ABSTRACT
This paper describes a dictionary-like reference tool that is designed to help users find information that is similar to what one would find in a dictionary when looking up a word, except that this information is extracted automatically from large corpora. For a particular vocabulary item, a user can view frequency information, part-of-speech distribution, word-forms, definitions, example paragraphs and collocations. All of this information is extracted automatically from corpora and most of this information is extracted from Wikipedia. Since Wikipedia is a massive corpus covering a diverse range of general topics, this information is probably very representative of how target words are used in general. This project has applications for English language teachers and learners, as well as for language researchers.

Categories and Subject Descriptors
K.3.1 [Computer Uses in Education]: Computer-assisted instruction (CAI)

General Terms
Human Factors, Design, Theory

Keywords
Corpora, Wikipedia, Language teaching

1. INTRODUCTION
It is easy to use Natural Language Processing (NLP) methods to determine frequent vocabulary terms in various genres of English (e.g., see [1] for methods of extracting frequency lists of ngrams from corpora). One can use vocabulary frequency lists of this sort for language teaching. In order to become proficient in English, English language learners need to know words that occur frequently in English in general, and also possibly words that occur in particular target genres of interest to the learner. However, simply giving students lists of words, a few examples, and telling them to learn the words probably is not sufficient. One does not become proficient in a language simply by memorizing a list of vocabulary terms.

Language learners (e.g., beginning to intermediate English language learners) frequently come across vocabulary terms that they do not know. In this case, a dictionary can come in handy. However, especially for a language learner, the information in a dictionary, while useful, may not be sufficient to truly acquire a new vocabulary term. A dictionary definition may be difficult to understand. Examples contained in a dictionary may not provide sufficient context. It may be difficult to learn a vocabulary item simply from reading about it in a dictionary.

In this paper, we describe a Corpus Reference Tool which provides information of the sort that one can find in a dictionary, except that some of this information is more detailed, and this information is extracted from large corpora. An important feature of this application is that most of the information it provides is extracted from Wikipedia. Since Wikipedia is a massive corpus that covers a diverse range of topics, this information is likely representative of how target vocabulary terms are used in general, and thus, this information can be of use for helping language learners. This application provides a user with lists of vocabulary terms in target genres. For a target term, selected either from a vocabulary list or input directly by the user, a user can view part-of-speech (POS) information, morphological information, definitions, collocations, and example paragraphs. Example paragraphs, extracted from Wikipedia, are particularly useful since they provide more context than sentence length examples that are generally provided by dictionaries or concordancers.

This project has several primary goals: a) to get a computer to automatically create useful reference materials for language teaching, learning, and research, b) to create reference materials of higher quality than can be created without the aid of large corpora, c) to create software that is easy to use and that provides useful information to a user, and d) to create a product that is language and genre independent and thus can be utilized for various languages and genres. With regard to d), if you have a suitable corpus (e.g., Wikipedia) and NLP tools for a particular language, then you should be
able to automatically reproduce this reference tool for that language. Thus, although this application is designed for English, a similar application could be produced for another language if there are large corpora (such as Wikipedia) and NLP tools for that language.

In this paper, we describe the various features of this Corpus Reference Tool, we describe how these features work, and we discuss some applications for this tool. The organization of this paper is as follows. In section 2, we describe the primary resources that this application uses. In section 3, we describe the various features of this tool and some of their possible applications. Section 4 is the conclusion.

2. RESOURCES
In this section, we describe the primary resources that are used by this Corpus Reference Tool. This software was created in Python\(^1\) and incorporates corpora and some NLP tools.

This application makes use of the corpora shown in Table 1. The Wikipedia corpus\(^2\) consists of all of the English Wikipedia text at the time that the file was downloaded. We stripped away most tags via a Python script. Wikipedia, a huge corpus of natural language that covers a variety of subjects, serves as the primary source of information for this project.

We also used two other corpora for the purposes of creating lists of vocabulary terms. English Gigaword [3] is an enormous corpus that contains text from international English news services. The Professional English Research Consortium (PERC) corpus\(^3\) consists of 22 different corpora in various science and technology fields. We utilized two versions of this corpus: 1) the complete corpus, and 2) a portion of the PERC corpus consisting only of the Computer Science and Engineering sections.

