Crossing WordNet with Crosswords, Netting Enhanced Automatic Crossword Generation

Aoife Aherne and Carl Vogel

Computational Linguistics Group, Trinity College, Dublin, Ireland,
aoaherne@cs.tcd.ie, vogel@tcd.ie

Abstract. This paper reports on a system for automatically generating and displaying crosswords from a system manager supplied database of potential clues and corresponding words that index those clues. The system relies on the lexical relations encoded in WordNet to enhance the aesthetics of the resulting crossword by making it easier to automatically identify a grid that may be populated with words and clues that have a thematic focus. The system architecture is provided in overview, as is an empirical evaluation of the results.

1 Introduction

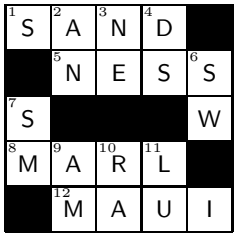
This paper builds on past work in automatic crossword generation, describing a system enhancement enabled by the availability of WordNet [1], and other freely available software resources. The paper begins by discussing the problem of automatic crossword generation, and some past developments in that area. The task is to use a database of answers and clues, based on a symmetric grid, such as the example in Fig. 1, automatically generating the grid, populating the grid with available clues, and formatting the presentation in useful ways. The grids are styled on the nature of *The New York Times* crossword puzzle, which typically have a theme. Further enhancing the aesthetics of the system, modest amounts of morphological analysis are incorporated to avoid the appearance of the answer or some morphologically related words in the clues. Complete details on the system are available; this includes user guidelines, installation and full implementation details, and inter-package interfacing.[1]

Several facts can be noted about the crossword in Fig 1. One is that it is automatically typeset in LaTeX, and another is that the answers are supplied. This is due to having interfaced the system constructed with LaTeXusing `crosswrd.sty`.[2] This package includes a binary flag which allows the user to print the solved puzzle or the open puzzle without the answers to the cues (see Appendix B and C for sample puzzles typeset w/ respect to one value for the flag or the other). Notice that a series of asterisks is used to block out instances of the answer itself or derived forms of it from the provided. clue. Another fact is that it has the unsatisfying property of numerous two letter answers. A more satisfying puzzle is given in Fig. 2

Our work on automatic crossword generation is inspired by that of [3] who developed a method for compiling fixed grids into Prolog predicates used to determine solutions (in the sense of an answer key, not about intelligent solving of preset crosswords but intelligent setting of crosswords to be solved: this is inherently a constraint solving problem in its own right) based on lists of lists corresponding to words, with shared variables constraining their interlocking. Subsequent work aimed to improve upon this [4] by beginning the optimization of grid/answer key constraint satisfaction. That work also addressed the aesthetics of grids — interestingly, high interlock in a crossword grid makes the puzzle easier for a human to solve, but more difficult for automated puzzle construction, because of the increased constraint search space. Since then, work has been

[1] See [2], available at: https://www.cs.tcd.ie/courses/csll/aoaherne0405.pdf.

[2] The author of crosswrd.sty, as distributed with the current version of MiKTeX, is Frank Mittelbach. Mittelbach extended Brian Hamilton Kelly's version of the *.sty file to ensure its compatibility with LaTeX2e, the latest version of LaTeX. Our system automaticaly constructs a text file with suitable text mark-up to automatically generate *.dvi, *.ps and *.pdf files compatible with this style file and corresponding to the automatically generated grids, solutions and clue sets. See [2] for full details.



ACROSS

1 a loose material consisting of grains of rock or coral (4)
5 a strip of land projecting into a body of water (4)
8 a loose and crumbling earthy deposit consisting mainly of calcite or dolomite; used as a fertilizer for soils deficient in lime (4)
12 the second largest of the Hawaiian Islands (4)

DOWN

7 a master's degree in science (2)
2 ***** associate degree in nursing (2)
9 modulation of the amplitude of the (radio) carrier wave (2)
3 a midwestern state on the Great Plains (2)
10 ancient hawk-headed Egyptian sun god; a universal creator (2)
4 an honorary degree in science (2)
11 a trivalent metallic element of the rare earth group; usually occurs in association with yttrium (2)
6 the compass point midway between south and west; at 225 degrees (2)

Fig. 1. Themed Crossword: Earth—Grid Containing 1 & 2 Letter 'Words'

ACROSS

1 a person or animal that is markedly unusual or deformed (5)
5 the pursuit and killing or capture of wild animals regarded as a sport (4)
6 a headlong plunge into water (4)
8 a living thing that has (or can develop) the ability to act or function independently (5)

DOWN

6 a brief swim in water (3)
2 the act of ****** as a sport (3)
3 a person trained to compete in sports (7)
7 a humorous play on words; 'I do it for the ****** of it'; 'his constant punning irritated her' (3)
4 a message whose ingenuity or verbal skill or incongruity has the power to evoke laughter (3)

Fig. 2. Themed Crossword: Sport—Grid Free of 1 & 2 Letter Words

done to both further optimize the automatically constructed Prolog code [5], and to render the system accessible by a web based interface, and further began working on the aesthetic issue of themed crossword construction with levels of solver based complexity entered as a parameter by the users. The data source for this was the 1917 Webster's Dictionary, which is available online, as it is out of copyright. Additional work [6] overhauled that system completely, providing a Web-based interface akin in functionality to many online newspaper crossword puzzles: potential for greater solver complexity provided by improving the efficiency of overall setting of the puzzle, verification for the solver of correctness of answers so far, a facility for giving up altogether. The present paper summarizes work detailed at greater length elsewhere [2], which extended the project further: enabling the server to accommodate password controlled discrimination of user types (e.g., a crossword solver in many cases should not have access to adding to or deleting from the database), thus assisting in pedagogical applications, such as vocabulary learning within subject disciplines or language learning by providing the teacher with access to do exactly that; enabling multi-platform presentation of the crosswords by automatically typesetting the puzzles using LaTeXand transformed into Postscript and PDF printable output, such that the output file can be easily printed and distributed in printed form as overnight homework assignments when users might not have access to the crossword server; and most importantly for the context of this paper, extremely enhanced control over the thematic nature of the puzzle, through relying on WordNet as a lexical resource in constructing occasionally enigmatic, but nonetheless thematic crossword puzzles.

The paper begins by describing the underlying constraint solving predicates implemented in Prolog; it then describes shortcomings of this system which are ultimately tied to the combination of complexity of grid generation and satisfying the constrains imposed by that grid on potential solution sets (and accompanying clues for the grid) [6, 7]. We reason that it is better to automatically construct the grid, and then attempt to find a solution set of answers and corresponding clues than to attempt the reverse — finding a symmetric grid to fit a list of words related by clues to a theme.

