How to express your feelings
(with a little help from Google)

Shaoqun Wu
Department of Computer Science, University of Waikato
Hamilton, New Zealand
shaoqun@cs.waikato.ac.nz

ABSTRACT
This paper describes an attempt to capitalize on the vast amount of
human-generated text readily available on the Web to help
language learners express their own feelings. We avoid errors,
idiosyncrasies, and other dross by employing various language
and grammar filters. We sort words and phrases by frequency of
use to ensure that only very common usage is included. The
work is based on a huge collection of n-grams published by
Google, along with their frequency on the web. We describe a
comprehensive query tool that allows language learners and
teachers alike to locate a particular word, its associated phrases,
clause patterns, synonyms and antonyms. In addition, we have
enabled sample sentences containing these patterns to be
retrieved from the Web and presented to the user. Finally, five
language activities have been designed to help learners master
important vocabulary and expressions. The work is at a
preliminary stage and no user tests have yet been performed.

Categories and Subject Descriptors
J.1 [Administrative Data Processing]: Education

General Terms
Design, Human Factors, Management.

Keywords
Lexical acquisition, Language learning

1. Introduction
Everybody wants to talk about their own feelings. But in a
foreign language, it’s not easy. Language learners often
complain that they cannot express what they mean, think and
feel. For a simple question like “How are you today?” some
learners will answer factually (“My left ankle is sore from all
that dancing last night”), some cursorily (“I’m OK”) and some
facetiously (“I’m feeling sexy”). But language learners find it
hard to go beyond, to talk about their feelings at greater depth.
And the same applies to many other areas of self-expression.

Part of the reason is that learners have not experienced enough
of the appropriate language to express how they think, feel, or
react in ways that sound natural [3]. As Moskowitz [2] suggests,
curricular material tends to focus on facts and everyday
transactions, and only rarely touches on vocabulary that is
appropriate for communicating how we actually feel. To help
rectify this problem she proposes ways of integrating a
humanistic approach to language teaching with a planned
curriculum to promote study of self-actualization and self-esteem,
so that students can express themselves meaningfully and tell
you how they feel.

It is important to recognize the value of being able to talk about
your feelings, and to provide language learners with linguistic
resources—even limited ones—to express these central aspects
of daily life. But can computers help?

This paper proposes to identify common word sequences
starting with the word I, and use them to help learners acquire
important “I-vocabulary” and “I-expressions.” We focus on the
most commonly used English words and their associated
expressions because fluency does not grow out of acquiring an
extensive lexicon of difficult words, nor even from familiarity
with the most common words. Instead, it depends upon an
internalized repertoire of phrases and expressions composed of
common words and used in everyday life [4].

How can we capture the language of everyday life? Our
approach is to capitalize on the vast set of n-grams collected
from the Web that Google has recently made available.¹ Our
work is in three parts. First, we have developed a
comprehensive query tool that allows language learners and
teachers alike to locate a particular word, its associated phrases,
clause patterns, synonyms and antonyms. Second, we have
enabled sample sentences containing these patterns to be
retrieved from the Web and presented to the user. Third, five
language activities—Sentence heads, Semi-fixed expressions
with identifier, Related verbs, De-lexicalised verbs and Double
gapping and common adjectives—have been designed to help
learners master important vocabulary and expressions.

This work is being undertaken as part of a PhD research project
that I began one year ago. I have implemented the first part
mentioned above in the form of a specialized query tool called
Query-Express. The second part has been tested manually but
not yet completely automated—although it is not envisaged that
this will be a difficult step. The third part will fit into a more
general system for lexical acquisition that I am building, called
FLAX for the Flexible Language Acquisition Project.²

This paper describes the three components in turn. First we look
at the n-grams that Google has kindly supplied and explain how
to extract a useful sub-collection for language learning. We then

¹ The Google n-gram collection is available on six DVDs from
http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId
=LDC2006T13
² www.nzdl.org/flax

This paper was published in the proceedings of the New Zealand
Computer Science Research Student Conference 2008. Copyright is held
by the author/owner(s).

describe a specialized query tool that learners can use. Next we
describe how the results from the tool can be augmented by
retreiving relevant material from the Web. Finally we describe
the five language activities in which learners can participate.

