A Pragmatic Analysis of Crossword Puzzle Difficulty

by

Jocelyn Adams

Professor Ted Fernald, Advisor

A thesis submitted in partial fulfillment
of the requirements for the
Degree of Bachelor of Arts
in Linguistics

Swarthmore College

Swarthmore, Pennsylvania

December 2014

Table of Contents

Table of Contents i

Abstract iii

Acknowledgements iv

1 Introduction 1

2 Background 3

2.1 Corpus . . . 3
2.2 Notation . . . 4
2.3 Puzzle rules and conventions . . . 5
2.4 Crosswords as communication . . . 7
2.5 Prior research . . . 11

3 Difficulty factors 13

3.1 Cultural vs. lexical knowledge . . . 13
3.2 Ambiguity . . . 15
3.3 Obscurity . . . 17
3.4 Underspecificity . . . 17
3.5 Difficult answers . . . 18
3.6 Trickiness . . . 19
3.7 Factor interdependencies . . . 21

4 Results 23

4.1 Cultural vs. lexical entries . . . 23
4.2 Ambiguity . . . 24
4.3 Obscurity . . . 27

4.4 Underspecificity 30
4.5 Trickiness 31
4.6 Analysis of results 32

5 Conclusion **34**
5.1 Pragmatics of crossword puzzles 34
5.2 Future work 35

References **36**

Abstract

The language of crosswords is intended to trick readers to some degree while still eventually leading them to the correct answer. While crosswords are governed by fairly strict rules dictating word length, grid structure, and clue content, clues are often purposefully ambiguous and provide a level of context that would be insufficient for normal conversation. In many newspapers, including the *New York Times*, daily puzzles become more difficult as the week progresses. For a puzzle later in the week, the creator may assume that solvers are seasoned enough to know the basic crossword patterns and can begin to play with conventions.

This thesis explores how the methods by which crossword creators write difficult clues mirror the Gricean maxims. It breaks down crossword entries into several difficulty factors: requirement of cultural knowledge, obscurity, ambiguity, underspecificity, and trickiness. Each of these factors is explored at length and then analyzed quantitatively using a large data set of entries from *New York Times* crosswords.

My findings indicate that many of these factors are interrelated. The distinction between puzzles at each day of the week is more nuanced than some abstract, monotonically increasing "difficulty" function. Some of the factors turn out to not be highly correlated to difficulty at all. In some cases, clues later in the week appear to favor entertainment value over sheer difficulty level.

Additionally, I show how pragmatic theories can be applied to a highly restricted and unnatural form of communication. My work extends similar research into the use of Gricean maxims in jokes and advertisements. This thesis combines natural language processing approaches on a full corpus with detailed manual analysis, demonstrating that computational methods can be used to address an area such as pragmatics.

Acknowledgements

Thank you to my advisor, Ted Fernald, to my student readers, Jacob Collard and Tamsin True-Alcalá, to my faculty reader, Brook Lillehaugen, to Nathan Sanders for creating and sharing this LaTeX template, and to Jane C. Betts and Jack Adams for acquainting me with the joy of puzzling.

Chapter 1
Introduction

In 1975, H. Paul Grice introduced the idea of a Cooperative Principle in casual conversation. It consists of four maxims that guide productive communication and govern the assumptions that people may make about what is implied through speech. The main purpose behind the Cooperative Principle is to make conversation easy and efficient (Grice 1975).

The conversation between crossword creator and crossword solver is not meant to be easy. The easiest crossword puzzle would tell you exactly what to write in every square, but that would be no fun at all! Instead, the creator and solver have a different, unique agreement: the creator will write interesting, educational, and playful clues and the solver will think critically about them. Because of the additional structural information that a crossword puzzle supplies, the clues need not be fully straightforward. The creator can exploit exactly the kinds of ambiguity, obscurity, and underspecificity that Grice warns against in order to create an entertainingly difficult experience for the solver. Crosswords are published at varying difficulty levels to allow solvers to commit as fully as they please to the crossword "game." At higher levels, even more attention, creativity, and expertise is required of both the creator and the solver.

Gricean maxims are intended to explain spoken conversation and pertain specifically to utterances. Crossword clues are neither spoken nor are they necessarily statements. A clue alone does not usually entail anything about the world without its answer. It is up to the solver to come up with the other half of the crossword entry "statement" that makes the proposition true. In order to allow solvers to complete the entry, the clue must be cooperative with its implicatures. In the absence of a proper entailment, each clue's implicatures are even more crucial.

Prior research has explored how Gricean maxims explain implicatures in jokes and advertising. In this thesis, I explore the possibility that the maxims are present in crosswords

as well. Within the context of crossword clues, the Cooperative Principle is relaxed somewhat — in fact, lexical ambiguity is encouraged — but elements of it must remain for the crossword to be fair. A crossword entry should not imply anything false, for example.

Consider one of the most famous crossword clues of all time, published in the *New York Times* on Election Day in 1996. The clue in question was "Lead story in tomorrow's newspaper(!), with [ELECTED]," and the creator constructed the puzzle in such a way that the answer could be either BOBDOLE or CLINTON (Reynaldo 2007). Having multiple possible answers was a clear breach of both the Cooperative Principle and the rules of crossword building, and plenty of readers were outraged when they initially believed the creator had been so audacious as to declare a winner early. However, this entry was also extremely clever and well-executed and has become famous as an example of the kind of creativity and humor found in crosswords.

This thesis explores how the conversation that transpires between crossword creator and solver evolves throughout the levels of difficulty. In Chapter 2, I introduce the language of crossword puzzles and relevant literature. In Chapter 3, I describe the factors that contribute to the difficulty of crossword clues and provide examples. Then in Chapter 4, I apply quantitative analysis to a set of clues and answers from *New York Times* crossword puzzles to measure the effects of the various difficulty factors.

Chapter 2
Background

2.1 Corpus

As part of a tool to help crossword creators find clues for their puzzles, Matt Ginsberg compiled a database called Cluer that collected hundreds of thousands of clues and answers from several popular newspaper crosswords spanning from the 1950's to 2012. Each entry in the free database contains clue, answer, newspaper of publication, day of the week (i.e., perceived difficulty), and year. For the purposes of this thesis, I restrict my research to the *New York Times* (NYT) puzzle, which is held by many as the "gold standard of American crosswords" (Reynaldo 2007). For further consistency, I only examine clues from puzzles published during Will Shortz's tenure as editor. Shortz has been serving as editor since 1993. The editor makes all of the executive decisions about which puzzles to publish and where they should rank on the weekday scale. In fact, Shortz alters about 50% of the clues in any given puzzle to adjust its difficulty (Reynaldo 2007).

The full corpus used here contains 738,110 <clue, answer> pairs from between 1993 and 2012. This includes 468,256 distinct clues and 110,319 distinct answers. I will often use the word "entries" to refer to distinct <clue, answer> pairs. Unless otherwise noted, any entry used as an example in this thesis is an actual entry that has been used in one or more New York Times crossword puzzles, retrieved from the Cluer database (Ginsberg 2012).

I wrote programs to process and analyze the data in Python. I used the Official Scrabble Dictionary as a definitive set of common English types when examining whether cultural or lexical knowledge was required to solve a clue (Scrabble 1995). In analyzing ambiguity of clues, I used the WordNet system to assess the number of possible senses of a given word (Princeton University 2010). I also consulted several corpora of English to assess how obscure certain words are, and used the Natural Language Toolkit (NLTK) library for

processing (Bird et al. 2009).

2.2 Notation

Within this thesis, <clue, answer> pairs in examples are written as "clue → ANSWER". All punctuation and capitalization in the clue is as it originally appeared. Answers are always written in small caps, e.g. WORD. In some writing about crosswords, answers are written in full capital letters, but as clues occasionally contain fully capitalized words, I use small caps to avoid confusion. Accented letters, apostrophes, word boundaries, or other characters that may occur with an answer word or phrase in its common form are not included. For example, ALA in example A would more accurately be written as À LA in context, but it is left in its simplified character form so as not to differentiate it from the identical answer in example B.

A. Pie ___ mode → ALA

B. Yellowhammer State: Abbr. → ALA

This notation does not imply a one-to-one relationship between the clue and answer; that is, the same clue could have been used for different answers as well.