We subsequently describe enhancements to this system more fully presented elsewhere [2]. These enhancements include differential access (mediated by passwords, and password control) to database controllers and puzzle solvers. The work enhances the potential for constructing aesthetically suitable crosswords with randomly seeded grids and thematic orientation by moving beyond the original lexical resource, the 1917 *Webster's Dictionary*, through appeal to hierarchical and lateral relations encoded in WordNet. Typesetting facilities are also described, but the heart of this work is in amplifying the quantity of words & clues related to the theme, thus reducing the time to generate a puzzle and enhancing its qualitative stature as a satisfying puzzle to solve. Accordingly, the system is informally evaluated in an experiment in which participants were asked to guess the theme of the puzzle. Finally, we suggest directions for future work.

2 Automatic Crossword Generation

Consider the Grid in Fig. 3.[3] Obviously, it has no blanks, and thus describes a puzzle in which each of the four 'down' answers intersects with each of the four 'across' answers. The approach taken by Berghel to solve grids like this was to construct a Prolog Horn clause as in Fig. 4, and evaluate it with respect to a Prolog database of facts `word`, and related to a Prolog built-in predicate which relates a list of ASCII codes to an atomic word (the first argument). The predicate `word/n` is true of the list of ASCII representations of the letters in a word recorded in the Prolog database as an assertion of the fact (e.g. `word(112,97,99,101)` corresponds to the word 'pace').

An obvious computational time complexity improvement can be gained by using the 'fail-first' strategy of constraint logic programming. That is, verify the satisfiability of each of the 'word' predicates before attempting the 'name' predicates to generate the 'returned' values as the list of slot fillers. An advance of prior work [5, 6] was in controlling the order of the 'word' predicates themselves, following the 'fail-first' ethic. Other constraints were also introduced to maximize the

[3] This figure is borrowed from, and associated discussion encapsulates aspects of related work [7].

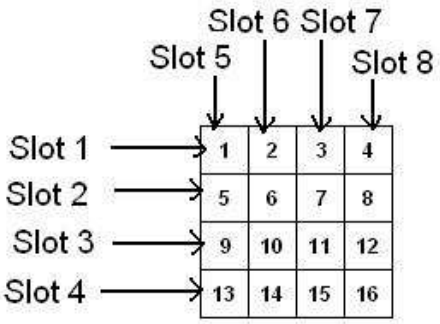


Fig. 3. A Four by Four Grid.

chances of generating a viable grid and potential solution within a reasonable amount of time: no duplicate words are allowed in the solution;[4] words that cause impossible combinations in interlinking words are disallowed; heuristics are used to evaluate particular words as candidates for appearing in the solution; backtrack to replace a prior word only if the word has an impact on the slot that is currently unfillable and hence forcing backtracking.[5] See [7] for details about some of these strategies. Here we comment just on the heuristic, as this relates to an important issue in the WordNet-enabled enhancement described later.

The 'best' choice of a word suggested for the grid is one that is not already suggested as part of the solution. The proposed word is examined for each of its letters that interlink with other slots in the grid. For each such intersecting slot, a record is kept of the number of words that are compatible as an intersecting word. The word for the original slot that has the least number of compatible intersections is most likely to cause trouble later on. Thus, that heuristic value is maximized when choosing the filler for the slot to be intersected with. This is only one such strategy that we have explored to date. Other improvements on the original [3, 4] which we take as the starting point for our work are discussed further elsewhere [5–7]. This section is intended just to give the reader a sense of the underlying constraint satisfaction problem, implemented in the context of Prolog's underlying depth-first theorem prover, in fitting clues to a grid, via words that fit the grid. Grid generation by random seed is described by [5].

3 Integrating WordNet to Improve Crossword Generation

The system whose evolution has been described so far represents an advance, but also has room for improvement. One range of issues inheres in the system presentation and overall functionality. Another is in the limitations imposed by the lexicon used as the principle data source for English crosswords. Yet another that we address here is in morphological analysis within the clues: if the clue contains the answer or a derivative form, then the puzzle becomes aesthetically less pleasing. The full system offered by [6] is very well and clearly coded and also clearly presented. The user interface provided is more than adequate. Information pertinent to crossword generation and crossword solving are intuitive and easy-to-use. However, more explicit instruction could be provided particularly if this tool was to be used in a classroom environment, as suggested in [6]. This is particularly true of information relating to the option of the user supplying their own database in text-file format.

Secondly, on a more practical issue, users cannot ever retrieve or save a copy of the crossword to complete at a later date — they are obliged to finish the crossword in one sitting. An ideal

[4] This trait is shared with the *New York Times* crossword.

[5] Note that symmetry can involve disconnected sections of the grid (recall Fig. 1); hence, backtracking across a segment that is independent of one causing problems is a waste of time.

```
solution([SLOT_1, SLOT_2, SLOT_3, SLOT_4, SLOT_5, SLOT_6]):-

        word(C1, C2, C3, C4),
        name(SLOT_1, [C1, C2, C3, C4]),

        word(C5, C6, C7, C8),
        name(SLOT_2, [C5, C6, C7, C8]),

        word(C9, C10, C11, C12),
        name(SLOT_3, [C9, C10, C11, C12]),

        word(C13, C14, C15, C16),
        name(SLOT_4, [C13, C14, C15, C16]),

        word(C1, C5, C9, C13),
        name(SLOT_5, [C1, C5, C9, C13]),

        word(C2, C6, C10, C14),
        name(SLOT_6, [C2, C6, C10, C14]),

        word(C3, C7, C9, C15),
        name(SLOT_7, [C3, C7, C9, C15]),

        word(C4, C8, C12, C16),
        name(SLOT_8, [C4, C8, C12, C16]).
```

Fig. 4. Berghel's Representation of the 4x4 Crossword Grid

solution to this problem would be the automatic generation of printable versions of the crossword at the users request. In a classroom environment this would be particularly useful, allowing an entire class to work

Thirdly, in [6] the solution to the clue very often appears in the clue text. Graham removes exact occurrences of the solution using the removeAnswer FromClue() method in the Solution class, but has not dealt with the solution in its plural, past participle or present participle forms.

A security issue, again, with respect to potential use as a language learning tool in a classroom environment is raised with access to the database. [6] did not restrict database access. Access must be restricted such that only those with sufficient privileges (e.g. "root" /"teacher") may add/modify databases on the system. Those with insufficient privileges (e.g "student") should only be able to generate and solve crossword puzzles.

The work reported in [6] includes exceptional advances in improving the execution time of the program through modification of the Prolog code that is generated. Despite the provision of a viable algorithm, the crosswords produced are very loosely related and at times totally unrelated to the theme specified. However, the fault does not lie within the algorithm employed, it lies with the dictionary database it exploits. Due to the nature of the dictionary, including its limited size, it was necessary, in order that enough words be returned to create a crossword, to select all words and definitions within which the string representation of the target word could be found. Hence, given the theme "sport", words like "transport" and indeed clues containing "transport" would be selected. Similarly the target theme "cat" returns words and clues containing the substring "cat". Dictionary size also contributes greatly to the overall result. The in corporation of a WordNet database, a lexical resource more similar to a thesaurus than to a standard dictionary, would go a long way towards solving this problem.