2. N-grams from Google
The Google n-gram corpus contains word n-grams in English.
Their size ranges from unigrams (single words) to five-grams,
and their observed frequency counts on the Web are also
supplied. These counts were generated by Google from
approximately one trillion word tokens of text from publicly
accessible web pages. This is a staggeringly large corpus of
natural English. N-grams appearing less than 40 times were
discarded from the corpus before publishing it.

Table 1 shows the number of various different units in this
corpus. The number of n-grams increases as the value of n
increases; it peaks at the value of four and then begins to drop.
In the file supplied by Google, each n-gram occupies one line:

<table>
<thead>
<tr>
<th>word_1 &lt;space&gt; word_2 &lt;space&gt;… word_n &lt;tab&gt; count</th>
</tr>
</thead>
</table>

The corpus comprises approximately 24 GB of compressed text
files.

3. N-grams for language learning
Because language learning demands n-grams that include as
much contextual information as possible, n-grams of length one
to four were discarded. (This had the useful side-effect of
reducing the collection to a manageable size.) Then three further
selection steps were applied:

1. select five-grams that start with the word I
2. select five-grams that consist of the Brown vocabulary
3. select only grammatically correct five-grams.

In the first step, five-grams that start with the word I are
identified. All characters are transformed to uppercase, and
initial I AM is replaced with I AM. The second step filters each
word in the five-gram though a standard word list. This
eliminates n-grams that include non-word strings, and ones that
include unusual words—for only a certain range of vocabulary
is useful for language learning. We used a word list containing
the 47,224 words in the well-known million-word Brown
corpus of natural English [5].

Finally, the remaining five-grams are parsed into linguistic
phrases by the OpenNLP sentence chunker [6] to identify
grammatically correct ones. This eliminates ill-formed
expressions such as I all most passed out that contain grammar
errors. To do this a string matching algorithm is defined using
the following regular expression:

\(\langle\text{NP}\rangle\langle\text{ADVP}\rangle\langle\text{VP}\rangle\langle\text{PP}\rangle\langle\text{ADJP}\rangle\langle\text{PRT}\rangle\langle\text{S}\rangle\langle\text{BAR}\rangle\)

According to this expression, a grammatically correct
expression must start with a noun phrase, indicated by NP, plus
optional adverb phrases (ADVP), a verb phrase (VP), and
optionally, a terminating noun, prepositional (PP), adverb,
adjecive (ADJP), particle (PRT) phrase or clause (SBAR).

Examples of valid five-grams are

[NP I] [VP AM] [ADJP AFRAID] [VP TO ANSWER]
[NP I] [ADVP ENTHUSIASTICALLY] [VP RECOMMENDED] [NP IT]

In these examples, the syntactic tags that the OPENNLP
chunker assigns to individual words have been removed for
clarity. An original chunked five-gram includes these tags, and
looks like:

[NP I/PRP] [VP ASKED/VBD] [PP FOR/IN] [NP A/DT
ROOM/NN]

In total, these three selection steps reduced the number of five-
grams from the 1.2×10^9 mentioned above to just under 100,000.

4. Querying the collection
The idea behind this project is to use the collection of selected
n-grams as a resource for language learning. We now examine
how learners employ the specialized query tool, Query-Express,
to study the language patterns that are associated with a
particular word. For example, one might study which adverbs
are frequently used by native speakers when they say they are
disappointed about something.

4.1 Particular words
We have identified five ways in which learners can examine
the usage of a particular word: its prefix and suffix patterns; phrases
and clauses in which it occurs; and its synonyms and antonyms.
We examine each one in turn.

4.1.1 Prefix patterns
Given a word, learners can study the language patterns that most
frequently precede it. The word is submitted to Query-Express
along with its syntactic class (such as verb or adjective) and the
number of results to return. For example, Table 2 shows the
result of querying for prefix patterns associated with the
adjecive disappointed, sorted by frequency. One striking
finding is the prevalence of the modifier very: nearly half of the