To establish a ground-truth difficulty value for an entry, it is possible to calculate an "average" day of publication. The dataset includes weekday of publication for each instance of an entry, recorded as an integer between 1 and 6. Since many <clue, answer> pairs have occurred in multiple puzzles, it is possible to take the average of these publication days. Entries with a higher mean day of publication are intended to be harder to solve, on average. Sometimes this difficulty score is relevant to an example; if so, it will be written in parentheses immediately following the entry.

A. Plain to see → OVERT (2.5)

This approach makes the assumption that difficulty increases linearly as the week progresses. While there is no way to definitively prove this, it seems fair to assume that the editors aim for a linear progression of difficulty to allow solvers to improve their skills gradually over time. Assuming a linear relationship is not therefore problematic.

What is more potentially troublesome is the fact the day of publication is based on the entire puzzle, not just one entry at a time. It is entirely possible, and in fact likely, for a

Friday or Saturday puzzle to have several straightforward entries mixed in to allow solvers to gain some traction. Similarly, an easy puzzle is likely to have a couple of difficult entries placed where a solver could easily use intersecting words to get the answer. A more precise analysis could consider the difficulty of an entire puzzle, but such a strategy would likely abstract away too much from the low-level linguistic features that contribute to difficulty of individual clues. Furthermore, the data set is large enough that I do not expect intra-puzzle variations to have a considerable effect on the kinds of patterns that emerge among clues from different days.

2.3 Puzzle rules and conventions

The *New York Times* has well-defined rules governing its crossword puzzles. The puzzles are submitted by individual creators, then chosen and edited by Will Shortz, who has been editor since 1993. All Monday through Saturday puzzles are 15 x 15. Typically the pattern of black squares (word boundaries) is rotationally symmetrical, but occasionally it may be reflectionally symmetrical if the theme requires. All answers must be at least 3 letters, and every letter must be part of an answer both Across and Down. The puzzle must be fully interlocking, so no parts of the puzzle are completely isolated (Reynaldo 2007).

Beyond the structural requirements, there are certain patterns that entries are expected to follow. These are not hard and fast rules, but especially in easier puzzles it is highly unlikely to see any deviation from the norm. Answers may consist of multiple words, abbreviations, and words that would typically be written with accents, apostrophes, or diacritic marks. Marking multiple-word answers is not required.

The most common type of entry is a pair of syntactically and semantically interchangeable phrases as clue and answer. Syntactically, the clue and answer are each a phrase of the same type (e.g. adverb phrase or noun phrase) with matching person and number. It should be possible to replace the clue phrase in an arbitrary sentence with the answer phrase without creating an ungrammatical sentence. Semantically, the clue and answer should also refer to the same entity or overlapping sets of entities.

A. Places of refuge → LAIRS

B. In a talented manner → ABLY

Another fairly common type of clue is a complete sentence containing a pronoun, which is to be replaced by the answer.

C. It makes the heart grow fonder → ABSENCE

D. Van Gogh threatened him with a razor blade → GAUGUIN

Similarly, if the phrase containing the answer is a title, quote, or idiom, the clue may contain a blank indicating where the answer appears in the phrase. A more general clue clarifying the intended meaning of the entire phrase may follow in parentheses. This type of clue is called fill-in-the-blank.

E. P. C. Wren novel "Beau ___" → GESTE

F. ___ few rounds (spar) → GOA

By convention, if an answer is abbreviated, the clue should indicate it by including an abbreviated word or otherwise acknowledging the abbreviation.

G. New Deal pres. → FDR

H. Homecoming attendee, for short → ALUM

Answers in a foreign language must also be indicated in some way, either directly or more often by referring to a name or place associated with that language.

I. Girlfriend, in Grenoble → AMIE

J. Juan's aunt → TIA

Often, tone and style of clue and answer also tend to match. An informally worded clue indicates the answer may be slang.

K. 100 smackers → CNOTE

L. Marvy → FAB

If the words in a clue are in quotes, the answer is likely to be an utterance with a similar meaning to the quoted clue.

M. "Hang in there!" → BESTRONG

N. "Holy cow!" → EGAD

A similar type of clue involves an utterance in square brackets. The answer to this type of clue is likely an action representing the sentiment suggested by the bracketed clue.

O. [I don't care] → SHRUG

P. [How dare you!] → SLAP

Clues that involve a play on words are particularly tricky and thus are conventionally marked with a question mark (?) at the end. For an in-depth analysis of these types of clues, see the section on trickiness in Chapter 3.

2.3.1 Themes

Many puzzles include a theme that unites several of the longer answers. Typically, Friday and Saturday puzzles (the most difficult) do not have a theme. This is sometimes considered a factor in their difficulty. The larger, middle-difficulty Sunday puzzles always have a theme. Often the theme is quite creative and involves puns, wordplay, or even occasionally putting multiple letters in one square (known as a "rebus" puzzle). This thesis does not consider how a theme impacts the difficulty of a puzzle, nor does it explore the patterns and conventions within theme clues and answers. For consistency, clues and answers from Sunday puzzles are not included in the dataset. For an analysis of crossword themes from a cognitive science perspective, see Nickerson (2011).

2.4 Crosswords as communication

It is not immediately clear that crossword puzzles are a valid form of communication at all. They involve language, certainly, but are not discourse in any traditional sense. Though the argument can be made that crosswords are educational, they are primarily intended as an entertaining pastime, not a source of crucial or new information.

Regardless of the very low stakes of the information communicated, a crossword puzzle does communicate. It communicates for the sake of communicating — to prove that two individuals who have never met can decode one another's understanding of language. Every crossword has two participants: the creator of the puzzle and the solver. The creator cooperates with the solver through the puzzle's clues, in a manner somewhat resembling a teacher eliciting an answer on an exam. Clues must be constructed according to the rules put forward by the newspaper. They must be factual and follow expected patterns, and

be accessible to a wide range of readers without much context. Ideally, they will also be entertaining. While the language is certainly unnatural, the context of a crossword puzzle creates expectations that pragmatically inform the solver's understanding of its clues. When a crossword clue is solved, the creator has succeeded in predicting the solver's linguistic process. The ultimate communication of any crossword is the simple fact that such a feat is possible.

2.4.1 Pragmatic theory outside of discourse

There is ample precedence for studying aberrant forms of language through pragmatic theory. Gricean maxims have been used to analyze many different kinds of language besides utterances, including the language of advertisements. The kinds of phrases used in advertising are somewhat analogous to crossword clues in terms of their brevity and indirectness. For advertisements to be successful, they must balance being direct and oblique by appealing to both reason and humor (Simpson 2001). This often necessitates flouting the maxims of manner and relevance.

For example, Weiner and De Palma investigate the pragmatic features of certain riddles as a "kind of back door approach to understanding the rules of natural language"(Weiner and De Palma 1993). They study riddles that depend on lexical ambiguity specifically, and find that most of these riddles succeed by purposefully appealing to one sense of a word when the answer depends on a different sense.[1] Because this purposeful ambiguity is generally prohibited by rules of discourse such as the Cooperative Principle, the answer comes as a clever surprise. One of the paper's most relevant observations is the importance that the ambiguity in a riddle be context-independent, so it "works" even if it comes out of nowhere. Riddles are quite similar to crossword entries in many ways, as both share the purpose of eliciting a single brief answer, though not all crossword clues require disambiguation. Most are far more straightforward, but still necessitate context-independence.

Another highly regulated and unnatural context in which pragmatics have been applied is the area of word problems in mathematics education. Math word problems are somewhat similar to crossword clues in that they are meant to elicit a specific answer from the reader. Word problems are longer than crossword clues, and often are set up to be as unambiguous as possible so as to allow only one possible answer. They also, of course, tend to require computation of some kind. Unlike riddles and advertisements, word problems follow the

[1] For example, "What has an eye but cannot see?" "A needle." The riddle sets up the vision-related sense of *eye* by using the related word *see*, but the answer subverts this expectation.

Cooperative Principle dutifully. Their language is artificially constructed to mimic a system of equations, with exactly enough information presented clearly so that the language does not interfere with the math. The only Gricean maxim that word problems consistently flout is quality, in fact, because their content is almost always completely imagined (Gerofsky 1996). Word problems are not meant to be believed, and thus any truth value (beyond the truth of mathematical equivalence) is either not present or irrelevant.

In a way, crossword puzzles are a combination of straightforward word problem-like entries and more clever riddle-like entries. Each of these presents their own pragmatic peculiarities.

