A Wikipedia-based English-Language Reference and Teaching Tool

Jason Ginsburg
University of Aizu
Tsuruga, Ikki-machi
Aizuwakamatsu, Fukushima, Japan
jginsbur@gmail.com

ABSTRACT

This paper describes software under development that automatically extracts information from corpora that can be useful for language learners and language researchers. This software provides users with lists of words (extracted from several large corpora) that are frequently used in various English genres. In addition, the software extracts the following types of information from Wikipedia: 1) word frequency: users can view lists of frequent words, 2) word forms: users can view different morphological forms of target words, 3) part of speech information: users can view the part of speech distribution of target words, and 3) example paragraphs: users can view sample paragraphs containing different usages of target words.

Keywords

corpora, Wikipedia, language teaching, language research

1. INTRODUCTION

Large natural language corpora can be excellent resources for language teaching, since they contain real language that a language learner could benefit from being able to understand. This paper discusses a reference tool that attempts to harness the power of large corpora (Wikipedia in particular) as a resource for language learners, as well as for language researchers. Wikipedia is particularly useful because it contains language written by many different people and it contains texts that cover a wide variety of subjects. The reference tool described in this paper extracts target information from corpora, stores this information, and allows users to easily access this information.

This project has several primary goals: a) get a computer to automatically create useful reference materials for language teaching, learning, and research, b) create reference materials of higher quality than can be created without the aid of large corpora, c) create software that is easy to use and that provides useful information to a user, and d) 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 corporus (e.g., Wikipedia) and appropriate Natural Language Processing tools for a particular language, then you should be able to automatically reproduce this reference tool for that language. Thus, although this reference tool is designed for English, a similar tool could be produced for another language if there are large corpora (such as Wikipedia) and natural language processing tools for that language.

In addition, this software could have applications for the developing world. It could be made available for free on the Web, and it could be used to help people in the developing world learn a language (English, etc.).

The organization of this short paper is as follows. In section 2, I describe the primary materials that this software is built with. In section 3, I explain the various functions of the software. Section 4 is the conclusion.

2. MATERIALS

This project utilizes the following corpora:

Table 1: Corpora

Corpus	Size (tokens)
English Wikipedia [3]	1,003,961,037 (after cleaning)
English Gigaword [2]	2,698,666,444
PERC Corpus [4]	27,211,335
PERC Corpus: Computer Science and Engineering sections	3,915,319

I downloaded Wikipedia and filtered out most Wikipedia tags. English Gigaword is an enormous corpus consisting of texts from international English news services. The Professional English Research Consortium (PERC) Corpus consists of texts from English academic journals in various fields of science and technology. PERC is comprised of 22 different corpora in various science and technology fields. I 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.

The main components of this software were created using the Python programming language. In addition, I utilized the following software from the Natural Language Toolkit [1] to extract relevant information from Wikipedia: a part-of-speech tagger, a stemmer (the Porter Stemmer) and a WordNet Lemmatizer.

3. FUNCTIONS

3.1 Display

A display (Figure 1), allows a user to make selections and conduct searches for various bits of information.

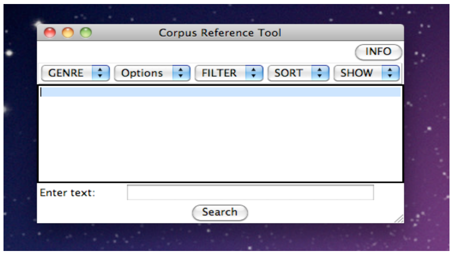


Figure 1: Display

3.2 Word list functions

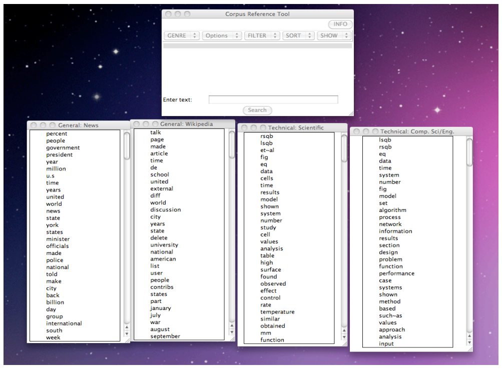
The “Genre” function (Figure 1) displays word lists, of various genres, which were created from the corpora given in Table 1. When a user selects a genre, a portion of the word list is shown. The word lists for each genre were created from the relevant corpora as shown in Table 2.