We address the system architecture modifications briefly in §3.1, followed by discussion of the enhancements enabled by WordNet in §3.1. It will be seen from examples in the appendices that we are still only part way to solving some of these problems (e.g. morphological analysis of clues for derivatives of the answer itself), and indeed, clue 7 down in Fig 2 demonstrates this also.

3.1 The Revised System Architecture

The system as currently implemented and tested (see §5), can be described in schematic terms as in Fig.5. Features currently available to the system manager include: population and de-population

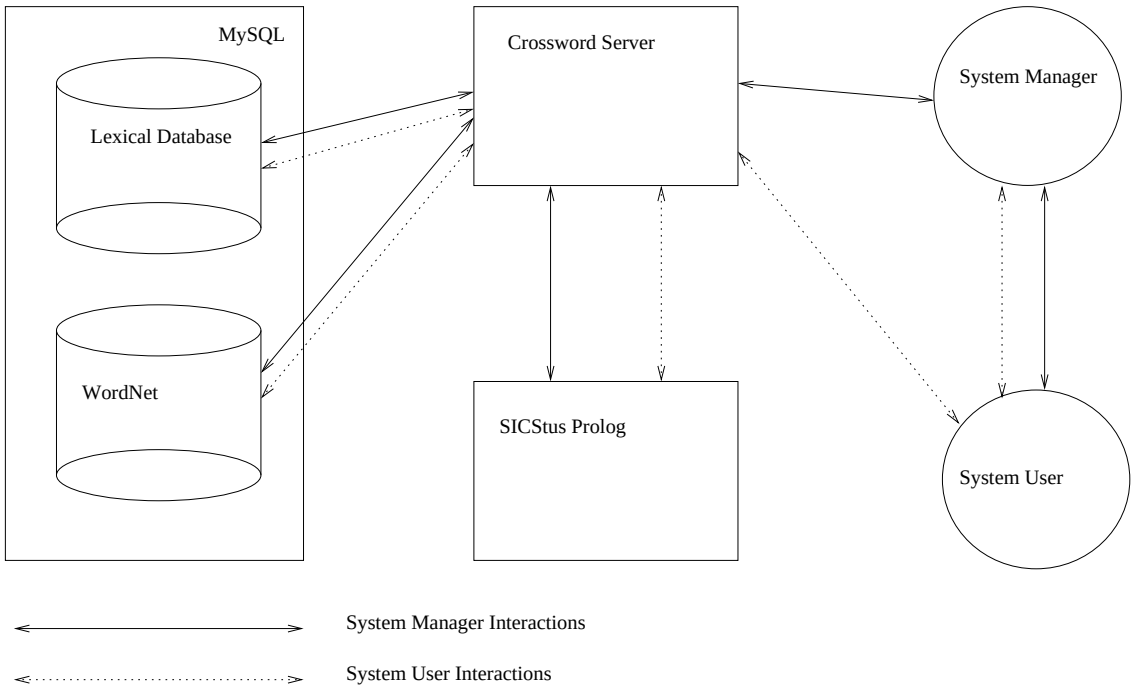


Fig. 5. Current System Architecture

of the system databases; permission or denial of user access via password management; generation of puzzles; printing of puzzles; printing of solutions. Features available to users include: generation of puzzles; online testing of interim solutions; printing of puzzles.

Currently, pending resolution of interface problems connected with institutionally available versions of MySQL, SICStus Prolog and Java, the system is implemented under linux, and is not yet available under Solaris.

Why Integrate WordNet? WordNet is a semantic web whose structure mirrors current psycholinguistic models of how lexical information is stored in the human brain. The structure of WordNet is closer to that of a thesaurus than to that of a traditional dictionary. Hence *meaning searches* as opposed to *word searches* are carried out on WordNet. This allows the retrieval of words related to a particular theme. Essentially, this allows us to bootstrap the lexical resource supplied by the original dictionary into a far larger resource.

Secondly, the source code for this linguistic tool is freely available. In addition numerous other developers have made available a variety of interfaces to WordNet including a MySQL database version, generated from the Prolog source code. These projects can be sourced from [8].

Table 1 outlines the size of the WordNet2 database and the number of relationships that exist between words and meanings. Note that each *unique string* is considered as a single word entry. Table 1 is sourced from [8].

Some Implementation Details The WordNet hierarchy forges links between words that are interchangeable in a given context. These words point to the same synset. Those words that are

Part of Speech	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	114648	79689	141690
Verb	11306	13508	24632
Adjective	21436	18563	31015
Adverb	4669	3664	5808
Totals	152059	115424	203145

Table 1. WordNet Statistics

synonymous are grouped into synsets which represent the concept described by their shared meaning. Not only is there a relationship between words in a particular synset there exists numerous links between the synsets themselves. Of particular pertinence is the hyponym and hypernym hierarchy. A hyponym synset of a particular synset contains words that are more specific with respect to an underlying theme than the original synset. A hypernym synset of a particular synset contains words that are less specific with respect to an underlying theme than the original synset.

```
Word = 'dog'
Gloss: a member of the genus Canis;
```

```
Hyponyms of 'dog': pooch, doggy
Gloss: Informal terms for dogs
```

Fig. 6. Hyponym & Hypernym Examples

```
Hypernyms of 'dog': canine, canid
Gloss: fissiped mammals with nonretractile claws[...]
```

Hence an algorithm exploiting this data structure would ensure that only words directly relevant to the theme be returned to the program.

3.2 Structural Aspects of WordNet

We take advantage of the organization in WordNet which groups words in synsets are according to meaning/concept, that it allows querying by definition, and that it posseses multiple entries for any given word supplying an ample search space from which to select words to fill the crossword grid. In essense, for every definition/meaning of a particular word or collocation there will exist a synset for that definition in WordNet. Consider the following query:

```
SELECT word, gloss, ss_type FROM 'wn_synset', wn_gloss WHERE word='sport' AND
wn_synset.synset_id = wn_gloss.synset_id
```

MySQL queries of for the puzzle is explained in more detail in [2]. In short this query searches for the word *sport* in the WordNet database and all of its associated definitions, indicating the part of speech that is assigned to the word in each instance. Word length is of primary importance to the crossword generator. The WordNet database does not contain any reference to the length of each individual word. By using the MySQL version of WordNet it became possible to easily manipulate the primary table "synset" such that the number of characters in each word was included as a tuple in the table relation. The Java methods constructed for the system queried not only for words that wre thematically related to each other but also words that pertained to the theme AND were of a length required to satisfy the grid.

The results returned to the query above are as follows:

Word	Definition	POS
sport	an active diversion...	n
sport	occupation of athletes who compete for pay	n
sport	verbal wit	n
sport	temporary summer resident of inland Maine	n
sport	organism with characteristics resulting from chromosomal alteration	n
sport	someone who engages in sports	n
sport	play boisterously	v
sport	wear/display in an ostentatious manner	v

Table 2. WordNet Results for "sport"

As can be observed in Table 2 WordNet contains eight separate entries for sport, each corresponding to a particular usage of the word. Each definition is identified by a unique numerical id. This allows the assignment of many words to the same synset and furthermore allows for the setting of relationships between synsets.