2.4.2 Felicity conditions in crossword clues

The equivalence relation between clue and answer means that unlike a mathematics word problem, any crossword entry has some entailment. A large subset of crossword entries, as discussed in Chapter 3, fall into the category of "lexical" entries, so that rather than making a proposition about the world as a whole, they make a proposition about the semantic equivalence of two words or phrases. For example, example A below implies that *astringent* means roughly the same thing as *harsh*. Even as an entire entry, it does not imply anything about the property of astringency or what entities might have that property.

A. Astringent → HARSH

Example B demonstrates an entailment about the world. The clue and answer paired together imply that there exist shoes made of suede. Based on how crosswords are understood, the relationship between clue and answer is not necessarily exact equivalence. It is sufficient for the clue and answer to overlap or be subsets of each other. Note that in this case, the clue does not entail that all shoes are made of suede or that all suede is made into shoes, but merely that some suede is also shoe material. This is a somewhat delicate balance. For example, some shoes are also made of duct tape, but intuitively *shoe material* would be a terrible clue for DUCTTAPE. The overlap must be extensive or salient enough, and there is not a clear guideline for how that is defined beyond what feels right to the creator.

B. Shoe material → SUEDE

Neither of the clues in A or B, taken alone, is a statement. These phrases do not entail anything, but they prompt an answer that will complete an entailment. In this way, crossword clues are similar to questions. Despite the fact that a question does not entail

anything and therefore cannot be false, it may contain presuppositions that can cause it to be infelicitous. These felicity conditions tend to arise from the hypothetical presence of a response to the question (Searle 1969). For example, consider the following exchange.

Speaker 1: Is your sister in town?
Speaker 2: Yes.

The exchange entails that Speaker 2 has a sister. This entailment would remain even if Speaker 2 had answered "No." In this case, Speaker 1's question presupposes that Speaker 2 has a sister. In the event that this presupposition is false, it would be an infelicitous question. Thus, the existence of Speaker 2's sister is a felicity condition on the question.

C. John Updike story set in a grocery store $\rightarrow$ AANDP

The full entry in example C entails that "A and P" is a John Updike story set in a grocery store. Again, the clue alone does not entail anything; however, it does contain a couple of presuppositions. For example, the clue in C presupposes that a story exists which is written by, or perhaps about, John Updike. If not, the clue would have no true answer, making it infelicitous.

The context of a crossword puzzle clue allows for the further presupposition that the referent of John Updike is the famous author, and not someone else who happens to be named John Updike but is not commonly known. If a different John Updike were being referred to, the clue would need to specify this in order to be consistent with Grice's maxim of quantity. In addition, a clue about someone who is not well known would be infelicitous, just as it is infelicitous to ask a question to someone who obviously cannot know the correct answer.

Questions in traditional speech acts may have even more felicity conditions; for example, it may be infelicitous to ask a question to which the asker already knows the answer. In certain situations, such as examinations, questions may "break" that condition (Searle 1969). A teacher knows the answer to a question before asking it, and therefore uses the question not to learn the answer, but to assess whether a student knows the answer. Because of their educational nature, these questions are not typically considered infelicitous. Crossword clues are another situation where the creator may "ask" the clue while already knowing the answer without being infelicitous, because the purpose of the clue is not to gain information but to entertain the solver.

2.5 Prior research

The interlocking nature of crossword puzzles presents an interesting problem for computational linguistics, and the bulk of research involving crosswords has focused on algorithms to leverage string interlocking for automatic generation and solving of puzzles. Most of these papers are fairly firmly rooted in the field of computer science, with only a surface analysis of the linguistic phenomena present in crossword puzzles.

In particular, the studies of automatic crossword generation have included only a rudimentary description of the clue-generating process. One group made progress in expanding linguistic research into crossword generation when they integrated WordNet, a lexical database, into their generation system for creating clues (Aherne and Vogel 2003). They also made some interesting advances in automatic theme generation. However, their system extended only as far as matching the syntactic role of the clue and the answer and did not address how clues may be calibrated to different levels of difficulty.

2.5.1 Automated solving

Most automatic crossword solvers use a combination of word interlocking and web scraping to come up with solutions. In creating the original crossword-solving software PROVERB, Greg Keim et al. determined several common syntactic categories for clues and used these to determine how to query their data to generate candidate answers (Keim et al. 1999). With regard to NYT puzzles specifically, they discovered that fill-in-the-blank clues were more prevalent early in the week and clues with question marks ("tricky" clues) were more prevalent later in the week.

A system proposed recently by Barlacchi et al. (2014) leverages the convention that clues be syntactically interchangeable with their answers to rank candidate answers. Their system builds a shallow syntactic tree for each clue representing the syntactic context in which they would expect to find an answer. They then search the web for text matching the clue phrase and parse it into a similar tree, identifying which words in the surrounding text are likely to fulfill the syntactic role of the answer. This pre-parsing allows the system to find matches with a close syntactic context.

2.5.2 Efron difficulty study

In 2008, Miles Efron attempted to create a probabilistic model of crossword difficulty based on a database of New York Times puzzles. The study used the day of publication as ground

truth difficulty level. A model was trained on a portion of the data set and then evaluated on a smaller test set. Overall, the system performed fairly well when the granularity of difficulty was reduced from six levels to three: easy, medium, and hard. When given a set of <clue, answer> pairs, it almost never misclassified easy sets as hard or vice versa.

The basic intuition used in the study is that <clue, answer> pairs are easier to solve if the clue almost always appears in conjunction with the answer. The primary metric the study used was thus a coincidence factor calculated by determining how many web documents containing the clue also contained the answer. While the intuition is generally valid, this method lacks a sophisticated understanding of the various factors that may contribute to a clue's difficulty. For example, a very obscure word is likely to have a high coincidence factor because of its small range of possible contexts. However, obscurity also makes a clue more difficult because a higher level of vocabulary is required. One of my central goals in this thesis is to explore all of the factors that contribute to difficulty and how they interact. I do not aim to create a metric like Efron's, but rather to examine the existing metric and explore its subtleties.

Chapter 3
Difficulty factors

Based on my initial analysis of the dataset, the tips provided by Amy Reynaldo in her book *How to Conquer the New York Times Crossword Puzzle*, and my own experience solving crossword puzzles, I have identified what I consider to be the major factors that may contribute to the difficulty of solving a clue. Most of these factors involve the clue itself, but the nature of answer also contributes to the difficulty.

In this section, I present a detailed description of each of these factors and their subcategorizations, along with examples. These factors often overlap and interact, and should not be considered independent of one another. In fact, as I explain in section 3.7, some of these factors are inversely related.

3.1 Cultural vs. lexical knowledge

Broadly, <clue, answer> pairs can be divided into two categories: lexical and cultural. Lexical includes English vocabulary, slang, and common abbreviations and phrases. Cultural includes names of specific individuals, places, acronyms, foreign words, and specific titles or song lyrics. Nickerson (2011) calls this second category "academic" or "literary" knowledge, however, I believe "cultural" more accurately describes the broad range of fields referenced in puzzles (sports, pop culture, geography, etc.) These categories are partly subjective and may overlap. For example, foreign words or brand names that have been incorporated into the general lexicon (e.g. *karaoke* or *xerox*) may arguably fall under either category.

The categorization can almost always be determined from just the clue, but occasionally a clue that appears purely lexical will have a culturally specific answer. The following examples demonstrate the possible combinations of lexical and cultural categories.

1. *Lexical clue, lexical answer*: Aquatic mammal → OTTER

2. *Cultural clue, lexical answer*: Lady Macbeth wanted one out → SPOT

3. *Lexical clue, cultural answer*: Garden figure → EVE

4. *Cultural clue, cultural answer*: Verne captain → NEMO

Examples 2, 3, and 4 may be considered cultural entries, because they require some cultural knowledge. From my own perspective as a solver, I find cultural entries more difficult because they require detailed knowledge of many different fields. However, some solvers appreciate how unambiguous the trivia-like cultural clues tend to be compared to lexical clues. Will Shortz in particular has been lauded for attempting to bring more diverse pop-culture content into the NYT puzzles, thus making them accessible to a wider audience (Reynaldo 2007).

It is also easier to find the answer to a clue by looking it up if it contains a cultural reference. Many crossword solvers view looking up answers on the Internet or by other means as cheating, but the fact that such an option is available does make harder puzzles more tractable for intermediate-level solvers. For the purposes of this thesis, I assume that solvers are not looking up answers, and crossword creators likely assume this as well.