Table 1. Number of units in the Google n-gram collection

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>tokens</td>
<td>1,024,908,267,229</td>
<td>10^{12}</td>
</tr>
<tr>
<td>sentences</td>
<td>95,119,665,584</td>
<td>0.95×10^{9}</td>
</tr>
<tr>
<td>unigrams</td>
<td>13,588,391</td>
<td>0.014×10^{9}</td>
</tr>
<tr>
<td>bi-grams</td>
<td>314,843,401</td>
<td>0.3×10^{9}</td>
</tr>
<tr>
<td>trigrams</td>
<td>977,069,902</td>
<td>1.0×10^{9}</td>
</tr>
<tr>
<td>four-grams</td>
<td>1,313,818,354</td>
<td>1.3×10^{9}</td>
</tr>
<tr>
<td>five-grams</td>
<td>1,176,470,663</td>
<td>1.2×10^{9}</td>
</tr>
</tbody>
</table>

Table 2. Prefix patterns associated with the adjective disappointed

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>I AM VERY DISAPPOINTED</td>
<td>47588</td>
</tr>
<tr>
<td>I WAS VERY DISAPPOINTED</td>
<td>33238</td>
</tr>
<tr>
<td>I AM DISAPPOINTED</td>
<td>31236</td>
</tr>
<tr>
<td>I AM EXTREMELY DISAPPOINTED</td>
<td>11599</td>
</tr>
<tr>
<td>I WAS SO DISAPPOINTED</td>
<td>9842</td>
</tr>
</tbody>
</table>

DJP|PRT|SBAR)*
How to Express Your Feelings (With a Little Help from Google)

208,691 occurrences of the word disappoint in the collection co-occur with this adverb. For verbs, learners can also specify to Query-Express a word family to which the verb belongs. A word family consists of a headword along with its inflected forms. For example, the verb enjoy’s family members are: enjoyed, enjoying, enjoys.

Query-Express adopts a straightforward method of determining the inflection forms of verbs. It simply combines three word family lists that have been downloaded from the Complete Lexical Tutor website:

1. the most frequent 1000 headwords
2. the most frequent 2000 headwords
3. Coxhead’s Academic Words [1].

This approach may seem limited because the combined list covers only about 2,560 headwords. However, this seems adequate for the purpose of language learning because our focus is to help learners master the most common words.

The results of querying the verb enjoy with and without the word family option are shown in Table 3. Without it, phrases that precede the word enjoy are shown. With it, additional results appear that correspond to syntactic variants of the term, like enjoyed and enjoying. People seem remarkably positive about their enjoyment—only one negative statement (I did not enjoy) occurred in the ten most common “enjoy-grams.”

Prefix Frequency
I ENJOY 507952
I DID NOT ENJOY 107616
I REALLY ENJOY 84977
I DID ENJOY 42446
I ALSO ENJOY 24986

Table 4 Top patterns that follow the words love and hate

Suffix Frequency
HATE MYSELF AND WANT 15713
EVERYTHING ABOUT YOU 7408
IT WHEN 5570
MYSELF FOR LOSING 4981
FRANK IF I 4866
LOVE YOU BECAUSE I 40814
YOU LIKE CRAZY 21264
YOUR WORK THE 18421
A MAN 9922
THOSE SHOES 8017

Table 5 Phrasal verbs of the verb go

Phrasal verbs Frequency
I AM GOING OUT 70584
I WENT ON 34118
I WENT IN 27603
I AM GOING DOWN 27471
I AM GOING UP 17183
I AM GOING BACK 16263
I AM GOING OFF 13260
I AM GOING OVER 7866

Table 6 Clauses involving the verb believe

Clause patterns Frequency
THAT IT IS 108363
THERE IS A 58841
THIS IS THE 49540
IT IS A 39937
IT WOULD BE 29736
THAT I HAVE 24277
THAT IF YOU 23306
THAT WE SHOULD 20044

208,691 occurrences of the word disappoint in the collection co-occur with this adverb.

For verbs, learners can also specify to Query-Express a word family to which the verb belongs. A word family consists of a headword along with its inflected forms. For example, the verb enjoy’s family members are: enjoyed, enjoying, enjoys.

Query-Express adopts a straightforward method of determining the inflection forms of verbs. It simply combines three word family lists that have been downloaded from the Complete Lexical Tutor website:

1. the most frequent 1000 headwords
2. the most frequent 2000 headwords
3. Coxhead’s Academic Words [1].

This approach may seem limited because the combined list covers only about 2,560 headwords. However, this seems adequate for the purpose of language learning because our focus is to help learners master the most common words.