Organisation of the WordNet Hierarchy The constituent synsets of WordNet are inter-linked by relations. These relations are either semantic or lexical in nature. Lexical relations hold between word forms, semantic relations hold between word meanings. All of these relations are reflexive. What follows is a brief outline of each relation that exists within the WordNet structure taken from [8, pg 3].

The *synonymy* relation facilitates the grouping of words that are interchangeable in a particular context into the same synset. *Antonymy* is a lexical relation that concerns itself with adjectives and adverbs, positing antonymous adjectives or adverbs in opposition to each other. *Hyponymy & Hypernymy* relations organise nouns and verbs into a hierarchical structure through which individual synsets are related to each other. Hyponyms are more specific than the base word, hypernyms are more general than the base word. *Meronymy & Holonymy* relations are also used — 'part of' and 'constituted by'. *Entailments* are also included. These relations allow traversal from solution words and accompanying clues in both vertical and horizontal directions, yielding a broader set of words that may be appealed to as a solution for a grid, and potentially more enigmatically satisfying, yet thematically related crosswords. WordNet is not without its limitations, however. The freely available versions of WordNet currently most extensive and are developed in North America for English only. Since WordNet is hand collated, some concepts have been developed in greater detail than others. This is highlighted in [9]. The example given to illustrate this phenomenon is the following:

- Drogheda: site of 16th century battle
- Limerick: poem or a port city

Given that both Drogheda and Limerick are Irish towns of similar size, it is unusual that Drogheda not be considered a city like its south-western counterpart.

Why a MySQL version of WordNet2? There are many versions of WordNet available: Windows, Linus, Prolog versions etc. Other individuals have developed a WordNet2 Java interface. This is a series of Java classes that can be used to query a WordNet database. Given that the Crossword Generator is coded almost exclusively in Java it would seem logical to use the Java interface available to query the WordNet database. However, this approach was rejected on the following grounds:

- Employing the Java interface would result in the importation of a number of classes — many of which are not required by the Crossword Generator

– The methods within the Java interface do not correspond exactly to the requirements of the algorithm described in Section 4. Methods would require modification.

For these reasons, it was decided that the inclusion of a WordNet2 MySQL database and the provision of suitable methods to query this MySQL database constituted a more efficient means of integrating WordNet with the existing system.

3.3 Integrating WordNet

The following software packages were required to integrate WordNet to the exisiting system:

– WordNet2 MySQL database

This database was sourced from [10] and was developed from Prolog files used in the implementation of the Prolog version of WordNet. It is a freely available tool.

– Java.sql

Java.sql is a class within the Java hierarchy that facilitates communication between a Java program and a MySQL database. Queries can be executed on the database and the result set can be returned to the Java program for interpretation.

– Java.io

Java.io, a constituent class of the Java hierarchy makes possible "input" and "output" operations. This is of particular importance when working with server technology as information must be passed from, in this case, the program/applet to the server and from the server to the program/applet.

4 Basic Algorithm

1. User submits a valid theme
2. MySQL base query for primary synsets
3. MySQL query for hyponym synsets of each synset in 2
4. Iterate 3 five times
5. MySQL query for hypernym synsets of each synset in 2
6. Iterate 5 five times
7. Selecting a word of appropriate length from each synset returned by steps 3 & 5

4.1 Algorithm Explained

– A *Valid* Theme

As described in the Section 3.1, the semantic relations pertinent to the retrieval of words belonging to a particular theme are the "Hyponym" and "Hypernym" relations. These relations can only be applied to singular common nouns and verbs. You cannot for example search for a hyponym of the preposition "at" nor the pronoun "she". Hence the user must submit either a singular common noun or a verb to the program for the algorithm to function correctly.

– *Primary* Synsets

The term "primary" synsets is attributed to the synsets in which the subject word occurs. This is the starting point of future queries.

– *Hyponym & Hypernym* Synsets

A hyponym synset of a base synset contains words that are more "specifically" related to a particular topic that those words in the base synset. Similarly, a hypernym synset of a given synset will contain words that are more "generally" related to a particular theme than those in the initial synset.

Using the *primary synsets* as an intial starting point, steps 3 and 5 search for hyponym and hypernym synsets of the primary synsets. The first iteration of step 3 and 5 queries for hyponym and hypernym synsets of the primary synsets returned in step 2. However successive iterations of steps 3 and 5 use the last set of hyponym/hypernym synsets returned as their starting point. This results in the creation of chains of related hyponym and hypernym synsets.

– Words of *Appropriate* Length

This algorithm must satisfy two separate constraints. Firstly, the words returned must be thematically related to each other. Secondly, the words returned must be able to satisfy the constraints of the random grid generated. That is to say, only words whose length corresponds to the requirements of the grid are accepted.

4.2 An Iterative Problem

The number of synsets returned by a particular query cannot be determined in advance. This computational problem was solved in the following manner. Each query is executed once only, however the result set returned is iterated through to determine its length. Next the data structure size is modified to accommodate the result set. Finally the result set is iterated through a second time and the synsets returned are stored in the data structure.

4.3 The Magic Number 5

As can be observed in the algorithm, the searches for hyponym and hypernym synsets are iterated five times each. This is to ensure that all words associated with the submitted theme are returned. The approach adopted follows statistics related to the WordNet 1.7 hierarchy presented by [11, 9].[6] Of particular pertinence is the following:[7]

– 78.5% of synsets are "leaf-synsets"
– Maximum distance from any synset to a leaf synset is 5
– Minimum distance from any synset to a leaf synset is 2

Leaf synsets are situated at the extremes of the WordNet hierarchy. In simple terms, no other synset inherits from a leaf synset: there are no hyponym synsets of leaf synsets. 78.5% of synsets constitutes a very large majority of synsets. Of even greater importance is the fact that the maximum distance from any synset to a leaf synset is five synsets. Hence by iterating the hyponym and hypernym searches five times, full coverage is ensured.