The distinction between cultural and lexical entries is related to the pragmatic idea of a "common ground," or a collection of mutual knowledge between speaker and listener (Simons 2006). Often the common ground of an utterance can be built up through the context of a larger conversation. With crossword clues, context is extremely limited and common ground must be established within the space of a few words.

With lexical clues, theoretically the only common ground presupposed is knowledge of the English language. Cultural clues extend the common ground to include specific named entities. This is why *lexical clue, cultural answer* entries are particularly hard: the common ground appears to be established as the basic English lexicon, but is in actuality larger.

In Chapter 4, I investigate whether the ratio of cultural to lexical clues changes as the week progresses. Identifying such a trend would provide insight into what creators and editors expect to be more difficult. Unlike the other difficulty factors discussed in this chapter, there is not necessarily a clear hypothesis as to what the results of this analysis should show.

3.2 Ambiguity

The second specification in Grice's maxim of manner is "Avoid ambiguity" (Grice 1975). Much of the ambiguity inherent in language can be resolved by context. For example, sentence B does not violate this submaxim as long as it is preceded by sentence A, because the Maxim of Relation allows us to assume that sentence A informs sentence B. Sentence B without the context provided by sentence A is quite a flagrant violation of the Maxim of Manner, however, as it has at least two possible meanings: we were watching when she ducked, or we saw the pet duck that belongs to her. Sentence A allows us to assume that in this case, the speaker means the former.

A. John threw a Frisbee at Mary.

B. We saw her duck.

By nature, crossword clues have very little context. Clues are for the most part independent of each other and have a highly limited length because of space limitations. All clues must fit into the small space allotted by the newspaper. Ambiguity is thus unavoidable, but it is also useful for crossword creators wishing to create more difficult clues. Creators may specifically invoke ambiguity to make the "conversation" with the solver more challenging or entertaining.

Crossword clues are not the only situation in which ambiguity is desirable. Much research has been done on how Grice's maxims are flouted and violated in jokes (Attardo 1993). Puns in particular make heavy use of deliberate ambiguity as a source of humor. Just as someone listening to a joke finds humor in an unexpected but still logical punchline, the crossword solver finds entertainment in an unexpected but still logical answer. This is particularly true for tricky clues, which rely on the surprising ambiguity where one sense is much more likely than another based on context.

There are two main classes of ambiguity from a linguistic perspective, syntactic ambiguity and lexical ambiguity. Crossword clues leverage both forms.

3.2.1 Lexical ambiguity

Lexical ambiguity occurs when a word has multiple distinct meanings that are possible in the context provided. The ambiguity may be due to homonymy, where meanings are unrelated, or due to polysemy, in which case the multiple meanings are related. Lexical

ambiguities can often be found in crossword clues by examining clues that have been used for multiple different answers.

A. Kind of paper → TERM

B. Kind of paper → RICE

C. Clipped → SHORN

D. Clipped → TERSE

Often, one meaning is more abstract. For example, *terse* is a metaphorical interpretation of *clipped*. One avenue for future research could be to explore whether abstract meanings are more likely to occur in more difficult entries.

Note that a single clue having multiple possible answers does not always indicate that it is ambiguous. In many cases, a clue has one possible interpretation that may be fulfilled by several synonymous answers.

3.2.2 Syntactic ambiguity

Syntactic ambiguity occurs when there are multiple possible syntactic parses of a sentence or phrase. Because crossword clues are only a few words long, the potential for syntactic ambiguity is quite high. Lacking a full context, the solver must infer syntactic role from morphology and their knowledge of the lexicon.

In general, crossword clues should be interchangeable with their answers. Thus the syntactic role of the clue can give a fairly major hint. In some cases, the syntactic context of the clue allows the solver to make predications about the morphology of the answer. For example, a solver might fill in the plural "s" of an answer that must be plural without even knowing what the answer is. Sometimes such premature decisions prove incorrect in the case of irregular forms, but in general, knowing the tense, person, or number of an answer can be quite helpful.

As puzzles get more difficult, the proportion of one-word clues increases. One likely reason for this trend is that it is easier to hide the syntactic role of a single word. Once there are multiple words in the clue, the solver can begin to build an idea of the syntactic context in which the answer is likely to be found.

Distinguishing between syntactic ambiguity and lexical ambiguity of clues is beyond the scope of this paper, but it is worthwhile to note the distinction.

3.3 Obscurity

The first stipulation of Grice's maxim of manner is "Avoid obscurity of expression" (Grice 1975). Essentially, don't expect more of your partner in conversation's vocabulary and world knowledge than you ought to assume. Crossword creators, on the other hand, can expect quite a lot from the few brave solvers who attempt the puzzles later in the week. A classic way for a creator to make clues more difficult is to elevate the level of vocabulary expected of the solver. Using a more obscure or archaic term in a clue requires the solver to be better read or educated, and in some cases, familiar with common Latin and Greek roots.

Word obscurity is not easily quantifiable, because the "difficulty" of a word is dependent on subjective experience. However, some methods exist. Generally, calculating a word's overall frequency in a large corpus is a good measure of how well known it is.

Note that a frequency-based approach does not consider that certain senses, contexts, or connotations of clues may be more obscure and that this may also increase a clue's difficulty (see more about the interdependency of ambiguity and obscurity in section 3.6 below).

Even more difficult than measuring the obscurity of a lexical item is measuring the obscurity of a cultural entity (e.g. a geographical location, famous person, or book title). There have been studies that attempt to measure an individual's fame based on web searches, but their results are uncertain. As such, this thesis will not explore how cultural obscurity impacts difficulty.

3.4 Underspecificity

Underspecificity can be distinguished from ambiguity in that the words in an underspecific clue have an unambiguous syntactic role and semantic sense, but could still be referring to several possible answers.

If the creator wishes to make a clue easier, one way is to add more details, thus expanding the possibility that the solver will be able to identify the answer. Stripping away all extraneous detail until the clue is minimally descriptive requires the solver to extrapolate more. A minimal clue may or may not have multiple possible answers. What is important is that the solver believe there may be multiple answers, and that supplementary information that might jog the solver's memory be withheld.

Underspecificity is directly related to Grice's maxim of quantity, which essentially states that the speaker should share exactly as much information as needs to be shared, no more and no less (Grice 1975). Grice himself questions whether the requirement not to share

too much is actually part of the Cooperative Principle or just good social skills; however, a crossword clue is one context where providing too much information does breach the agreed-upon principles. If too much information is given, the puzzle is not fun anymore. Here, both sides of Grice's Maxim are certainly in play.

For example, the following are several clues for the answer KYRA, in ascending order of average day of publication. All of these arguably fall within the reasonable bounds of too much and too little information, but within that range, there is quite a bit of flexibility.

A. Actress Sedgwick of "The Closer" → KYRA (1.5)

B. Actress Sedgwick → KYRA (3.11)

C. "The Closer" star Sedgwick → KYRA (4.0)

D. Sedgwick of the screen → KYRA (6.0)

Note that clues A and B specify gender[1], A and C specify a particular context, and D specifies neither. This factor does not consider whether the information about gender or context is deemed more valuable, though that could be something interesting to examine in future research.

In order to identify clues following this pattern, I identified multiple clues for the same answer where the words in one clue form a proper subset of the words in other clue. For full methodology and results, see Chapter 4.

3.5 Difficult answers

There are fundamental differences in the shape and answer space of easy puzzles and difficult puzzles. Changing the clues can calibrate puzzle difficulty within a couple of days, perhaps, but the answers in a Saturday puzzle are still going to be longer and more obscure than those in a Monday puzzle.

Longer words are typically harder to guess, because there are more valid letter permutations. The Official Scrabble Dictionary contains 1311 3-letter words and 33315 7-letter words, which is not even counting all of the proper nouns and phrases that appear as valid

[1] In most contexts, it is generally accepted and advisable practice to use *actor* to refer to someone of any gender in the acting profession. However, crossword creators, for the exact purpose of making more informative clues, tend to use *actor* exclusively to specify male actors and *actress* to specify female actors. Alternately, a phrase like *of Hollywood* or a more neutral descriptor like *movie star* could be used.

crossword answers (Scrabble 1995). Furthermore, if the solver uses an intersecting answer to fill in one letter, they gain more information (proportionally) for a three-letter answer than a seven-letter answer.