In addition, we used the language resources for this project given in Table 2. Versions of the Porter Stemmer and a WordNet Lemmatizer, obtained from the Natural Language Toolkit [1], were used to extract morphological forms of tokens. We used the Stanford POS Tagger [6] to POS tag Wikipedia, and we used the Princeton WordNet Gloss Corpus\(^4\) to provide definitions for target vocabulary terms.\(^5\)

<table>
<thead>
<tr>
<th>Table 1: Corpora</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Corpus</strong></td>
</tr>
<tr>
<td>Wikipedia</td>
</tr>
<tr>
<td>English Gigaword</td>
</tr>
<tr>
<td>PERC Corpus</td>
</tr>
<tr>
<td>PERC Corpus: Computer Science and Engineering Sections</td>
</tr>
</tbody>
</table>

We next describe how we incorporated these various components into this Corpus Reference Tool.

3. FUNCTIONS
In this section, we describe the functions of the Corpus Reference Tool. We explain what the functions are, how they work, and what some of their potential uses are.

3.1 Search Functions
This Corpus Reference Tool has a GUI (Figure 1) that enables a user to find vocabulary terms and search for information about vocabulary terms.

Figure 1: Corpus Reference Tool GUI

A user can choose to view a word list which shows the most frequent words in different genres, extracted from corpora

\(^1\)http://www.python.org/

\(^2\)This was downloaded from http://en.wikipedia.org/wiki/Wikipedia:Database_download#English-language_Wikipedia on September 6, 2010 and consists of all "articles, templates, image descriptions and primary meta-pages" of Wikipedia.

\(^3\)http://scn.jkm21.com/~percinfo/

\(^4\)http://wordnet.princeton.edu/glosstag.shtml

\(^5\)The WordNet Lemmatizer and WordNet Gloss Corpus are tools that were created from WordNet [5], [2].
(see Table 1 for details about these corpora). The lists can be sorted in terms of frequency or alphabetically and stop words can be filtered out or presented (default is without stop words). In addition, a user can type into the text box a term of his/her choosing.

These wordlists can be used to find the most common vocabulary terms in particular genres of English, and thus they can be used for teaching English. For example, the most common terms in Wikipedia, due to its size, are, with some exceptions due to the nature of Wikipedia, most likely to be common words that are used in English in general.6

Once a user selects a word from a word list or inputs a word in the search box, a display appears (Figure 2 shows a window for the word ‘talk’). By clicking on the appropriate button on the left side of the window, the display enables a user to view the following types of information for a target term: Frequency, Word Forms, POS distribution, Definitions, Examples, and Collocations. We discuss these functions in the following sections.

Figure 2: Information Window for Vocabulary Term

3.2 Frequency
A user can view frequency information about a target word. The screenshots of GUIs in Figures 3 and 4 include frequency information for ‘show’ and ‘walk’. For example, of all unigrams that occur in Wikipedia, ‘show’ is the 118th most frequent, if you exclude stop words, and it is the 247th most frequent, if you include stop words. The tool also shows the total number of times that the word ‘show’ occurs and the percentage of the corpus (the entire Wikipedia text) that is taken up by ‘show’. Figure 4 shows that ‘walk’ is much less frequent; it is the 2606th most frequent word, if you exclude stop words.

The frequency of a word in Wikipedia, in most cases, is likely representative of the general frequency of that word in English, considering that Wikipedia is so large. Language teachers can use this information to help determine what vocabulary to teach and language researchers (linguists, lexicographers, etc.) who are interested in how often particular words are used may also find this information to be useful.

![Figure 3: show: Frequency, Word Forms, Part of Speech distribution](image)

![Figure 4: walk: Frequency, Word Forms, Part of Speech distribution](image)

3.3 Word Forms
The Corpus Reference Tool also allows a user to view word form (morphological) information. Figures 3 and 4 include word forms for ‘show’ and ‘walk’ that are automatically extracted from Wikipedia utilizing the Porter Stemmer and the WordNet lemmatizer. All tokens in Wikipedia were stemmed with the Porter Stemmer and lemmatized with the WordNet lemmatizer. Then for each token, lists of morphological forms were formed by a) finding the stem and lemma of the token, and b) finding all tokens that have an identical stem or lemma with the token. For example, the following tokens are grouped together since they all are stemmed or lemmatized as ‘show’ at least once: ‘show’, ‘showed’, ‘showing’, ‘shows’, ‘showings’, ‘showe’, ‘showes’, and ‘showss’. Similarly, the following tokens are all stemmed or lemmatized as ‘walk’ at some point: ‘walking’, ‘walked’, ‘walk’, ‘walks’, ‘walke’, ‘Walke’, ‘walkes’, ‘walkings’, ‘walkely’.