4.4 Removing Multiple Forms of Solution From A Clue

The work presented in [6] does not successfully remove all forms of the solution from the clue. To partially address this problem, it was decided to extend the *removeAnswerFromClue* method, a member method of the *Solution class*. Use was made of the replaceAll(string this,string that) method available in the Java.String class. This method replaces occurrences of the string "this" with the string "that". The aim was to remove the solution whether it occurred at the end or in the middle of a sentence, in its plural, past particible, present particible, adverbial and superlative forms. [6] successfully removes occurrences of the solution in all parts of the clue, but does not deal with other morphological forms of the solution. An outline of the extension to the removeAnserFromClue method is given below and involves searching for representative strings:

[6] This work grew out of a need to quantify relative specificity of terms when attempting to calculate the content-based similarity of pairs of texts [12, 9].
[7] Other recent work also takes advantage of or otherwise analyzes the topology of WordNet, e.g. [13–15]

- Solution in middle of sentence
  - whitespace + answer + whitespace
- Solution at the end of a sentence
  - whitespace + answer + “.” or “!” or “?” or “:” or “;”

The occurrences of plurals, present and particible are found in a similar manner and are sought both in the middle or at the end of the sentence:

- Plural Forms in the middle of a sentence
  - whitespace + answer + “s” + whitespace
- Past & Present Particibles
  - whitespace + answer + “ed” or “d” or “ing” + whitespace
- Adverbial & Superlative Forms
  - whitespace + answer + “ly” or “ily” or “er” or “est” + whitespace

This method clearly does not take into account those words whose plural or other forms do not follow general morphological rules. In many instances exceptional cases result in a complete change to the word itself — (e.g. eat : ate) thus ensuring that the actual solution would not appear in the clue text, therefore coverage of exceptional cases is in many instances not required. Obviously, there remains room to expand on this (see the Appendices for examples of clues which slip through this minimal morphological analysis (e.g. ‘cup’ is blanked from the clue while ‘cupped’ remains in the sample of usage accompanying the definition). Possibly simpling parsing the definitions, currently used as the text of the clues, to eliminate sample usage would ameliorate the situation without recourse to deep morphological analysis, but certainly deeper analysis is necessary, both for regular patterns and exception cases. Recording exceptions needn’t be computationally expensive if a minimized trie data structure is used, or an effective caching mechanism.

5 Independent Analysis of Themed Crosswords

In order to determine whether or not the system actually succeeds in creating themed crosswords an experiment was carried out. In this experiment a total of twenty individuals, both male and female, spread between the ages of 18 and 60 were presented with a completed crossword puzzle produced by the program. Their task was to “Guess the Theme”. There was no time limit set for this task, but participants were asked to follow their “gut” instinct. Participants could respond in sentence form, single word or a couple of words that they felt best described the theme of the crossword. A total of eight crosswords were presented to the participants. The following themes were covered:

- Sport
- Water
- Music
- Pick
- Match
- Food
- Animal
- Colour

The themes were selected so as to provide a varying level of abstraction and hence difficulty. Of the themes selected sport, water, pick, match and colour can be interpreted as a noun or a verb. Food, animal and music are all nouns. It would be expected that the first group pose more problems than than the second given the extra degree of complexity. Furthermore, given that pick is primarily employed as a verb the crossword generated contains numerous verbal synonyms and

is hence more difficult to identify its theme. The crosswords presented to the participants can be viewed in Appendix A.[8]

Participants were asked to state whether or not they ever studied linguistics and if so in what capacity. After the experiment, the subjects noted which crossword they found most difficult and which they found easiest.

5.1 Interpretation Of Responses

In order to accurately analyse results a sliding scale was employed. It was decided that responses that included the exact word entered be considered as positive results. Words that could be considered as being part of the same domain were also accepted but accorded less importance. An example of a word/phrase belonging to a particular domaine is: "athletic skill", a member of the "sport" domain. Domain membership was verified using WordNet's hyponym and hypernym semantic relations. Needless to say responses that were outside of the subject domain were considered as being negative results.

5.2 Results

- Crossword 1: Sport
  - Sample of Participant Responses:
    - sport, activity, athletic skill, actions
- Crossword 2: Water
  - Sample of Participant Responses:
    - liquid, water, bodies of water, flowing, rivers & lakes, nature
- Crossword 3: Music
  - Sample of Participant Responses:
    - music, musical forms/terminology, work
- Crossword 4: Pick
  - Sample of Participant Responses:
    - actions, directions, hobbies, making decisions, sorry haven't a clue
- Crossword 5: Match
  - Sample of Participant Responses:
    - being influenced by others, ranking of respect, recognition, fitting in, dunno, changing form,imitate
- Crossword 6: Food
  - Sample of Participant Responses:
    - food, food & drink, eating & drinking
- Crossword 7: Animal
  - Sample of Participant Responses:
    - birds, insects & lizards, animals, flying creatures, small animals, birds, insects of the animal kingdom, feathers & fins
- Crossword 8: Colour
  - Sample of Participant Responses:
    - colour, colouring, in colour, colour & mediums used by artists, painting
- Overall Results

The crosswords generated to test participants views of thematic coherence of the puzzles are provided in Appendix A. Statistical results for each of the crosswords are detailed in Table 3.

[8] Appendices B and C demonstrate sample puzzles seeded by themes both with and without the solution flag set to positive.

n=20	Correct	Domain	Incorrect
Crossword 1	70%	20%	10%
Crossword 2	30%	70%	0%
Crossword 3	90%	10%	0%
Crossword 4	0%	10%	90%
Crossword 5	0%	10%	90%
Crossword 6	90%	10%	0%
Crossword 7	60%	40%	0%
Crossword 8	70%	20%	10%

Table 3. Crossword Evaluation Results

5.3 Discussion of Results

The results presented in Table 3 are largely predictable.

Themes were largely successfully chosen by the participants for crosswords 1, 2, 3, 6, 7 and 8 — only 10% of responses to crosswords 1 and 8 were entirely unrelated to the theme.

Crosswords 4 and 5 scored poorly given the high level of abstraction associated with the theme words selected: pick and match. Although the participants were not informed as to the syntactic category that the theme words belonged to, the responses returned for crosswords 4 and 5, although outlandish, were largely verbal.

Whether the participant had or does study linguistics had no bearing on the overall results. However, upon completion of the evaluation, the participants were given the correct theme for each crossword. In general, participants "understood" why a particular theme could be associated with a crossword, but linguists were naturally quicker to recognise the themes relevance to the crossword.

6 Conclusions

Obviously, WordNet has been incorporated in a wide range of applications. This paper supplies yet another application that makes crucial advantage of the lexical resource. The fact that use of WordNet is cost-free and that its source code is well documented many researchers have adopted this tool for use in a variety of areas including:

- Disambiguation of Meaning
- Semantic Tagging
- Information Retrieval
- Conceptual Identification
- Machine Translation
- Document Classification

In addition, EuroWordNet, a European project based on WordNet, has adopted WordNet's basic structure and is developing WordNets for several European languages. At present this resource is not under general release and a fee is paid depending on the number of synsets requested.

WordNet developers are currently developing READER, a system designed to aid language learners in the acquisition of vocabulary. This system is at present unavailable.

Through the integration of WordNet, the crossword generator now successfully generates "themed" crosswords. The findings of the crossword evaluation survey confirm this fact in that subjects generally converged on descriptions closely related to the seed themes, except where the seed was perhaps too abstract in nature. This should be analyzed further in terms of the location of the seeding theme with respect to its location in WordNet's topology, paying attention to the fact that the topology is not evenly distributed in its coverage. The algorithm employed for the extraction of thematically related words ensures that all relevant words are returned to the program for

consideration. Adaptation of the WordNet2 MySQL database contributes to this process. Another distinct advantage of the current system is that it opens door to easily integrating EuroWordNet and its Inter-Lingual-Index, which will allow the production of multi-lingual themed crosswords, where the clues could be presented in one language (e.g. French), with the solutions expected in the target language (e.g. English).