The results of querying the verb enjoy with and without the word family option are shown in Table 3. Without the option, phrases that precede the word enjoy are shown. With it, additional results appear that correspond to syntactic variants of the term, like enjoyed and enjoying. People seem remarkably positive about their enjoyment—only one negative statement (I did not enjoy) occurred in the ten most common “enjoy-grams.”

### 4.1.2 Suffix patterns

Prefix patterns allow users to find out what words and phrases precede a particular word. Conversely, suffix patterns allow them to examine the words and phrases that follow it. For example, Table 4 shows the top five patterns that follow the word love and hate. Such queries not only tell us the most common things that people love or hate, but also help language learners choose appropriate words when they want to express similar feelings. It is interesting that people like to say they hate themselves! Also, there is evidence here that women tend to talk more about their feelings than men (a man, and maybe those shoes).

---

3 In contrast, my supervisor tells me to avoid the word very in academic writing.

4 [www.lexutor.ca/lists_download](www.lexutor.ca/lists_download)
Phrasal verbs comprise a verb combined with an uninflected preposition, adverb, or adverbial particle [7]. These constructions present learners with a challenging task because there are many of them, and some are difficult to master. For example, learners often become confused by colloquial expressions involving phrasal verbs, such as *go off* and *get off*. Query-Express allows learners to retrieve n-grams that contain phrasal verbs. They do this by specifying a particular verb, for example, *go*. As shown in Table 5, *go* is frequently used as a phrasal verb and can be associated with most prepositions, including *out*, *on*, *in*, *down*, *up*, *back*, *off*, and *over*.

### 4.1.4 Clauses
Clause patterns that are commonly associated with a verb constitute the basic sentence structures, and learning them naturally helps improve learners’ communicative fluency. Query-Express allows clause patterns involving a particular word to be found. For example, Table 6 shows the result of seeking clauses involving the verb *believe*. These are useful sentence structures that learners can use in conversation (or in writing) when they want to say that they believe something.

### 4.1.5 Synonyms and antonym
How can we help learners express feelings that they find difficult to articulate, such as sadness, in rich and appropriate ways? We are experimenting with the use of WordNet to retrieve synonyms or antonyms of a particular word. WordNet is a large lexical database of English, in which nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Once synonyms and antonyms of a given word have been determined, they can be issued as query terms to the n-gram collection to determine whether they are actually used by native speakers, and if so, how.

At the time of writing, this has not been fully implemented. The results shown below were obtained by manually querying the adjective *sad* in WordNet, and using its synonyms and antonyms to retrieve samples from the n-gram collection. WordNet returns these words:

| sad synonyms: | bittersweet, depressing, depressive, gloomy, saddened, doleful, mournful, heavyhearted, melancholy, melancholic, pensive, wistful, tragic, tragical, tragicoical, sorrowful, deplorable, distressing, lamentable, pitiful, sorry |

When they are looked up in the collection, only *depressing*, *gloomy*, *melancholy*, *pensive*, *pitiful*, *sorry*, *glad*, *joyful*, and *good* occur. The phrases found are shown in Table 7. This tool can help learners eliminate idiosyncrasy like “Gadzooks! I am feeling aright today.”

The whole process will be automated once we have tested and refined this idea.

### 4.2 Specific word classes
Learners can issue queries for a particular word class, although currently the system only supports verbs and adjectives. For example, the top ten most frequently used adjectives are:

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>SURE</td>
<td>744952</td>
</tr>
<tr>
<td>I AM PRETTY SURE</td>
<td>439738</td>
</tr>
<tr>
<td>I AM QUITE SURE</td>
<td>133955</td>
</tr>
<tr>
<td>I WAS SURE</td>
<td>60761</td>
</tr>
<tr>
<td>GLAD</td>
<td>3149592</td>
</tr>
<tr>
<td>I AM SO GLAD</td>
<td>765576</td>
</tr>
<tr>
<td>I AM VERY GLAD</td>
<td>199951</td>
</tr>
<tr>
<td>I AM REALLY GLAD</td>
<td>138388</td>
</tr>
<tr>
<td>I AM JUST GLAD</td>
<td>112529</td>
</tr>
<tr>
<td>I AM ALSO GLAD</td>
<td>33553</td>
</tr>
</tbody>
</table>

Any given n-gram could contain more than one verb or adjective. However, only the main ones are extracted. Identifying these involves two steps:

1. identify main verb or adjective phrases
2. extract verbs or adjectives.