Since all NYT crosswords (except Sunday puzzles, which are not considered in this thesis) are a 15x15 grid, it follows that the longer the answers are, the fewer clues there must be. As such, number of clues is highly correlated with average answer length. Efron's study that attempted to create a metric for calculating crossword difficulty found that number of clues on its own was a fairly good predictor of a crosswords day of publication (Efron 2008). Number of clues is also quite simple and efficient to calculate.

While it is important to keep in mind that the kinds of answers differ greatly throughout the week, this thesis is primarily concerned with how clues for the same answers vary across days of publication. Generally the experiments are controlled to consider only identical answers or identical clues at once.

3.6 Trickiness

By convention, clues that are particularly misleading or "tricky" are punctuated with a question mark (?). These clues tend to follow a few common patterns. The prevalence of tricky clues in a puzzle is one of the factors most strongly correlated with day of publication (Efron 2008). Many different types of clues can fall under this category. Often, theme clues are marked as tricky, especially when the theme relies on a pun or non-standard orthography. Overall, however, there are a couple of very common trends among tricky clues. I examined the 200 most common tricky clues, excluding those with multiple possible answers in order to partly control for ambiguity. Of these clues, almost two thirds consisted of a common two-word phrase or compound word. Among those, a third referred to a linguistic aspect of one of the words in some way.

3.6.1 Irrelevant surface meanings

As mentioned in section 3.2, tricky clues flout the ambiguity clause of the maxim of manner in a similar way to jokes and puns. The most frequently seen pattern for a tricky clue is a compound word or common phrase which has an established meaning somewhat abstracted from the original two words. Consider the following <clue, answer> pairs:

A. Wet bar? → SOAP

B. Horseplay? → POLO

C. Knight time? → YORE

D. One taking a bow? → EROS

E. Butler's quarters? → TARA

Each of these clues has a well-defined surface meaning. For example, *wet bar* typically refers to a counter for mixing beverages that includes a running sink. The sense of *bar* is highly unlikely to be anything but a counter for mixing beverages within the context of the two-word phrase *wet bar*. *Bar* by itself has multiple senses, but the collocation feature of being preceded by *wet* is a very strong indicator as to which sense is intended. By marking the clue as tricky, the creator indicates that the solver should abandon any preconceptions about how the two words interact with each other.

Interestingly, answers for tricky clues tend to depend on an interpretation that is actually more literal than the initial reading. For example, our typical definition of *horseplay* does not involve horses, but rather general rowdy or boisterous activity. The interpretation prompted by clue B above is more literal, because horses do in fact play polo.

There are variations within this type of clue that can alter the difficulty. For example, example C above is easier to understand because the spelling is different from the surface meaning the clue is attempting to invoke, *nighttime*, so it is easy even without the help of the question mark for the solver to see that the answer will not actually have to do with *nighttime*. On the other hand, example D is made more difficult by the fact that *bow* in the surface meaning of *taking a bow* is pronounced /baʊ/, but the *bow* associated with Eros is pronounced /boʊ/. In this case, the surface meaning is particularly distracting. Another method to increase difficulty is demonstrated in example E. Reynaldo (2007) calls this technique "capitalization obfuscation," as the creator takes advantage of the convention of capitalizing the first word of every clue to hide a proper noun. Here, the solver will likely read the common noun *butler* because of the surface meaning of the phrase *butler's quarters* as a place where a servant or butler lives. In fact, the clue is using the proper noun *Butler*, as it references the character Rhett Butler from the film *Gone with the Wind*.

3.6.2 Linguistic clues

Another common type of clue to be marked as tricky is the linguistic clue, or any clue for which the answer references the words of the clue itself rather than the entity or entities

signified by it. English contains a great deal of 3-letter affixes that crop up often in crossword puzzles because they contain common letters. For example, ESE and EER are among the most common crossword answers. Some linguistic clues are overt and not marked as tricky:

A. Prefix with conservative → NEO

Because affixes are so common as crossword answers, clues like the clue in example A become easily tired. One way to make such a clue more difficult and interesting, then, is to introduce a level of trickiness. These clues have the same pattern as other tricky clues in that they form common two-word phrases. In addition, they do not indicate explicitly that they are metalinguistic. The solver must rely on words like *ending*, *stub*, *back*, or *continuation* to indicate that the desired answer is a suffix.

B. Velvet finish? → EEN

C. Minnesota twins? → ENS

In example B, the presence of the modifier *velvet* initially suggests that *finish* will have the connotation of a surface texture or coating. However, *velvet* is not referring to the material but the word itself, and *finish* thus serves as a synonym for *suffix*. The clue as a whole suggests the word *velveteen*. Other variations that do not involve affixes are also possible, such as example C, which makes reference to the double *n* in the word *Minnesota* while on the surface referring to the state's sports team.

Similarly to how people refer to words metalinguistically in conversation by using a different tone or emphasis and in writing by using quotation marks, a crossword creator must explicitly mark when an answer relates to the words in a clue rather than their meaning. Failing to sufficiently introduce the linguistic context goes against convention. However, stating explicitly that the solver should be looking for a prefix, for example, may make the clue too easy. Marking a clue as tricky allows for the possibility of such a context without overtly stating it, allowing for increased difficulty without breaking the basic principles of cluing.

3.7 Factor interdependencies

In considering what makes clues difficult, one may come across two hypotheses that, while not mutually exclusive, seem somewhat logically opposed: 1) Hard clues will use obscure

words, therefore requiring a higher level of vocabulary to solve, and 2) Hard clues will use words with many possible meanings, which tend to be more common (for example, *cut*). The opposition of these two ideas is critical in understanding how these difficulty factors interact.

Firstly, it is fairly common knowledge that more common words are more likely to have irregular forms. Words where, for instance, the past and present tenses are sometimes identical are not likely to be rare. These words are often found in the most difficult clues, though, because of their ambiguity.

In fact, as a general rule, obscurity is actually correlated with high specificity and low ambiguity. Words that are highly specialized are not used as frequently, and so become obscure, yet are less likely to be vague or ambiguous. The fact that a word is unambiguous may actually cause it to be less well-known.

Finally, note that ambiguity and underspecificity, though distinct, are not inversely related in the same way. A decrease in specificity often occurs by removing semantic content or syntactic context, which may lead to ambiguity. It is important to keep these relationships in mind when examining how puzzles on different days of the week make use of different kinds of clues.

Chapter 4
Results

In order to assess if the difficulty factors proposed in Chapter 3 are reflected in the data, I ran several experiments. I wrote these experiments in Python with the aid of the Natural Language Toolkit (NLTK) library (Bird et al. 2009). In each experiment, I endeavored to control for all of the possible difficulty variables so that the results reflect the impact of the difficulty factor in question. Often this required filtering out any tricky clues, using clues with only one word, or using clues that have only appeared with one possible answer.

4.1 Cultural vs. lexical entries

I did not expect a particularly strong correlation between percentage of cultural entries (as compared to lexical entries) and day of publication, because neither type is inherently more difficult.

A general approach to classifying <clue, answer> pairs is to strip them of punctuation and capitalization and then search the Official Scrabble Dictionary for each word. If any word is not found, the entry is labeled as cultural. If all words in the clue and answer can be located in the dictionary, it is labeled as lexical. The Scrabble dictionary is particularly useful because it excludes proper nouns, which are a strong indicator of a cultural entry (Scrabble 1995).

This process is complicated considerably by the unmarked word boundaries within answers. To truly identify lexical clues as such, all possible word divisions of an answer would have to be attempted. In addition, many cultural entities have names that are also part of the common lexicon, such as *Bill Gates*.

Identifying capitalized words could also flag proper nouns, which are a strong indicator that an entry is cultural rather than lexical. Unfortunately, capitalization information is not

encoded into answers, and all clues begin with a capitalized letter.

The raw results for this experiment are displayed in Table 4.1. Generally there appear to be slightly more cultural entries than lexical entries, and the proportion increases as the week progresses. The difference between proportions on Tuesday and Wednesday is not statistically significant, nor are the differences among proportions on Thursday, Friday, and Saturday. There is a statistically significant difference among these three groups of day (Monday; Tuesday and Wednesday; and Thursday, Friday, and Saturday) with $p < 0.01$. The R-score for all six data points is 0.8575, which indicates significant positive correlation at $p < 0.05$.