Table 2: Genres

Genre	Source	Size (types)
General: Wikipedia	Wikipedia	72,081,643
General: News	English Gigaword	55,522,640
Technical: Science	PERC	860,653
Technical: Computer Science /Engineering	PERC Computer Science and Engineering portions	101,657

These word lists for various genres can be useful for students and researchers who want to see what words tend to be used in particular genres of English. For example, the “Technical: Science” list can be useful for English language learners who are studying an area of science.

This software allows a user to manipulate a target word list using various functions (Figure 1). The “Options” function allows a user to choose whether or not to see the frequency of each word, as it is found in the particular corpus. For example, a user can view a list of frequent words in the “General: News” genre that also displays the frequency of each word (the frequency corresponds to the frequency of that word in the English Gigaword corpus). The “Filter” option allows a user to choose whether or not to include stop words (frequently occurring words such as *the*, *a*, *am*, *are*, etc.) in the word lists. The “SORT” function enables a user to choose whether words should be presented in order of frequency or in alphabetical order. The “SHOW” function allows a user to determine how many words are displayed in a list. A user can then select a word from a list for a target genre (see Figure 2). In addition, the “Search” box enables a user to input a target word of his/her choice (top of Figure 2).



General: News
percent
people
government
president
year
million
u.s
time
years
united
world
news
state
york
states
minister
officials
made
police
national
told
make
city
back
billion
day
group
international
south
week

General: Wikipedia
talk
page
made
article
time
de
school
united
external
diff
world
discussion
city
years
state
delete
university
national
american
list
user
people
contribs
states
part
january
july
war
august
september

Technical: Scientific
rsqb
lsqb
et~al
fig
eq
data
cells
time
results
model
shown
system
number
study
cell
values
analysis
table
high
surface
found
observed
effect
control
rate
temperature
similar
obtained
mm
function

Technical: Comp. Sci/Eng.
lsqb
rsqb
eq
data
time
system
number
fig
model
set
algorithm
process
network
information
results
section
design
problem
function
performance
case
systems
shown
method
based
such~as
values
approach
analysis
input

Figure 2: Choosing a target word

3.3 Search Functions

Once a user selects or inputs a word, a window appears that provides a user with target information about a word. Figure 3 shows a window for the word *analysis*.

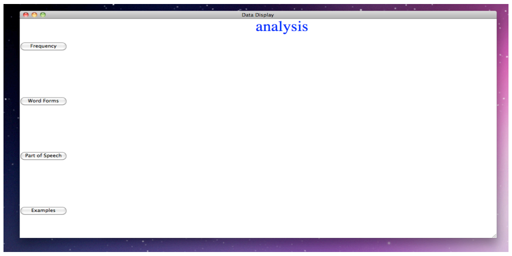
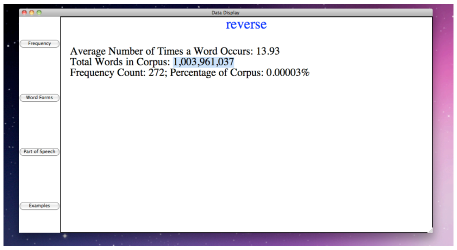


Figure 3: *analysis*

The user then can choose to view information, which is automatically extracted from Wikipedia, about the target word.

