1. Introduction
This goal of the project is to build a tool that will help learners or non-native speakers of English. Quite frequently, a learner may have a general idea of a word’s meaning, but may not be sure of how to use the word in a sentence. Furthermore, since a word is defined by the surrounding context, it is not sufficient to know just the meaning of the word in order to be able to use it correctly.

In this paper, we describe a tool that illustrates the usage of a word by finding commonly used expressions with the word and providing example sentences with such expressions. We believe that the tool will be useful for both learning and teaching of English. Moreover, native speakers wishing to find common usages of a word may also benefit from the tool.

The rest of the paper is organized as follows: first, we describe resources that are similar to the tool that we develop. We then give an overview of the system and briefly describe each component. We provide example output from the program and offer a discussion of the results. We conclude with possible directions for future work.

2. Similar tools
While the function of our tool is similar to that of dictionaries, we distinguish the following key differences: first, we focus on the usage of words and not on defining their meanings. Second, we rank expressions based on frequency. Finally, because we extract example sentences from context, we believe our examples are more “natural” than those provided in a dictionary.

One can argue that identifying commonly used expressions can be accomplished with the use of a search engine. However, while a search engine searches for documents that contain a certain word, we are interested in determining which words appear frequently in the vicinity of the target word. In fact, as we describe below, we use a search engine to build a corpus of documents with the target word, and then apply our algorithm to identify common expressions.

3. Implementation
3.1 System Overview
Fig. 1 gives an overview of the system. The basic components include corpus generation, text pre-processing, Inverted Index generation, and ranking the expressions. Our interface is over the web to allow for easy query submission and result viewing. We now describe each component.
3.2 Corpus Generation
We refer to the sentence where a target word appears as the context. A corpus is a collection of contexts containing the target word.

Unlike a search engine, we are interested in all contexts where the target word appears, irrespective of the document frequency of the target word in a given document.

We explore two ways of collecting contexts for a target word: using a static collection of documents and querying the web. In the first case, we use a corpus of newspaper articles (NewspaperCorpus) of 3.5 Million words. The advantage of a static corpus is the large amount of data. The disadvantage is that it has a limited domain. In the second case, we submit the target word as a query to Yahoo! search engine to retrieve documents containing the word. Given a collection of documents, we extract all contexts containing the target word to build a corpus (YahooCorpus). Using a search engine allows one to find information about any word. By contrast, the vocabulary of NewspaperCorpus is rather limited. One problem with this approach is that it takes time to download the documents. We therefore retrieve only the first fifty web pages returned from Yahoo!. Using only fifty documents limits what word expressions actually appear in the results.

3.3 Pre-processing
The pre-processing includes two stages:
1) extracting text from html
2) cleaning the text
The first stage is omitted for a static corpus. Cleaning includes removing non-alphabetic symbols and lowercasing.

3.4 Inverted Index

We search a corpus to determine the most typical context words by extracting words within a certain window of the target word and ranking words based on their frequencies. We collect counts for all words and word pairs in a given position in the context window. We then build a dictionary keyed on words/word pairs that contains information about frequency in specific position and sentences where the word/word pair appears. Only a limited number of sentences are kept for each word (word pair)/position combination. The structure allows us to extract sentences for each expression efficiently, once the ranking is performed.

3.5 Ranking expressions

We refer to the target word combined with the context word (or words) as an expression. The rank of an expression is determined by three factors:

- **TF**: raw count
- **IDF**: \( \log \frac{\text{# sentences in corpus}}{\text{# sentences containing the word}} \)
- **Position Normalization**: Reward context words closer to the target

These factors reflect our belief in how a common expression can be identified. In particular, one would expect that words representing a commonly used expression occur frequently with each other. To reduce the weight of frequent words, such as function words, we use inverse document frequency. IDF is computed in advance using the newspaper collection described in 3.2. To compute inverse document frequency of a word, we treat each sentence in the corpus as a document. We multiply IDF scores of individual words to compute IDF for a word pair. Finally, position normalization rewards those context words and pairs that occur closer to the target word. The total score of an expression is the product of these factors.

3.6 Search Interface

Our python script allows the user to search our corpus via the web and to set the context window size. We expect this option will typically not be used, as the default size of two words before and after the target word should be sufficient.

The web interface is used to output the top ranked expressions and examples. The search brings back a list of results where the target term is highlighted within the context of the sentences it was matched in. The list is broken into two sections: one with the list of 2-word expressions and the other of 3-word expressions. We decided to do this because in some cases the user might not care about seeing 3-word expressions and might only be interested in the 2-word ones.