Day of publication	Percent of entries labeled as "cultural"
Monday (1)	50.33%
Tuesday (2)	53.17%
Wednesday (3)	53.04%
Thursday (4)	54.94%
Friday (5)	55.09%
Saturday (6)	54.55%

Table 4.1: Percentage of cultural entries through the week

On examining the labeled data, I discovered many more false positives (lexical entries that had been labeled as cultural) than I found false negatives (cultural entries that had been labeled as lexical). It is likely that the total proportion of cultural entries is less than indicated in the results in Table 4.1. It is also possible that mislabeling is correlated with day of publication and therefore the observed trend is a result of the faulty system rather than a real correlation. For example, puzzles later in the week with their longer answers may be more likely to have multi-word lexical answers that would be falsely categorized as cultural.

4.2 Ambiguity

4.2.1 Counting number of possible answers

One simple approach to determining how ambiguous a clue may be is counting the number of answers it has been used for. If a word can serve as the clue for many different answers, it would make sense that the word is ambiguous. Even if ambiguity were not present, having multiple possible answers makes a clue objectively more difficult even for a solver with a complete knowledge of crossword history. Note that because crossword answers are not

a representative corpus of English words, we should not expect the number of recorded answers to correlate perfectly with number of hypothetically valid answers.

The results in Table 4.2 break down into three statistically significant categories: Monday and Tuesday; Wednesday, Thursday, and Saturday; and Friday. Within these categories, differences between the observed values are not statistically significant. Perhaps the most surprising result is the significantly fewer number of answers for Friday clues compared to its surrounding days. It is not clear why Friday puzzles in particular would show this discrepancy.

Day of publication	Avg. number of answers for clue
Monday (1)	1.82
Tuesday (2)	1.83
Wednesday (3)	2.05
Thursday (4)	2.03
Friday (5)	1.98
Saturday (6)	2.03

Table 4.2: Number of possible answers for all clues

Overall, the R-value is 0.7414, which is not statistically significant with $p > 0.05$. As it turns out, the number of possible answers is not correlated with a clue's day of publication.

This result is unexpected. Regardless of the linguistic reasoning behind a clue's ability to suit multiple answers, it seems obvious that deciding among more options would be more difficult. One potential explanation, and limitation to this experiment, is that all of these multiple answers for easier clues could have differing word lengths, thus would not be confused during an actual experience of solving the clue.

4.2.2 Counting the senses for one-word clues

It is also possible to identify ambiguous clues more directly. The WordNet database from Princeton catalogues distinct senses for a large English lexicon (Princeton University 2010). Each word is associated with a list of possible senses, including example sentences. Counting these senses provides a metric for measuring the lexical ambiguity of a particular word.

Clues with more than one word would be significantly more difficult to analyze, because the surrounding context may narrow down the possible senses of a word considerably, so the experiment only considers one-word clues. It also excludes words with non-alphabetic characters such as hyphens and apostrophes in order to be able to search WordNet. Clues with question marks were excluded as well, to control for the trickiness factor.

This was one of the most successful experiments. The results in Table 4.3 clearly demonstrate that more difficult clues tend to have more senses. A standard T-test for comparing two independent means with unequal variances was performed for each pair of consecutive days. Each resulted in a p-value below 0.05. The differences among days are all therefore statistically significant. The R-score is 0.9897, which is a strong positive correlation with $p < 0.001$.

Day of publication	Avg. number of senses of clue
Monday (1)	4.54
Tuesday (2)	4.80
Wednesday (3)	5.33
Thursday (4)	5.86
Friday (5)	6.67
Saturday (6)	6.92

Table 4.3: Number of senses for one-word clues

Some of the most unambiguous clues have in fact been used in reference to many different answers. For example, according to WordNet, *oodles* has only one sense: "a large number or amount." However, it has been the clue for 16 distinct puzzle answers, including ATON, LOTS, and MANY (Princeton University 2010). Clearly, having few potential senses does not always correlate with having few potential synonyms. From the perspective of a solver, it might be easier in practice to solve polysemous clues with limited potential synonyms for each sense than unambiguous clues with a surfeit of potential synonyms. From the perspective of the creator and editor, however, polysemous clues apparently seem more difficult.

Note that this experiment was performed on a different, smaller set of clues than the experiment that counted possible answers. To ensure that the better results are due to counting senses and not to the data set used, I repeated the multiple-answer experiment on just the set of one-word clues that appeared in WordNet. All results are statistically significant with respect to all other values, with $p < 0.05$. The R-value for these six data points is 0.7944, which is not a strong enough correlation to be statistically significant.

As seen in Table 4.4, the overall values increased compared to the search on all clues, supporting the idea that one-word clues are more ambiguous in general. However, there is still not a consistent pattern as to how number of answers correlates to day of publication. In general there appears to be a slight positive correlation, but there is also an unexplained spike on Wednesday.

Day of publication	Avg. number of answers for clue
Monday (1)	4.29
Tuesday (2)	4.55
Wednesday (3)	5.24
Thursday (4)	4.82
Friday (5)	5.07
Saturday (6)	5.19

Table 4.4: Number of possible answers for one-word clues

4.3 Obscurity

One major challenge for measuring obscurity in relation to clue difficulty was accounting for different levels of difficulty among answers. Harder crosswords do have longer and often more obscure answers, which may be more likely to need obscure clues. To analyze the effect of word obscurity on the clues independently, this potential for skewing the data must be accounted for.

In order to avoid biasing the experiment towards more obscure answers, rather than clues, I identified a set of answers that had a variety of different clues with a range of average days of publication. Collecting answers with a different clue for each day of publication was too limiting, so clues were condensed into three broader difficulty categories based on average day of publication. These categories cover approximately equal spans of publication days and do not overlap. An answer's clues were only considered if the answer had at least one clue in each of these three categories. If multiple clues for an answer fell into one category, one was chosen at random.

All clues also had to pass a basic vocabulary test which ensured that each word was represented in the Official Scrabble Dictionary. As discussed earlier, this provides a high likelihood that most of the clues in question were lexical, not cultural (recall that the cultural classifier had far more false positives than false negatives). Finally, all of the answers considered had to have a medium average difficulty to avoid skewing the data. The final experiment analyzed 12543 distinct answers, and measured obscurity of one early, midweek, and later clue for each answer.

The second major challenge was choosing an appropriate corpus to base the judgements about obscurity on. The best choice would probably be the Corpus of Contemporary American English (COCA), but rights to the data are highly limited and expensive. Instead,

I considered three fairly different corpora: the Microsoft N-grams corpus from Bing(Wang et al. 2010), the Project Gutenberg corpus from NLTK (Bird et al. 2009), and the Open American National Corpus or OANC (Ide and Suderman 2012). These three corpora represent a range of different styles of English. The Bing corpus primarily includes modern American English, including slang and certain technological terms, while most of the Project Gutenberg texts are over 50 years old and tend to include antiquated terms. The American National Corpus represents a middle ground, with text drawn from a variety of written texts and spoken transcripts. I considered a clue "obscure" if any of the words in it did not appear in the most frequent 20,000 words from the corpus; for the PG corpus, due to limitations of the data available, the cutoff point was after the most frequent 1,500 words.

Day of publication (approximate)	**PG**	**OANC**	**Bing**
Early (avg < 2.5)	46.59%	46.34%	43.03%
Midweek ($2.75 <$ avg < 4.25)	41.24%	40.21%	37.93%
Late (avg > 4.5)	44.27%	44.25	40.93%

Table 4.5: Percent of obscure clues based on various corpora

The results in Table 4.5 are somewhat surprising. With each of the corpora, clues early in the week show the highest level of obscurity. In addition, the midweek clues have a lower level of obscurity than either the early or later clues, so there is not a clear correlation in either direction. According to a standard Z-test for comparing two population proportions, the distinctions among proportions for the three categories are all statistically significant with $p < 0.001$. This significance holds for all three corpora.

Clearly the data do not support the hypothesis that crossword creators use more obscure words to make clues more difficult. One likely explanation for this is the tendency discovered in section 4.2 for clues later in the week to be more ambiguous rather than more obscure.