Note that the word form lists for target words includes some odd forms such as ‘showe’ and ‘walke’. The word ‘showe’ is stemmed/lemmatized as ‘show’, and thus it is considered a form of ‘show’. In (1a), ‘show’ is a misspelling of ‘showed’.

(1) (a) Researched showe7 that Piagetian intelligence is correlated but it is not identical with psychometric intelligence and IQ.

(b) “Eshowe”, is the oldest town of European settlement in Zululand. Eshowe’s name is said to be inspired by the sound of wind blowing through the more than 4 of the indigenous Dhlinza Forest, the most important and striking

7“Research showed . . . “
feature of the town. Although the name is most likely to be derived from the Zulu word for the xysmalobium shrubs, “showe” or “shongwe”.

“Walke” also appears in Wikipedia. It is stemmed and/or lemmatized as “walk”, thus resulting in ‘walk’ appearing in the word forms list for ‘walk’. Examples with “walke” are shown in (2a-b). (2a) contains the name ‘Walke’ and (2b) contains ‘Walke’ as part of the title of an old publication; this appears to be an old spelling of ‘walk’.

(2) (a) Three ships in the United States Navy have been named for Rear Admiral Henry A. Walke.

(b) *“A Weekes Worke, And A Worke For Every Weeke”, 1628, 1650? (“A Weekes Worke Containing Rules And Directions How To Walke In The Ways Of Godliness Both To God And to Men”)

Odd word forms, such as ‘showe’ for ‘show’ and ‘walke’ for ‘walk’ could be filtered, based on the fact that they tend to occur very infrequently in Wikipedia. For example, ‘walk’ appears 37,920 times and ‘walke’ appears only 109 times. The word ‘show’ appears 403,053 times and ‘showe’ only appears 12 times. However, these ‘odd’ word forms may be of use for identifying certain misspellings (1a), old spellings (2b), etc. This word form function could be improved so that it generally excludes incorrect word forms using algorithms to exclude names and misspellings. We leave this for future work.

Word forms are extracted automatically from Wikipedia and thus they likely give a good representation of possible morphological forms of target vocabulary terms. This information may be of interest to language teachers, learners, and researchers.

3.4 Part of Speech

This application enables a user to view the POS distribution for target words. To obtain the POS distribution, we tagged the entire Wikipedia corpus with the Stanford POS tagger. The large tagset (the Penn Treebank tags [4]) used for the Stanford POS tagger was simplified for ease of use. For each word type, we calculated the percentage of occurrences of each POS tag for that word; this distribution can be viewed by a user. POS tags that occur less than 3% of the time are not displayed. See Figures 3 and 4 for the POS distributions for the words ‘show’ and ‘walk’ respectively. Both of these words are used as nouns and verbs, and this distribution most likely gives a good picture of the general POS distribution of target words, since the distribution is calculated from an enormous corpus. This POS information can be useful for language researchers and educators, in giving a general idea of the POS distribution of target terms in English. In addition, as will be discussed in the following sections, this tool enables a user to view definitions and examples for particular POS versions of target terms.

3.5 Definitions

For each vocabulary term, a user can view definitions, when available, for particular POS tags of the term. The user chooses a particular POS tag for a word, and then definitions for that POS tag for that word are displayed. See Figures 5 and 6 (see the next page) for examples of definitions for the noun and verb forms of ‘show’ and of ‘walk’, respectively.

Definitions are extracted from the Princeton WordNet Gloss Corpus. This corpus is especially useful since it provides definitions that are categorized into adjectives, adverbs, nouns, and verbs. Thus, for words that have POS tags of these basic categories, definitions can be extracted. When the user chooses a particular definition of a target word, the program searches through the Princeton WordNet Gloss Corpus for the same word with the same POS tag, and prints the relevant definitions to the GUI.

3.6 Examples

One of the most useful components of this Corpus Reference Tool is its ability to show example paragraphs for particular POS categories for target words. Examples containing noun and verb versions of ‘show’ and ‘walk’ are presented in Figures 7 and 8. This function takes advantage of the structure of Wikipedia, in which text is stored in paragraphs (there is reflected in the POS distributions, with ‘show’ being used about 78% of the time as a noun and 22% of the time as a verb, and ‘walk’ being used about 75% of the time as a verb and 25% of the time as a noun.