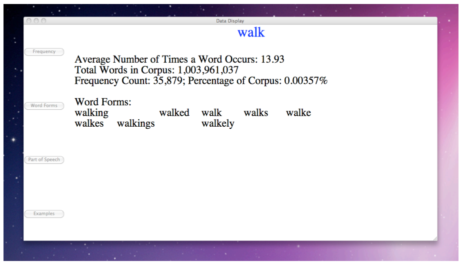
Clicking on the “Frequency” button (Figure 4) displays the frequency (and other relevant information) of the selected word, as calculated from the Wikipedia corpus. For example, the word *reverse* occurs 272 times in Wikipedia. Frequency information was obtained by calculating the unigram frequencies for all tokens in Wikipedia.



```
Average Number of Times a Word Occurs: 13.93
Total Words in Corpus: 1,003,961,037
Frequency Count: 272; Percentage of Corpus: 0.00003%
```

Figure 4: Frequency of *reverse*

The “Word Forms” button displays various forms of a word (Figure 5). This information is automatically extracted from Wikipedia in the following way. All words in Wikipedia were stemmed with the Porter stemmer and lemmatized with the WordNet lemmatizer. Then all words that have an identical stem/lemma were grouped together to create lists of morphological forms of each word in Wikipedia. Note that for *walk*, the algorithm finds some erroneous forms, such as *walke,* but this results from the fact that *walk*e appears in Wikipedia (since Wikipedia is so large, erroneous word forms, resulting from



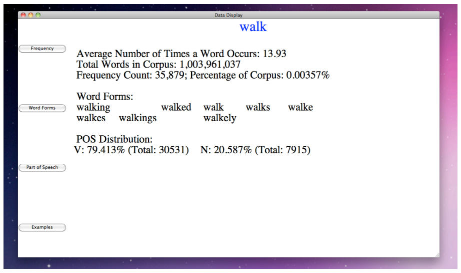
```
Average Number of Times a Word Occurs: 13.93
Total Words in Corpus: 1,003,961,037
Frequency Count: 35,879; Percentage of Corpus: 0.00357%

Word Forms:
walking              walked    walk      walks     walke
walkes    walkings             walkely
```

Figure 5: Word Forms for *walk*

misspelling, etc. appear).

The "Part of Speech" button displays the Part of Speech (POS) distribution of target words. This function is designed to show the frequency of the various POS categories for a target word, where POS labels and frequency of each POS label are automatically extracted from Wikipedia. In order to calculate this information, the entire Wikipedia corpus was tagged with the Natural Language Toolkit built-in POS tagger. Then I created a file that contains the POS frequencies for each tag for a particular token. For example, the POS distribution of the word *walk* is shown in Figure 6.



walk

Average Number of Times a Word Occurs: 13.93
Total Words in Corpus: 1,003,961,037
Frequency Count: 35,879; Percentage of Corpus: 0.00357%

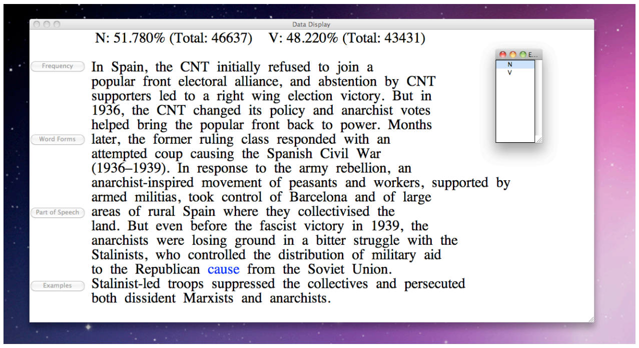
Word Forms:
walking walked walk walks walke
walkes walkings walkely

POS Distribution:
V: 79.413% (Total: 30531) N: 20.587% (Total: 7915)

Figure 6: Part of Speech distribution for *walk*

The "Examples" button presents a user with paragraph length texts containing target words. This function is unique. Concordancers are used for corpus research and language teaching. They provide at most sentence length examples. However, Wikipedia enables extraction of larger paragraph length examples. Examination of a paragraph enables a user to get a better idea of how a target word is used, since there is more natural language discourse presented to a user than would be presented with a sentence length example.