Figure 4.1 shows the 50,000 most frequent words from the Microsoft N-grams Bing Corpus (Wang et al. 2010) along with the number of senses listed for each word in WordNet (Princeton University 2010). Notice that more frequent words tend to have more senses, while as the frequency decreases, the number of senses stabilizes around 10.[1] The negative correlation is only moderate, but the because of the very large sample size it is statistically significant with $p < 0.05$. This correlation makes intuitive sense from a couple of perspectives. Firstly, if a word has multiple senses, there are more contexts in which it might

[1] Note that WordNet's distinctions between senses are quite granular, often leading to a higher-than-expected number of senses overall. The word with the most senses is *break* at 75, closely followed by *cut*. Morphological variants of these two words alone account for most of the data points above 50 senses.

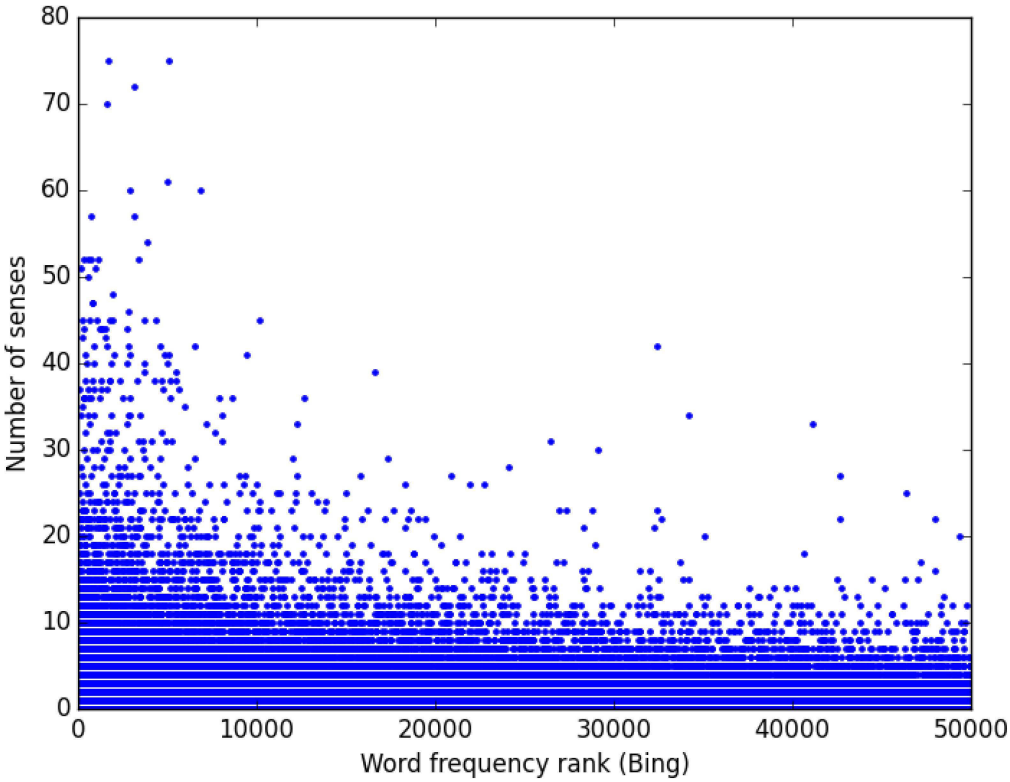

Figure 4.1: Number of senses of words by frequency

appear and its frequency would therefore be higher. Additionally, words that are used very commonly may evolve more nuanced senses over time.

This trend may be responsible for the lower obscurity of mid- to late-week clues. If more ambiguous clues are being used, their obscurity will be lower. The two difficulty factors may then have an inverse relation.

The somewhat more difficult to explain phenomenon is that midweek clues are less likely to be obscure than clues earlier or later in the week. Looking back on the ambiguity experiments, the midweek clues tended to have higher numbers of possible answers than clues from the end of the week. Midweek clues also have a higher proportion of cultural clues. It's possible that because creators rely heavily on ambiguity and cultural knowledge as a source of difficulty in midweek puzzles, obscurity is rarer. For puzzles late in the week, a combination of obscurity and ambiguity is used. This trend may be an intentional pattern or just a result of tradition and intuition as to what constitutes a difficult puzzle.

4.4 Underspecificity

Recall two of the examples from Chapter 3.

A. Actress Sedgwick of "The Closer" → KYRA (1.5)

B. Actress Sedgwick → KYRA (3.11)

Based on the intuition that clue A is more informative and easier than clue B, I performed a search on the data to find clues whose words were a proper subset of other clues for the same answer. These clues would hypothetically be more difficult, because they are less informative or specific.

The search is case-insensitive and ignores words like *the*, *a*, and *of*. For example, clue A above forms the set {*actress*, *sedgwick*, *closer*} and clue B forms {*actress*, *sedgwick*}. Set B is a proper subset of set A, so the clues are added to the list of clues of interest. For each pair of clues that form a subset and superset, I compare the average days of publication. This procedure ignores subtleties like the difference between *star* and *actress*, but captures the basic difference in volume of information between clues. In over 70% of these pairs, the clue with more information (superset clue) was easier than its less informative counterpart. This is highly statistically significant, with $p < 0.000001$ using a binomial distribution.

Because underspecificity as defined here involves a dependency relationship between different clues for a single answer, it is not feasible to express more detailed results for underspecificity based on day of publication. Future work could create a more independent definition of underspecificity in order to examine this phenomenon more closely.

4.4.1 Clue length

It would follow that clue specificity is somewhat related to clue length. Longer clues have the potential to encode more information, possibly making them easier. However, clues later in the week tend to be slightly longer, which does not fit with our intuition about how more specific clues should be easier. The results in Table 4.6 show average clue length, in both characters and words. Clues with blanks and tricky clues were removed from consideration, as both factors have been shown to correlate with difficulty and may also influence length. Both measurements show a similar clue length for the first four days, and then a slight jump from Thursday to Friday. This jump is quite statistically significant for both words and characters, with $p < 10^{-100}$. Otherwise, distinctions between average clue length do not appear to be highly statistically significant.

Day of publication	Average clue length (characters)	Average clue length (words)
Monday (1)	16.28	2.70
Tuesday (2)	16.39	2.71
Wednesday (3)	16.13	2.68
Thursday (4)	16.29	2.70
Friday (5)	17.32	2.87
Saturday (6)	17.54	2.89

Table 4.6: Clue length through the week

Even removing tricky clues from consideration, clue length increases as the week goes on. It's possible that more words are necessary to convey the greater amount of information contained in the longer answers of more difficult puzzles. However, it's also likely that the content of the clue is more important to difficulty than the length of the clue. Consider the following clues, which are quite lengthy and informative, but both published late in the week.

A. Fictional school whose motto is "Draco dormiens nunquam titillandus" → HOGWARTS (6.0)

B. What a virtuous woman is worth more than, according to Proverbs 31:10 → RUBIES (6.0)

Both of these clues use a highly specific and informative but relatively obscure detail about the answer. Many of the more difficult, longer clues involve a quote by a famous individual or a passage from literature. These kinds of entries are actually fairly educational for the solver, who has likely heard of the answer in a separate context. Note that these clues may be easier to look up, but that is not related to their innate difficulty.

4.5 Trickiness

Previous studies have noted that number of tricky clues has a high correlation with difficulty of a puzzle (Efron 2008). The results in Table 4.7 confirm this, showing a low proportion of tricky clues early in the week and a higher proportion as the week goes on, leveling off around Thursday. These statistics were gathered simply by counting the percentage of clues ending in a question mark, which is the indicator of a tricky clue.

A standard Z-test for comparing two population proportions was performed for each pair of consecutive days. The increases in tricky clue proportions from Monday to Tuesday, Tuesday to Wednesday, and Wednesday to Thursday are statistically significant, with $p < 0.05$. The differences between proportions among Thursday, Friday, and Saturday are not statistically significant. The R-score for all six data points was 0.9134, indicating a strong positive correlation which is statistically significant at $p < 0.05$.

Day of publication	Percentage of tricky clues
Monday (1)	0.89%
Tuesday (2)	1.99%
Wednesday (3)	2.97%
Thursday (4)	3.81%
Friday (5)	3.80%
Saturday (6)	3.74%

Table 4.7: Clue trickiness through the week

I attempted to discover more detailed trends among the kinds of tricky clues used on each day, but after an analysis of the 200 most common tricky clues, no consistent patterns emerged. Linguistic clues based on affixes made up about 20% of tricky clues on each day. Clues early in the week were not significantly more likely to include indications of the true meaning in their spelling, and later clues were not significantly more likely to use an unclear pronunciation or capitalization obfuscation to hide the true meaning.