Careful theme selection produces interesting results. Conventional themes, like for example "language", "country", "food", "sport" produce consistently good crosswords, indicating that in a language learning environment, crosswords can be reliably produced for "popular" themes . More "original" themes, like for example "pick" or "match", provide highly-interesting, abstract crosswords which require a greater level of lateral thinking to determine their theme. These could be described as "verbally synonymous" crosswords, where light is shed on the relationship between verbs and expressions as opposed to noun-noun relationships. Other aspects of the existing system are hinted at here, but not fully detailed. For example, the Appendices show the fruit of the LaTeXinterface and modest morphological manipulation of clues. Mechanisms for differential permission to database access and update rights depending on user types and passwords are provided, but the implementation of those features is not detailed here. Finally a user's installation guide is available, as is a general user's help guide. See [2] for full details.

In further work, an obvious extension of the current program would be the inclusion of other WordNets. Given that all WordNets are built on the same underlying structure, this aspect should be relatively easy to implement. Of greater difficulty, however, would be the integration of WordNet's inter-lingual index, that links synsets in one language with corresponding synsets in another language. This would render the generation of multi-lingual themed crosswords possible. A more efficient algorithm for the generation of random symmetric grids free of one and two letter words would enhance the program's execution time. At present, grids are fully complete before they are tested for one and two letter words. Ideally, grids should be tested at regular intervals during the grid building process to remove doomed grids as early as possible. Decreasing the number of black squares present in a grid will inevitably result in fewer grids containing one and two letter words. An interesting study could involve the comparison of themed crosswords generated by the crossword generator and the New York Times themed crosswords. Not only would this highlight the quality or otherwise of the generated crosswords, it may provide an insight into which concepts require further development within WordNet. Expansion of the morphological analysis to avoid repeated word forms or elaborations of the theme in clues would also be valuable, as demonstrated by samples produced by the system included in the Appendices. Finally, overcoming the technical difficulties currently experienced with Linux operating systems the Java Virtual Machine and SICStus Prolog would make the program accessible to a wider audience.

References

1. Fellbaum, C., ed. In: WordNet: An Electronic Lexical Database. First edn. MIT Press (1998)
2. Aherne, A.: Generation of themed portable crosswords: A wordnet implementation. Centre for Computing and Language Studies, Trinity College, University of Dublin. Bachelor of Arts (Moderatorship) in Computer Science, Linguistics and a Language. Final Project Dissertation. (2002)
3. Berghel, H.: Crossword compilation with horn clauses. The Computer Journal **30** (1987) 183–88
4. Berghel, H., Yi, C.: Crossword compiler compilation. The Computer Journal **32** (1989) 276–280
5. Gibbons, A.M.: Random crossword generator with a web interface. Department of Computer Science, Trinity College, University of Dublin. Bachelor in Engineering. Final Project Dissertation. (2002)
6. Graham, Y.: Random crossword generator with a web user-interface. Centre for Computing and Language Studies, Trinity College, University of Dublin. Bachelor of Arts (Moderatorship) in Computer Science, Linguistics and a Language. Final Project Dissertation. (2003)
7. Graham, Y., Vogel, C.: Computer construction of crossword puzzles using horn clauses and constraint programming (2005)
8. WordNet2: WordNet2 Reference Manual. Princeton University, http://wordnet.princeton.edu/doc. (2004)
9. Devitt, A.: Methods for Meaningful Text Representation and Comparison. PhD thesis, Trinity College Dublin (2004)

10. AndroidTechnologies: WordNet2 MySQL Database. Android Technologies, http://www.androidtech.com/html/wordnet-MySQL-20.php. (2005)

11. Devitt, A., Vogel, C.: The topology of wordnet: Some metrics. In: Second International Wordnet Conference. (2004) 106–111 Brno, Czech Republic.

12. Devitt, A., Vogel, C.: Using wordnet hierarchies to pinpoint differences in related texts. In: Proceedings of EUROLAN 2003, Ontologies, and Information Extraction International Workshop. (2003) 37–44 Bucharest, Romania.

13. Agirre, E., Alfonseca, E., de Lacalle, O.L.: Approximating hierarchy-based similarity for wordnet nominal synsets using topic signatures. In: Second International Wordnet Conference. (2004) 15–22 Brno, Czech Republic.

14. Farreres, J., Gibert, K., Rodrigues, H.: Towards binding spanish senses to wordnet senses through taxonomy alignment. In: Second International Wordnet Conference. (2004) 259–264 Brno, Czech Republic.

15. Teich, E., Fankhauser, P.: Wordnet for lexical cohesion analysis. In: Second International Wordnet Conference. (2004) 326–331 Brno, Czech Republic.

A System Generated Crosswords Used for Evaluation

– Themed Crossword: 1

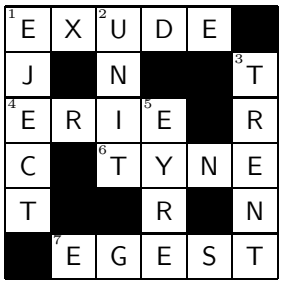
ACROSS

1 a person or animal that is markedly unusual or deformed (5)
5 the pursuit and killing or capture of wild animals regarded as a sport (4)
6 a headlong plunge into water (4)
8 a living thing that has (or can develop) the ability to act or function independently (5)

DOWN

6 a brief swim in water (3)
2 the act of rowing as a sport (3)
3 a person trained to compete in sports (7)
7 a humorous play on words; 'I do it for the ?????? of it'; 'his constant punning irritated her' (3)
4 a message whose ingenuity or verbal skill or incongruity has the power to evoke laughter (3)

– Themed Crossword: 2



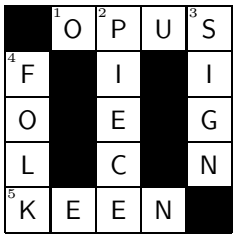
ACROSS

1 release (a liquid) in drops or small quantities; ‘ ********** sweat through the pores’ (5)
4 the 4th largest of the Great Lakes; it is linked to the Hudson River by the New York State Barge Canal (4)
6 a river in northern England that flows east to the North Sea (4)
7 eliminate from the body; ‘Pass a kidney stone’ (5)

DOWN

1 eliminate (substances) from the body (5)
2 an assemblage of parts that is regarded as a single entity; ‘how big is that part compared to the whole?’; ‘the team is a ******** ’ (4)
5 a shallow salt lake in south central Australia about 35 feet below sea level; the largest lake in the country and the lowest point on the continent (4)
3 a river in central England that flows generally notheastward to join with the Ouse River and form the Humber (5)