Main verb phrases are identified using this pattern:

- noun_phrase + adverb_phrases (optional)
- + verb_phrase (main)

For example,

- [NP I/PRP] [ADV enthuisiastically/RB] [VP recommend/VBP] [NP it/PRP]

The corresponding patterns for adjectives are:

- noun_phrase + adverb_phrases (optional)
- + verb_phrase + adverb_phrases (optional)
- + adjective_phrase (main)

---

5 See [http://wordnet.princeton.edu](http://wordnet.princeton.edu) for more information.
How to Express Your Feelings (With a Little Help from Google) 181

Adjectives are located by searching for JJ tags in main adjective phrases, and verbs by seeking VBP, VBD, VBN tags in main verb phrases.

5. Retrieving samples from the Web

It is of paramount importance to learn language in context. Consequently the extracted n-grams are of little pedagogical value in themselves because their context is lost when they are removed from their original document. The situation can be remedied by providing learners with a large volume of more extensive language samples that contain the phrases in question. For example, samples could be obtained from a general corpus such as the British National corpus, or from corpora provided by teachers.

In this project, we use the Web as the source of samples and develop a special retrieval tool that suits our purpose. This program takes a query term and submits it to the Google search engine; then uses a web mirroring program\(^6\) to retrieve the pages corresponding to the top few search results. These pages are parsed and the sentence that contains the target n-gram is extracted, along with the following sentence.

We illustrate the procedure using the fragment I was very happy because. A query

\[
\text{http://www.google.co.nz/search?hl=en}
\quad \text{&q=%22I+was+very+happy+because%22}
\]

is constructed and sent to the Google search engine. The returned page giving the search results is shown in Figure 1. From this, the following language fragments are extracted:

\begin{quote}
On my first day at S.M.S.H., I was very happy because I could meet my friends, and my mother and father work here. But I was a little bit nervous ...
\end{quote}

\begin{quote}
I was very happy because at least I can go home and I can stay as much as I want and still come back. I came here on Pakistani passport but with the British ... www.movingaway.org.uk/stories/story125/story123 story124.htm - 14k - Cached - Similar pages
\end{quote}

\begin{quote}
I was very happy because he is a very handsome man. My parents were also happy to see him even though he comes ... www.abcnews.com/option/index.php? a=lenews/women_i_want_to_leave_the_people_who_brought_me - 15k - Cached - Similar pages
\end{quote}

\begin{quote}
The Church of Jesus Christ of Latter-day Saints (Hong Kong, China) I was very happy because Heavenly Father answered my prayer. My mom taught me that we have to thank our Heavenly Father even if we do not get perfect scores ... www muchos.com/english/wordmem/onomast/corpus/ielts_samples/1010-primary.htm - 18k - Cached - Similar pages
\end{quote}

I am currently experimenting with reliable ways of extracting an appropriate amount of text.

6. Using the collection in language activities

Language learners can examine the n-grams and corresponding samples obtained from the web directly using the Query-Express query tool. Perhaps more usefully, this information can serve as material for exercises that are generated automatically by the system. We demonstrate five language activities whose exercises are generated in this way.

6.1 Sentence Heads

This activity focuses on common (and therefore important) sentence heads. It requires learners to match the first part of a sentence with its ending. The exercises are generated in three steps:

1. selecting the target sentence heads
2. retrieving sample sentences from the Web
3. splitting these sentences into two parts.

The person constructing the exercise—for example, a teacher—submits a list of words that provides the focus for the exercise. For example, they might choose the list wondering, trying, feeling, doing; the system will retrieve the n-grams containing those words. The teacher then has an opportunity to decide which sentence head patterns should be used in the exercise (if they do not, the most frequent patterns are used by default). Then the system retrieves sample text from the Web. It locates target sentences, splits them into two parts, and scrambles the parts.
example, common qualifiers for
qualifiers that are most frequently associated with
them. To retrieve the n-grams containing them, along
with the
language learners tend
to formulate more complicated expressions to achieve
these purposes—expressions that could easily be replaced by a simple
modifier like just. This section describes an exercise designed to
help learners master these common and useful qualifiers in
order to make what they are trying to express sound strong,
weak or negotiable in a simple and natural way.