This POS distribution most likely gives a good picture of the general POS distribution of target words, since the distribution is calculated from an enormous corpus. This POS information can be useful for language researchers and educators, in giving a general idea of the POS distribution of target terms in English. In addition, as will be discussed in the following sections, this tool enables a user to view definitions and examples for particular POS versions of target terms.

9 The Stanford POS tagger attains 96.97% accuracy on WSJ22-24 test set (http://nlp.stanford.edu/software/pos-tagger-faq.shtml). However, it is not completely accurate and numerous problems arise for POS tagging. The tagger tags “red” in (i) as a noun. Here, “red” could be considered an adjective that modifies “storm”, in which case the noun tag would be incorrect, or it could be considered to be part of a compound noun, in which case the tag would be correct. In (ii), “walk” is incorrectly tagged as a verb.

(i) “Red Storm” is a supercomputer architecture designed for the US Department of Energy’s National Nuclear Security Administration Advanced Simulation and Computing Program.

(ii) The 2009 Capital Challenge Walk MS raised half a million dollars for the National MS Society’s programs and initiatives.

10 Note that the definitions are presented in the order that they are found in the Princeton WordNet Gloss Corpus, and they are not necessarily in the order of the most common uses. For example, the first definitions that appear for the noun and verb forms of ‘walk’ are definitions related to baseball. The definitions related to moving feet do not appear until definition 3 for the verb form and definition 2 for the noun form. It may be desirable to order the definitions in a better manner.
are line breaks between paragraphs); thus paragraphs are simple to extract.

Examples are extracted in the following way. When the user chooses a particular POS tag for a particular word, the program searches through the Wikipedia corpus for an instance of the target word with the appropriate tag. If found, the paragraph containing the target word is printed to the GUI. A user can click on the same POS tag, or another POS tag, to view further example paragraphs. If a target word is frequent, then the user can potentially view many different example paragraphs.

This component of the tool can be especially useful for language teachers, learners, and researchers. Dictionaries and concordancers generally present, at most, sentence length examples. This Corpus Reference Tool presents paragraph length examples. Examples of this length contain more context than simple sentence length examples. In addition, these are examples of real language on various general topics, and thus, these examples can give a user a good idea of how a target word is used. Teachers can use these example paragraphs to develop lesson materials, students of English can use these examples to see how target words are used, etc.

3.7 Collocations

Lastly, this tool provides basic data on collocations for target words. A user can choose to view two or three word collocations, or without stop words. For a particular collocation, the user can view example paragraphs. Collocations, as defined for this tool, are simply frequent bigrams and
trigrams containing a target vocabulary word. Collocations with stop words may be useful for some users and not for others, so we’ve included the option of letting a user choose whether to view collocations with or without stop words. When a user chooses to view collocations, 10 collocations (the most frequent of the appropriate type) are displayed and the user can view 10 more collocations, etc., by clicking on the collocation type again. The user can also view example paragraphs containing a target collocation. Examples of 2 and 3 word collocations containing ‘show’, both without stop words, as well as example paragraphs containing the collocation “television show”, are shown in Figure 9.

![Collocations and Example Paragraphs with ‘show’](image)

The collocation function enables a user to see words that tend to occur together and to see example paragraphs with target collocations. This function can be useful for finding out what words tend to occur with a target word.

4. CONCLUSION

This Corpus Reference Tool provides a variety of information (word lists, frequency information, POS information, morphological forms, example paragraphs, collocations, and definitions) that is automatically extracted from corpora. We have presented a few examples of the output of this tool (see the Appendix for more examples). Wikipedia, from which most of this information is extracted, lies at the heart of this tool. Since Wikipedia is so large, this information is likely quite representative of how target vocabulary terms are used in general English.

One immediate way that this tool can be used is for development of language teaching materials. An instructor can use the tool to obtain word lists in a target genre and then find example paragraphs containing target words. These example paragraphs can be given to students to study, along with exercises. Exercises could include questions about the paragraphs that test for reading comprehension. In the future, we plan to add a function that automatically creates editable lesson materials. For example, example paragraphs for target terms could be automatically extracted and printed to a file, and vocabulary terms within a paragraph that may be useful to English language learners can be automatically identified, via the use of word frequency lists.