– Themed Crossword: 3



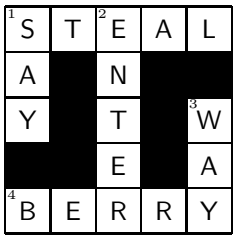
ACROSS

1 a musical work that has been created; ‘the composition is written in four movements’ (4)
5 a funeral lament sung with loud wailing (4)

DOWN

4 people descended from a common ancestor; ‘his family has lived in Massachusetts since the Mayflower’ (4)
2 a musical work that has been created; ‘the composition is written in four movements’ (5)
3 any communication that encodes a message; ‘signals from the boat suddenly stopped’ (4)

– Themed Crossword: 4



ACROSS

1 take without the owner's consent; 'Someone stole my wallet on the train'; 'This author stole entire paragraphs from my dissertation" (5)

4 pick or gather berries; 'We went ******** in the summer' (5)

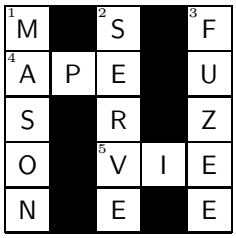
DOWN

1 express in words; 'He said that he wanted to marry her'; 'tell me what is bothering you'; 'state your opinion'; 'state your name' (3)

2 to come or go into; 'the boat entered an area of shallow marshes' (5)

3 doing as one pleases or chooses; 'if I had my ******* ' (3)

– Themed Crossword: 5



ACROSS

4 imitate uncritically and in every aspect; 'Her little brother **** her behavior' (3)

5 compete for something; engage in a contest; measure oneself against others (3)

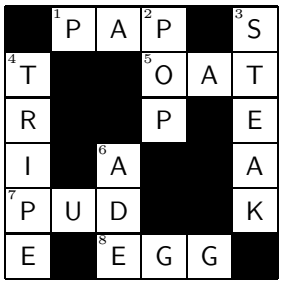
DOWN

1 a member of a widespread secret fraternal order pledged to mutual assistance and brotherly love (5)

2 be sufficient; be adequate, either in quality or quantity; 'A few words would answer'; 'This car suits my purpose well'; 'Will $100 do?'; 'A 'B' grade doesn't suffice to get me into medical school'; 'Nothing else will ********** ' (5)

3 a friction match with a large head that will stay alight in the wind (5)

– Themed Crossword: 6



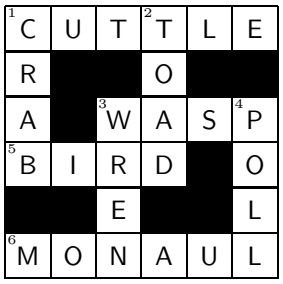
ACROSS

1 a diet that does not require chewing; advised for those with intestinal disorders (3)
5 seed of the annual grass Avena sativa (spoken of primarily in the plural as '********') (3)
7 (British) the dessert course of a meal (' ****** ' is used informally) (3)
8 oval reproductive body of a fowl (especially a hen) used as food (3)

DOWN

4 lining of the stomach of a ruminant (especially a bovine) used as food (5)
6 a sweetened beverage of diluted fruit juice (3)
2 a sweet drink containing carbonated water and flavoring; 'in New England they call sodas tonics' (3)
3 a slice of meat cut from the fleshy part of an animal or large fish (5)

– Themed Crossword: 7



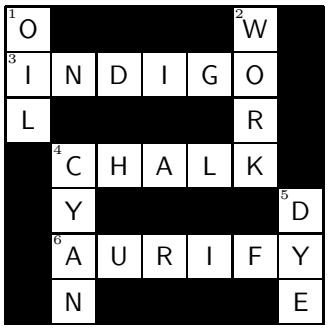
ACROSS

1 ten-armed oval-bodied cephalopod with narrow fins as long as the body and a large calcareous internal shell (6)
3 social or solitary hymenopterans typically having a slender body with the abdomen attached by a narrow stalk and having a formidable sting (4)
5 warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings (4)
6 brilliantly colored pheasant of southern Asia (6)

DOWN

1 decapod having eyes on short stalks and a broad flattened carapace with a small abdomen folded under the thorax and pincers (4)
3 and of several small active brown birds of the northern hemisphere with short upright tails; they feed on insects (4)
2 any of various tailless stout-bodied amphibians with long hind limbs for leaping; semi-aquatic and terrestrial species (4)
4 a tame parrot (4)

– Themed Crossword: 8



ACROSS

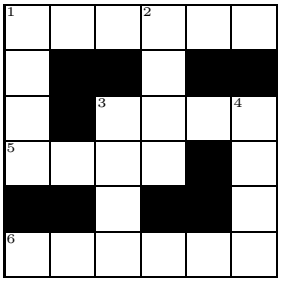
3 a blue dye obtained from plants or made synthetically (6)
4 a pure flat white with little reflectance (5)
6 turn golden (6)

DOWN

1 ******** paint used by an artist (3)
4 a blue-green that is one of the primary pigments (4)
2 have and exert influence or effect; 'The artist's ******** influenced the young painter'; 'She ******** on her friends to support the political candidate' (4)
5 a usually soluble substance for staining or coloring e.g. fabrics or hair (3)

B Sample Themed Crosswords without Answers

1. Themed Crossword: Country



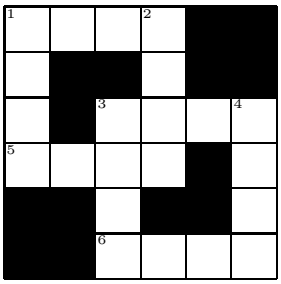
ACROSS

1 the capital and largest city of Zambia (6)
3 a landlocked republic in northwestern Africa; achieved independence from France in 1960; ******** was a center of West African civilization for more than 4,000 years (4)
5 a country of southeastern Asia that extends southward along the Isthmus of Kra to the Malay peninsula; 'Thailand is the official name of the former ******** ' (4)
6 a Scandinavian kingdom in the eastern part of the Scandinavian Peninsula (6)

DOWN

1 a mountainous landlocked communist state in southeastern Asia; achieved independence from France in 1949 (4)
3 the capital of Maldives in the center of the islands (4)
2 the biblical name for ancient Syria (4)
4 a theocratic islamic republic in the Middle East in western Asia; ******** was the core of the ancient empire that was known as Persia until 1935; rich in oil; involved in state-sponsored terrorism (4)

2. Themed Crossword: Language



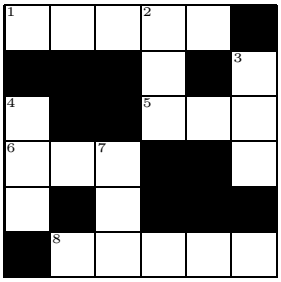
ACROSS

1 the words of something written; 'there were more than a thousand words of ******** '; 'they handed out the printed ******** of the mayor's speech'; 'he wants to reconstruct the original ******** ' (4)
3 the language of the nomadic Lapp people in northern Scandinavia and the Kola Peninsula (4)
5 Kamarupan languages spoken in western Burma and Bangladesh and easternmost India (4)
6 a family of American Indian languages spoken by Mayan peoples (4)