The teacher prepares a list of feeling-related words such as
annoyed, grateful, upset, and disappointed. The system uses
these to retrieve the n-grams containing them, along with the
qualifiers that are most frequently associated with them. For
example, common qualifiers for disappointed include so, very,
quite, pretty, rather. Then the teacher manually sorts the
qualifiers into categories according to their degree of strength.

Figure 3 shows a word selection and placement exercise that
asks learners to use identifiers to make expressions stronger or
weaker.

6.3 Related verbs

Learners often find it difficult to differentiate between words
with similar meanings, such as speak, talk, say, tell. This
activity is designed to help them distinguish one from the other
by studying collocation and sentence structures with which they
are commonly associated.

Given a list of such words, their most frequently used 5-grams
are retrieved. For example, Table 9 shows the top three most
commonly-used 5-grams of speak, talk, say, tell. Then sample
text containing these fragments is collected from the Web, and
this is used to construct the fill-in-the-blanks exercises shown in
Figure 4.

6.4 De-lexicalised verbs

One of the best ways to make your spoken English more natural
is to learn expressions that use the verb get. This is generally a
far more productive way for learners to spend their time and
energy than studying unusual new words.

We are working on automatically constructing different kinds of
get-related exercises. For example, Figure 5 shows an exercise
that asks learners to put the verb into its proper place, while
Figure 6 shows an exercise that asks learners to complete a
sentence. For each one the system begins by retrieving the most
frequently used five-grams that contain get and its inflected
forms gets and got.

6.5 Double gapping and common adjectives

It is very helpful for learners to enrich their vocabulary
knowledge by learning pairs of adjectives that commonly appear
together. Given two such adjectives, exercises can be
constructed by using n-grams that contain them joined by the
word and. Adjectives and sample sentences can be used to
create "double gapping" exercises such as the ones shown in
Figure 7.

7. Conclusion

This paper has described an attempt to use computers to help
language learners express their own feelings. The key, of course,
is not the computer per se but the vast amount of human-
generated text that is readily available on the Web. A second
essential enabling factor is the ability to avoid errors,
idioms, and other dross not by detecting them directly
(which would be very difficult) but using various language and
grammar filters. A third is the ability to sort words and phrases
by their frequency of use to ensure that only very common
usage is included. This is all made possible by the use of a huge
collection of n-grams, along with their frequency on the web:
we use the collection published by Google.
How to Express Your Feelings (With a Little Help from Google)

We are experimenting with the use of common word sequences starting with the word I. We have argued that these “I-grams” have real value: they can enable learners to study the usage of the most common words and expressions, and thereby help them express themselves articulately. Google’s n-gram collection reveals many interesting things, like what most people are worried about or afraid of, what they enjoy, love, and hate, what disgusts them, and so on. These are of great value not only to sociologists and psychologists, but also to language learners for the purpose of enriching their expressive repertoire.

The potential of this project goes well beyond what has been presented in this paper and intend to carry out further studies on linguistic features of these I-grams. For example, native speakers rarely use the Present Perfect tense when they say they are disappointed about something (I have been disappointed …), but they do use it when they say they enjoy something (I have enjoyed …). We are not quite sure where this project will lead us, but we are sure that it will transcend the question of expressing feelings which is the subject of this paper.

Many fictional robots in books and movies are able to feel “as a person does.” But in the science, despite 40 years of research in the area of artificial intelligence, little progress has been made on the realistic emulation of emotion. Surprisingly, it seems that though they may never have feelings of their own, computers can nevertheless help language learners express their feelings precisely and eloquently and thus participate more meaningfully in society.

8. Acknowledgements
I gratefully acknowledge funding from the NZ Marsden Fund and Foundation for Research, Science and Technology. I would also like to thank my supervisor Ian Witten for helping me to express myself in this paper.

9. References
[5] The Brown Corpus was the first of the modern, computer readable, general corpora. It was compiled by W.N. Francis and H. Kucera, Brown University, Providence, RI. See http://www.essex.ac.uk/linguistic/clmt/w3c/corpus_content/corpus_list/private/brown/brown.html for more information.
[6] OpenNLP hosts a variety of java-based NLP tools which perform sentence detection, tokenization, post-tagging, chunking, etc. See http://opennlp.sourceforge.net for more information.
[7] The definition of phrasal verb can be found at http://en.wikipedia.org/wiki/Phrasal_verb