Lastly, we note that this research has applications for a variety of fields. It has applications for applied linguistics, since one purpose is to aid in the development of materials that can be used for teaching and learning languages. It has applications for corpus linguistics and possibly theoretical linguistics, since it can be used for investigating parts of speech, morphology, semantics, discourse analysis, etc. It has applications for computer science (e.g., natural language processing) since it requires development of algorithms to extract natural language information from corpora. It also has applications for artificial intelligence, since a goal is to get a computer to automatically create useful reference and research resources - automatic creation of reference materials could in some sense be likened to ‘intelligence’. Further research and development remains to improve this tool and apply it for various uses.

5. REFERENCES


---

11Paragraphs are currently printed out to GUI but they can be easily printed to a file or a user can copy directly from the GUI.
Appendix: Further Examples of Data

<table>
<thead>
<tr>
<th>Word</th>
<th>fly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Details</td>
<td>Freq.: 2793; POS: V (81%), V (18%); Forms: fly flies flying flis fli flyings flied flys</td>
</tr>
<tr>
<td>Ex.(V)</td>
<td>An Air Force analysis (Project Red Baron II) showed that a pilot’s chances of survival in combat dramatically increased after he had completed 10 combat missions. Red Flag was created in 1975 to offer US pilots the opportunity to fly 10 realistically-simulated combat missions in a safe training environment with measurable results. Many aircrews had also fallen victim to SAMs and Red Flag exercises provided pilots experience in this regime as well.</td>
</tr>
<tr>
<td>Ex.(N)</td>
<td>“Tenkara lines” are special lines used for the fixed-line fishing method of tenkara. Traditionally these are furled lines the same length as the tenkara rod. Although original to Japan, these lines are similar to the British tradition of furled leaders. They consist of several strands being twisted together in decreasing numbers toward the tip of the line, thus creating a taper that allows the line to cast the fly. It serves the same purpose as the fly-line, to propel a fly forward. They may be tied of various materials, but most commonly are made of monofilament.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Details</td>
<td>Freq.: 655; POS: N(81%), V(18%); Forms: studis studying study studied studiing studys studyings studi studyed studies studiess</td>
</tr>
<tr>
<td>Ex.(N)</td>
<td>After the revolution, Ramon Castro was forced to abandon the family farm as it became legal property of the state. He continued to dedicate himself to the study of agriculture however, and is chiefly responsible for many of Cuba’s agricultural initiatives since the revolution.</td>
</tr>
<tr>
<td>Ex.(V)</td>
<td>“Lydie Salvayre” (b. 1948) is a French writer. Born in the south of France to Republican refugees from the Spanish Civil War, she went on to study medicine in Toulouse and continues to work as a practicing psychiatrist. She has been awarded both the Prix Hermes and the Prix Novembre for her work.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>fast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Details</td>
<td>Freq.: 2177; POS: Adj (45%), Adv(37%), N(18%); Forms: fast fastnesses fasting fasts fasted fastness fastings fastes faste</td>
</tr>
<tr>
<td>Ex.(Adj)</td>
<td>The “Breviarium” was completed by Eutropius within 369: Paeanius’ translation appeared before 380. The “Breviarium” was a compendium of ancient Roman history, used both as a textbook in schools and as a fast course on Roman history for the higher social classes (it was dedicated to Emperor Valens): Paeanius’ translation allowed Greek-speaking people to have a graceful version of this compendium.</td>
</tr>
<tr>
<td>Ex.(Adv)</td>
<td>Aside from her powers, Gypsy is an expert in hand-to-hand combat stealth. She’s also an accomplished acrobat, able to leap high, run fast, swim, and execute unexpectedly quick martial arts tactics with relative ease. Gypsy also has a strong aptitude in electronics and computers, and has become skilled in the use of firearms. She has been trained by Bronze Tiger, and in issues of “Birds of Prey”, Lady Shiva has offered to train her as well.</td>
</tr>
<tr>
<td>Ex.(N)</td>
<td>Benjamin of Tudela mentions an Ahimaaz ben Paltiel in Amalfi in southern Italy, in the year 1162 (see his “Travels,” ed. Asher, i. 13, 14). This may well have been a descendant of his earlier namesake; for it is known that two brothers of the grandfather of Ahimaaz ben Paltiel were sent with presents to Paltiel by the prince of Amalfi. In a list of twenty-two selihah (elegiac) poets (Italy, fifteenth century?), Ahimaaz ben Paltiel is mentioned as the author of two poems; and a Mahzor of the Roman rite attributes to him a selihah for the Fast of Esther.</td>
</tr>
</tbody>
</table>

<sup>12</sup>Freq. = frequency rank, not including stop words; POS = POS distribution; Forms = word forms, Ex. = example