this clue are: another word for 'donkey' is ass. 'Variety' is an anagram pointer indi- cating an anagram of 'grey' and, using shorthand notation ( explained later), 'R' is an abbreviation for right. Therefore the solution is: r + yegr + ass, a syno- nym for 'fodder'.

Shorthand and abbreviations

As was demonstrated in the above section concerning compound clues, abbreviations of common words are frequently used in crossword compilation. The fact that a word is to be abbreviated is rarely stated explicitly in a clue. Instead, regular crossword solvers acquire a 'mental thesaurus' of commonly abbreviated words. As well as the more obvious shorthand forms of certain words, such as m for metre or us for America, there are 'traditional' words that regularly appear when a given letter or group of letters is required. For example, beginner, student, novice, or learner are all words that are used when the letter 1 (for learner) is required. Abstainer is used to denote tt, and point, direction, quarter, or way can be used to represent the letters that stand for the points of the compass. In addition, way can stand for the abbre- viations of street (st) and road (rd). Other frequently used shorthands include kiss (k), pound (1),ring or hole (o), and key or note (the letters a to g). The roman numerals (i, v, 1, c, d, and m) are represented by their respective numbers or by the more general words number or many.