4. Performance Evaluation and Discussion

We evaluate the performance of the tool using NewspaperCorpus and YahooCorpus. The appendix presents output for the words come and notorious. We observe that the performance is generally better when NewspaperCorpus is used. This may be due to the fact that a lot of the documents retrieved with Yahoo! do not contain much text.

A search for notorious using NewspaperCorpus returns characteristic examples of the
word (most notorious criminals, notorious for, notorious debt). These seem like typical results from a newspaper article and illustrate the usage of this word.

A search over the whole Internet using Yahoo's API returns results which are all related in some way to the music band named Notorious. These results paint quite a different picture, as they do not actually give us a good sense as to how the word is used. Rather, the examples are mainly of a popular usage of the word in a proper noun which is not necessarily helpful to the target users of this program.

As for the word come, both corpora find useful expressions with this word, but the NewspaperCorpus is able to find more expressions. Most of the expressions are phrases containing prepositions. This is because prepositions occur frequently with verbs and are more common that other parts of speech.

One general problem with both corpora is that results contain a lot of noise, i.e., context words that do not constitute an expression with the target word. We believe this may be a side effect of using IDF, because it favors rare words. On the other hand, not using IDF may only find function words which do not give information about the target word. For example, when we do not use IDF with the verb come, we retrieve more phrases with prepositions. But when we use IDF we also find idiomatic expressions (come under fire, come full circle)

5. Conclusion and Future work

We have presented a tool that given a word will find typical usages of the word in natural language.

We believe there are several directions for future work. We would like to expand our search ability to multiple words or phrases. This feature will allow a user to verify the usage of a phrase instead of just a word and to find out whether a specific phrase is common. Additionally, we conjecture that querying a search engine with phrases will be more effective, as more text documents will be retrieved.

Finally, given that the results are noisy, we would like to further refine the ranking formula.

We would like to thank Professor Zhai for the helpful suggestions and for providing us with the newspaper corpus.

The tool can be accessed online at
http://csil-linux35.cs.uiuc.edu:4000/cgi-bin/web_word.py
http://csil-linux35.cs.uiuc.edu:4000/cgi-bin/web_word_yahoo.py
The first link uses the static newspaper corpus, the second – builds a corpus by querying Yahoo!

Team Contributions

Paul: Designed the web interface and wrote the CGI scripts to interact with the search code. Wrote the initial script to output and debug results in the terminal.

Alla: Wrote scripts to process html files and to clean up text, implemented TF/IDF, Inverted Index, and Yahoo API interaction.

Both: decided on what python data structures and methods to use. Wrote the proposal, presentation and final report.
Appendix:

- A search for notorious with NewspaperCorpus

http://csil-linux35.cs.uiuc.edu:4000/cgi-bin/web_word.py?q=notorious

Top 20 most common 2 word expressions

- notorious for (rank: 9.28, 5 matches)

  It was notorious for blustery weather year-round with winter temperatures dipping to 20 below zero.
  The arrests were made at a hotel and eight cottages in Pagsanjan, a riverside resort town 40 miles southeast of Mani.

Top 20 most common 3 word expressions

- colombia's notorious medellin (rank: 158.53, 1 matches)
- hurg _ notorious murderer (rank: 146.02, 1 matches)
- notorious medellin cartel (rank: 137.64, 1 matches)
- most notorious criminals (rank: 131.58, 2 matches)
- notorious prison island (rank: 128.39, 2 matches)
- helping _ notorious medellin (rank: 118.15, 1 matches)
- notorious mass murderer (rank: 115.95, 1 matches)
- america's _ notorious criminals (rank: 109.75, 1 matches)
- notorious medellin drug (rank: 87.99, 1 matches)
- mc. _ notorious detroit (rank: 85.28, 1 matches)
- nation's _ notorious criminals (rank: 85.19, 1 matches)
- notorious criminals including (rank: 74.06, 1 matches)
A search for *notorious* with YahooCorpus

http://csil-linux35.cs.uiuc.edu:4000/cgi-bin/web_word_yahoo.py?q=notorious

Word: notorious

Left search position: -2  Right search position: 2

Minimum number of matches to show example sentences: 4

Search

Top 20 most common 2 word expressions

**notorious **b.i.g (rank: 385.74, 28 matches)

  in just a few short years the *notorious b.i.g* went from a brooklyn street hustl
  all-too-brief odyssey almost imm
  movies a-z the *notorious b.i.g*
  box set *notorious b.i.g*

**notorious **_dvd (rank: 110.21, 16 matches)

  _notorious - dvd new
  _notorious 2001 dvd new sealed oop criterion
  _notorious criterion dvd - brand new amp sealed oop