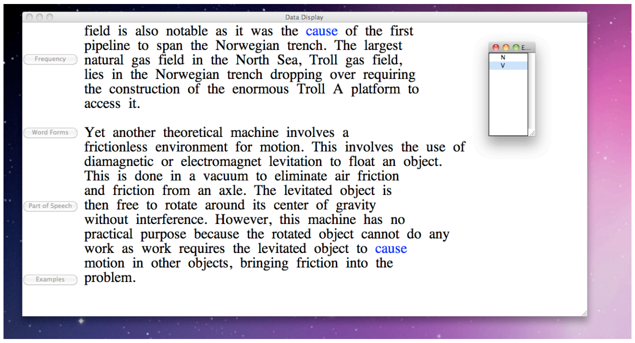
Example paragraphs are extracted in the following way. I created a file tha emulates a dictionary data structure, with keys and values. The key consists of a token and its part-of-speech tag; e.g., ('token', POS). Each key has a value, which is a list of all positions in all texts in which the particular token occurs with that particular POS tag; e.g., Value = ['1.txt','Line_29'], ['17.txt','Line_518', Line_519'], etc. An advantage of Wikipedia is that it has line breaks between paragraphs. This enables easy extraction of sample paragraphs. For example, Figure 7, shows an example paragraph for the word *cause* (which appears in blue text) with the POS tag Noun (N). Figure 8 shows an example paragraph for the same word with the POS tag Verb (V). A user can view further examples for a target word by clicking on the desired POS tag. Since examples are extracted from the enormous Wikipedia corpus, there are many example paragraphs for any frequent word.



N: 51.780% (Total: 46637) V: 48.220% (Total: 43431)

In Spain, the CNT initially refused to join a popular front electoral alliance, and abstention by CNT supporters led to a right wing election victory. But in 1936, the CNT changed its policy and anarchist votes helped bring the popular front back to power. Months later, the former ruling class responded with an attempted coup causing the Spanish Civil War (1936–1939). In response to the army rebellion, an anarchist-inspired movement of peasants and workers, supported by armed militias, took control of Barcelona and of large areas of rural Spain where they collectivised the land. But even before the fascist victory in 1939, the anarchists were losing ground in a bitter struggle with the Stalinists, who controlled the distribution of military aid to the Republican cause from the Soviet Union. Stalinist-led troops suppressed the collectives and persecuted both dissident Marxists and anarchists.

Figure 7: Examples: *cause* as a Noun



field is also notable as it was the cause of the first pipeline to span the Norwegian trench. The largest natural gas field in the North Sea, Troll gas field, lies in the Norwegian trench dropping over requiring the construction of the enormous Troll A platform to access it.

Yet another theoretical machine involves a frictionless environment for motion. This involves the use of diamagnetic or electromagnet levitation to float an object. This is done in a vacuum to eliminate air friction and friction from an axle. The levitated object is then free to rotate around its center of gravity without interference. However, this machine has no practical purpose because the rotated object cannot do any work as work requires the levitated object to cause motion in other objects, bringing friction into the problem.

Figure 8: Examples: *cause* as a Verb

4. CONCLUSION

This software provides frequency-based word lists, frequency information about target words, morphological forms of target words, POS information about target words, and examples containing target words. The current functions require further refinement and additional functions need to be added, such as a collocation finder and dictionary. In addition, the accuracy of the information presented in this software needs to be evaluated. I eventually hope to make this system available for students studying English.

Lastly, this research has applications for a variety of fields. It has applications for applied linguistics, since the purpose is to develop materials that can be used for teaching and learning languages. It has applications for 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 show how much this project can contribute to these fields.

5. ACKNOWLEDGMENTS

Thanks to the audience of the ACM Chapter Conference on eLearning and Technical Documentation for the Developing World for helpful comments and questions. I also thank the PERC Corpus project team for allowing me to use the PERC corpus.

6. REFERENCES

[1] Bird, S., Klein, E., and Loper. E. 2009. *Natural Language Processing with Python*. O'Reilly, Sebastopol, CA. http://www.nltk.org/

[2] Graff, D., Kong, J., Chen, Ke, and Maeda, K. 2007. E*nglish Gigaword Third Edition*. Linguistics Data Consortium, Philadelphia.

[3] http://en.wikipedia.org/wiki/Wikipedia:Database_download#English-language_Wikipedia

[4] http://www.corpora.jp/~perc