4.6 Analysis of results

Overall, my analysis points to a nuanced system of creating appropriately difficult clues. Most of the factors do not increase or decrease linearly or even steadily over the course of the week. Most factors are not even monotonic.

Some results did not show a statistically significant difference between days of publication, but a number of experiments showed significant differences. We can be certain that the NYT's claim of escalation in difficulty throughout the week is not entirely unfounded, nor is it a result of any single particular strategy.

Figure 4.2 collects normalized results from the difficulty experiments from Chapter 4. For the obscurity metric, I used the results from the OANC corpus, because they were the most statistically significant. Results based on percentages were all divided by 10 so that the factors could all be shown on the same scale, for purposes of comparison. More important

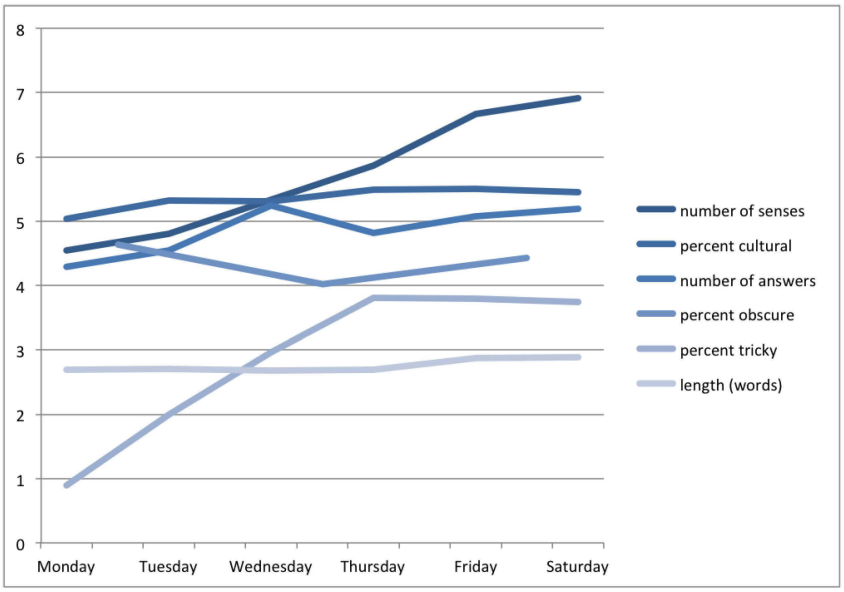

Figure 4.2: Normalized difficulty factors

than the magnitude of each point is the shape of each line, demonstrating how the difficulty factors all behave in different ways. Occasionally, factors share a peak or plateau, such as trickiness and obscurity. However, obscurity and number of answers appear to almost mirror each other. This is consistent with the inverse relationship between frequency rank and ambiguity as discovered earlier.

The target audience for each day of the week may be distinguished by more than just their crossword solving prowess. Midweek solvers may be considered more casual solvers who seek entertainment, thus creators provide them with puzzles full of pop culture and clever riddle-like ambiguities. By the end of the week, solvers are likely more serious and competitive, perhaps demanding a broader range of difficulty sources such as obscure vocabulary.

It is unlikely that crossword creators and editors keep all of these factors in mind when calibrating the difficulty of puzzles. Rather, they use experience, patterns, and intuition to determine what qualifies as “difficult.”

Chapter 5
Conclusion

5.1 Pragmatics of crossword puzzles

The concept of exploring crossword puzzle difficulty through the lens of pragmatics is not entirely intuitive. Puzzles are not speech acts and typically are not even considered a form of communication. However, over the past century, a cooperative principle of crosswords has developed that allows solvers to draw implicatures from even the briefest and vaguest clues. Often it still takes a few guesses to find the answer that works with the words around it, which is why so many people choose to solve the crossword in pencil. The creator's meaning is not fully revealed to the solver until the answer is confirmed by its intersecting letters. A completely and correctly filled-in crossword puzzle is a successful communication act, in which one individual has guided another in constructing a series of true statements.

How can a crossword clue provide enough information in a few words to prompt a stranger to pick exactly the word the creator intended? Crossword creators and editors accomplish this by balancing the rules that make understanding easy, in the form of Gricean maxims and conventions, with the rules that make understanding entertainingly challenging, in the form of flouting these same maxims and conventions.

The puzzles earlier in the week appear to serve as a kind of training for the more difficult puzzles later on, introducing these rules and conventions in a manageable setting. These results demonstrate that not only are puzzles published later in the week more difficult, they are also more fun. They have more unique and detailed clues; they rely on ambiguity and trickiness for difficulty rather than obscurity of vocabulary. Instead of using clues with an abundance of synonyms, which could be ostensibly more difficult to sift through, creators use clues with an abundance of senses, creating a dynamic experience for the solver.

5.2 Future work

The experiments performed in this thesis suffer from limitations that are for the most part described in earlier chapters. Further research could examine puzzles as a whole, looking at the balance of clue difficulties within one puzzle since that is how solvers experience it. This thesis does not consider puzzle themes, which certainly involve linguistic phenomena of their own.

Additionally, it would be interesting to perform similar experiments on another newspaper's puzzles to compare how the editor's individual style impacts how difficulty factors shift over the course of the week. As discussed earlier, much of the content of this thesis relies on data drawn from Will Shortz's personal beliefs about how difficult solvers will find each clue. Studying how crosswords function in other languages would also be an interesting avenue for future research.

A well-constructed, well-clued crossword puzzle can capture the exquisite intricacy of natural language. Despite the trend towards machine-generated and machine-solvable crosswords, puzzles have and always will be a fundamentally human affair, with which humans may explore the limits of functional communication.

References

Aherne, Aoife and Carl Vogel. 2003. WordNet enhanced automatic crossword generation. *Proceedings of the 3rd International Wordnet Conference* .

Attardo, Salvatore. 1993. Violation of conversational maxims and cooperation : The case of jokes. *Journal of Pragmatics* 19:537–558.

Barlacchi, Gianni, Massimo Nicosia, and Alessandro Moschitti. 2014. Learning to rank answer candidates for automatic resolution of crossword puzzles. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning* 39–48.

Bird, Steven, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Efron, Miles. 2008. Shannon meets Shortz: A probabilistic model of crossword puzzle difficulty. *Journal of the American Society for Information Science and Technology* 59:875–886.

Gerofsky, Susan. 1996. A linguistic and narrative view of word problems in mathematics education. *For the Learning of Mathematics* 36–45.

Ginsberg, Matt. 2012. Cluer Database. URL `http://www.otsys.com/clue/`.

Grice, H. Paul. 1975. Logic and conversation. 41–58.

Ide, Nancy and Keith Suderman. 2012. Open American National Corpus. URL `http://www.anc.org/`.

Keim, Greg A., Noam Shazeer, Michael L. Littman, Sushant Agarwal, Catherine M. Cheves, Joseph Fitzgerald, Jason Grosland, Fan Jiang, Shannon Pollard, and Karl Weinmeister. 1999. PROVERB: The Probabilistic Cruciverbalist. *American Association for Artificial Intelligence* .

Nickerson, Raymond. 2011. Five down, absquatulated: Crossword puzzle clues to how the mind works. *Psychonomic Bulletin and Review* 18:217–241.

Princeton University. 2010. About WordNet. URL `http://wordnet.princeton.edu`.

Reynaldo, Amy. 2007. *How to conquer the New York Times crossword puzzle: Tips, tricks and techniques to master America's favorite puzzle*. New York: St. Martin's Griffin.

Scrabble. 1995. *The Official Scrabble Players Dictionary*. Springfield, MA: Merriam-Webster.

Searle, John R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. 20. Cambridge University Press.

Simons, Mandy. 2006. Presupposition without common ground. *Ms., Carnegie Mellon University* .

Simpson, Paul. 2001. Reason and tickle as pragmatic constructs in the discourse of advertising. *Journal of pragmatics* 33:589–607.

Wang, Kuansan, Christopher Thrasher, Evelyne Viegas, Xiaolong Li, and Paul Hsu. 2010. An overview of Microsoft web n-gram corpus and applications. URL `http://research.microsoft.com/apps/pubs/default.aspx?id=130762`.

Weiner, E Judith and Paul De Palma. 1993. Some pragmatic features of lexical ambiguity and simple riddles. *Language & communication* 13:183–193.