DOWN

1 the dialect of Albanian spoken in southern Albania and in areas of Greece and Italy (4)
3 reading or glancing through quickly (4)
2 a branch of the Tai languages (4)
4 a dialect of the Chiwere language spoken by the ******** people (4)

3. Themed Crossword: Shape



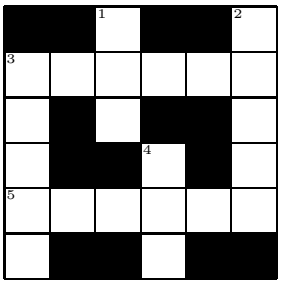
ACROSS

1 a round shape formed by a series of concentric circles (5)
5 form a knot or bow in; ' ****** a necktie' (3)
6 a continuous portion of a circle (3)
8 form metals with a swage (5)

DOWN

4 an angular shape characterized by sharp turns in alternating directions (3)
7 form into the shape of a ****** ; 'She cupped her hands' (3)
2 a groove or furrow (especially one in soft earth caused by wheels) (3)
3 a part of a forked or branching shape; 'he broke off one of the branches'; 'they took the south fork" (3)

4. Themed Crossword: Fruit



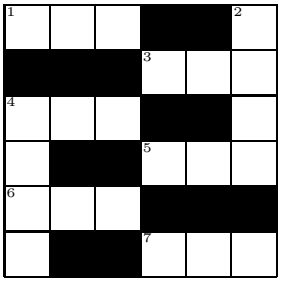
ACROSS

3 pod of the peanut vine containing usually 2 nuts or seeds; 'groundnut' and 'monkey nut' are British terms (6)
5 the small hard nutlet of a drupe or drupelet; the seed and the hard endocarp that surrounds it (6)

DOWN

3 any of various juicy purple- or green-skinned fruit of the genus Vitis; grow in clusters (5)
1 a several-seeded dehiscent fruit as e.g. of a leguminous plant (3)
4 the fruit or seed of a ****** plant (3)
2 dried plum (5)

5. Themed Crossword: City



ACROSS

1 a city in northwestern Iran; a place of pilgrimage for Shiite Muslims (3)
3 a city in the European part of Russia (3)
4 a port in southern Lebanon on the Mediterranean Sea; formerly a major Phoenician seaport famous for silks (3)
5 the former capital and 2nd largest city of Brazil; chief Brazilian port; famous as a tourist attraction (3)
6 the capital and largest city of Japan; the economic and cultural center of Japan (3)
7 a city in north central Morocco; religious center (3)

DOWN

4 a city in northeastern Egypt at the head of the Gulf of ******** and at the southern end of the ******** Canal (4)
2 a port in western Israel on the Mediterranean; incorporated into Tel Aviv in 1950 (4)

C Corresponding Sample Themed Crosswords with Answers

1. Themed Crossword: Country

ACROSS

1 the capital and largest city of Zambia (6)
3 a landlocked republic in northwestern Africa; achieved independence from France in 1960; ******** was a center of West African civilization for more than 4,000 years (4)
5 a country of southeastern Asia that extends southward along the Isthmus of Kra to the Malay peninsula; 'Thailand is the official name of the former ******** ' (4)
6 a Scandinavian kingdom in the eastern part of the Scandinavian Peninsula (6)

DOWN

1 a mountainous landlocked communist state in southeastern Asia; achieved independence from France in 1949 (4)
3 the capital of Maldives in the center of the islands (4)
2 the biblical name for ancient Syria (4)
4 a theocratic islamic republic in the Middle East in western Asia; ******** was the core of the ancient empire that was known as Persia until 1935; rich in oil; involved in state-sponsored terrorism (4)

2. Themed Crossword: Language

ACROSS

1 the words of something written; 'there were more than a thousand words of ******** '; 'they handed out the printed ******** of the mayor's speech'; 'he wants to reconstruct the original ******** ' (4)
3 the language of the nomadic Lapp people in northern Scandinavia and the Kola Peninsula (4)
5 Kamarupan languages spoken in western Burma and Bangladesh and easternmost India (4)
6 a family of American Indian languages spoken by Mayan peoples (4)

DOWN

1 the dialect of Albanian spoken in southern Albania and in areas of Greece and Italy (4)
3 reading or glancing through quickly (4)
2 a branch of the Tai languages (4)
4 a dialect of the Chiwere language spoken by the ******** people (4)

3. Themed Crossword: Shape

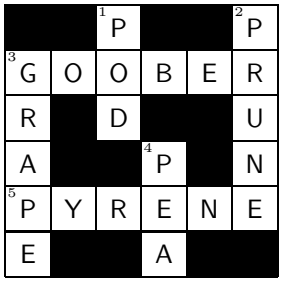
ACROSS

1 a round shape formed by a series of concentric circles (5)
5 form a knot or bow in; ‘ ****** a necktie’ (3)
6 a continuous portion of a circle (3)
8 form metals with a swage (5)

DOWN

4 an angular shape characterized by sharp turns in alternating directions (3)
7 form into the shape of a ****** ; ‘She cupped her hands’ (3)
2 a groove or furrow (especially one in soft earth caused by wheels) (3)
3 a part of a forked or branching shape; ‘he broke off one of the branches’; ‘they took the south fork’ (3)

4. Themed Crossword: Fruit



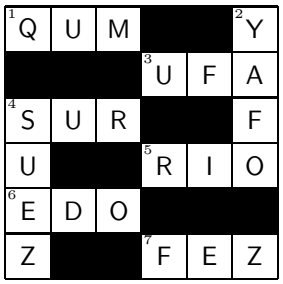
ACROSS

3 pod of the peanut vine containing usually 2 nuts or seeds; ‘groundnut’ and ‘monkey nut’ are British terms (6)
5 the small hard nutlet of a drupe or drupelet; the seed and the hard endocarp that surrounds it (6)

DOWN

3 any of various juicy purple- or green-skinned fruit of the genus Vitis; grow in clusters (5)
1 a several-seeded dehiscent fruit as e.g. of a leguminous plant (3)
4 the fruit or seed of a ****** plant (3)
2 dried plum (5)

5. Themed Crossword: City



ACROSS

1 a city in northwestern Iran; a place of pilgrimage for Shiite Muslims (3)
3 a city in the European part of Russia (3)
4 a port in southern Lebanon on the Mediterranean Sea; formerly a major Phoenician seaport famous for silks (3)
5 the former capital and 2nd largest city of Brazil; chief Brazilian port; famous as a tourist attraction (3)
6 the capital and largest city of Japan; the economic and cultural center of Japan (3)
7 a city in north central Morocco; religious center (3)

DOWN

4 a city in northeastern Egypt at the head of the Gulf of ******** and at the southern end of the ******** Canal (4)
2 a port in western Israel on the Mediterranean; incorporated into Tel Aviv in 1950 (4)