**notorious** 2001 (rank: 100.90, 9 matches)

  _notorious 2001 dvd 1946 hitchcock criterion brand new
  _notorious 2001 dvd
  _notorious 2001 dvd
A search for *come* with NewspaperCorpus

http://csil-linux35.cs.uiuc.edu:4000/cgi-bin/web_word.py?q=come

Top 20 most common 2 word expressions
to come (rank: 252.86, 242 matches)
  abrams also told the cable news network that if the government of president eric arturo delvalle says to american banks don't pay money to noriega then all those banks are going to come to the state department come out (rank: 244.39, 66 matches)
  but as polls began to come out showing bush with a solid lead in the state robertson toned down his optimism
come from (rank: 241.66, 97 matches)
  a doctor later diagnosed the bite as having come from a black widow spider according to matt parmenter leonard's attorney
come back (rank: 236.99, 51 matches)
  after i come back with a summation of my talks with representatives of the american government the cabinet will meet immediately and i think we'll have a decision on this subject shamir said on israel radio
come up (rank: 235.94, 70 matches)
  all stations in a block of states come up for renewal in the same year

come up with (rank: 726.60, 47 matches)
  a group of students venturing where government has been less than successful vows to come up with a series of tax forms that are understandable even to the high school dropout
views come closest (rank: 348.17, 3 matches)
come under fire (rank: 338.99, 8 matches)
  dallas police have come under fire by minority leaders and city hall in the past two years for being trigger-happy when confronting suspects especially blacks and hispanics
has come under (rank: 302.32, 13 matches)
  none of the speakers mentioned president kurt waldheim who has come under increasing pressure to resign because of allegations about his past while serving in the german army in the balkans during world war ii
crewcuts come pouring (rank: 248.25, 1 matches)
evangelists come full (rank: 238.15, 2 matches)
come under increasing (rank: 230.49, 4 matches)
  none of the speakers mentioned president kurt waldheim who has come under increasing pressure to resign because of allegations about his past while serving in the german army in the balkans during world war ii
going come out (rank: 205.43, 6 matches)
  they're all vying for market share leadership in one niche or another and not all of them are going to come out with gold medals he said
have come under (rank: 204.77, 9 matches)
  drug testing policies also have come under legal attacks led by the american civil liberties union
  exporters out of certain japanese markets such as highly controversial restrictions on imported beef citrus fruits and rice add to the trade imbalance problem and have come under heavy fire from the u.s
  gorbachev's policies of perestroika have come under a good deal of opposition however from the people in the economic administrative bureaucracies in moscow and in the soviet union's 15
republics he said
maybe _ come onto (rank: 199.03, 2 matches)
to come up (rank: 197.22, 28 matches)
and it needed to come up with about 30 million as its share of 120 million in overdue federal-
state medicaid payments
chocolate _ come paired (rank: 191.43, 1 matches)
come full circle (rank: 191.43, 2 matches)

• A search for come with YahooCorpus
http://csil-linux35.cs.uiuc.edu:4000/cgi-bin/web_word_yahoo.py?q=come

Top 20 most common 2 word expressions
come true (rank: 88.17, 13 matches)
dreams come true - aol music
dreams come true
get the full dreams come true biography
come down (rank: 50.71, 12 matches)
customs that come down from colonial times
come down on
come down to
come - (rank: 42.37, 5 matches)
come - u.s.nbspreleases
come - singlesnbspsampnbspsep's
to come off with honors
the end of the novel just doesn't come off
come off it
come blow (rank: 30.32, 4 matches)
imdb message board for come blow your horn 1963
come back (rank: 27.88, 6 matches)
you gotta go there to come back
the thrill of it all / lover come back / it happened to jane
to come back with a witty remark
come _ way (rank: 23.77, 10 matches)
come my way close your eyes come close to me and sleep i39ll be with you i39ll be with you
i39ll stay beside you so come my way
calling out can you hear me yea so come my way
come my way alone in this darkness come close to me now i39ll shine some light i39ll be with
you i39ll be with you i39ll stay beside you so come my way
come up (rank: 23.59, 7 matches)

Top 20 most common 3 word expressions
dreams come true (rank: 556.05, 5 matches)
more songs/videos from dreams come true
see full dreams come true discography
come my way (rank: 448.86, 10 matches)
come my way close your eyes come close to me and sleep i39ll be with you i39ll be with you
i39ll stay beside you so come my way
calling out can you hear me yea so come my way
come my way alone in this darkness come close to me now i39ll shine some light i39ll be with
you i39ll be with you i39ll stay beside you so come my way
come games.doc endobj (rank: 379.57, 1 matches)
come true discography (rank: 373.75, 2 matches)
come planet earth (rank: 362.21, 3 matches)
come blow your (rank: 341.71, 4 matches)