

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

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).

NZCSRSC 2008, April 2008, Christchurch, New Zealand.

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 student 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

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

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

describe a specialized query tool that learners can use. Next we show how the results from the tool can be augmented by retrieving 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:

word_1 <space> word_2 <space>... word_n <tab> count

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'M* is replaced with *I AM*. The second step filters each word in the five-gram through 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:

Table 2. Prefix patterns associated with the adjective *disappointed*

Prefix	Frequency
I AM VERY DISAPPOINTED	47588
I WAS VERY DISAPPOINTED	33238
I AM DISAPPOINTED	31236
I AM EXTREMELY DISAPPOINTED	11599
I WAS SO DISAPPOINTED	9842

$\backslash\backslash[NP\\s+((ADVP\\s+)*\\backslash[VP\\s+\\backslash[(NP|PP|ADVP|ADJP|PRT|SBAR)*$

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, adjective (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 adjective *disappointed*, sorted by frequency. One striking finding is the prevalence of the modifier *very*: nearly half of the

Table 3 Result of querying the verb *enjoy* (a) with and (b) without the word family option

Prefix	Frequency	Prefix	Frequency
I ENJOY	507952	I ENJOYED	712745
I DID NOT ENJOY	107616	I AM ENJOYING	118490
I REALLY ENJOY	84977	I HAVE ENJOYED	30335
I DID ENJOY	42446	I THOROUGHLY ENJOYED	16745
I ALSO ENJOY	24986	I ESPECIALLY ENJOYED	12997

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.³

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.⁴

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.lextutor.ca/lists_download

Table 7 Phrases containing synonyms and antonyms of *sad*

Five-grams	Frequency
I FIND EVEN MORE DEPRESSING	66
I GET GLOOMY WITH THE	172
I AM AS MELANCHOLY AS	105
I CONTINUED PENSIVE AND SAD	109
I AM JUST A PITIFUL	175
I AM SORRY TO SAY	86443
I AM GLAD TO SEE	156084
I AM JOYFUL TO CONFIRM	143
I AM NOT VERY GOOD	50920

4.1.3 Phrases

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.⁵ 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, saddening, doleful, mournful, heavyhearted,*

Table 8 The result of querying adjectives

	Suffix	Frequency
SURE	I AM SURE	7440952
	I AM PRETTY SURE	439738
	I AM QUITE SURE	133955
	I WAS SURE	60761
GLAD	I AM GLAD	3149592
	I AM SO GLAD	765576
	I AM VERY GLAD	199951
	I AM REALLY GLAD	138388
	I AM JUST GLAD	112529
	I AM ALSO GLAD	33553

melancholy, melancholic, pensive, wistful, tragic, tragical, tragicomic, tragicomical, sorrowful, deplorable, distressing, lamentable, pitiful, sorry

sad antonyms: *glad, joyful, good.*

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:

sure, glad, happy, sorry, interested, able, proud, afraid, aware, willing.

The result of querying adjectives is a list of adjectives and their associated sentence heads, sorted by frequency, shown in Table 8.

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] [ADVP **enthusiastically/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> for more information.

Web

[firstday](#)
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. ...
phomber.com/bsc/firstday/firstday.htm - 12k - [Cached](#) - [Similar pages](#)

[Moving Here | Stories | The Struggle to Settle](#)
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.movinghere.org.uk/stories/story243/story243.htm - ProjectNo=10&Identifier=stories/story243/story243.htm - 14k - [Cached](#) - [Similar pages](#)

["I want to leave the people who brought me..." / Women / News ...](#)
Then one day, the husband also returned, and I was very happy because he is a very handsome man. My parents were also happy to see him even though he comes ...
www.africansuntimes.com/index.php?ast/news/women/i_want_to_leave_the_people_who_brought_me - 16k - [Cached](#) - [Similar pages](#)

[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.lds.org.hk/english/members_column/random_sampler/yllst-primary.htm - 18k - [Cached](#) - [Similar pages](#)

Figure 1 Results of querying Google for *I was very happy because*

For example,

[NP I/PRP] [VP AM/VBP] [ADVP NOT/RB]
[ADJP SO/RB SURE/JJ]

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 program⁶ 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

<http://www.google.co.nz/search?hl=en&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:

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.

⁶ We use Wget, a network utility, to retrieve files from the Web.

Matching Beginnings with Endings

Beginnings	Endings
1. I am just wondering if	a. should be doing
2. I am trying hard to	b. you have to be very good in math to become a pyrotechnition?
3. I get the feeling that	c. be nice but cheney and bush are really pushing my buttons what do I do about it?
4. I am doing what I	d. everyone needs a next big thing, and if there is not one, they create it.

Figure 2 The Matching Beginnings with Endings exercise

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 ...

Then one day, the husband also returned, and I was very happy because he is a very handsome man. My parents were also happy to see him even though he comes ...

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 then 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.

Make the expression stronger by choosing a word (or words) in brackets and putting it in the right place:

1. I am upset because this might be the last time he is in the films (really, all).
2. I was disappointed but there was nothing I could do about it. (bitterly, absolutely)

Make the expression weaker by choosing a word (or words) in brackets and putting it in the right place:

1. I was upset when they told me. (extremely, a bit)
2. I was annoyed that he was late. (somewhat, rather)

Figure 3 The Choosing the right words exercise

Figure 2 shows a matching exercise that uses *wondering*, *trying*, *feeling*, *doing* in turn. This information is presented on-screen in the form of an interactive exercise, and the learner's job is to match which beginning goes with which ending.

6.2 Semi-fixed expressions with qualifiers

Native speakers tend to use qualifiers such as *quite*, *really*, *so* and *just* to strengthen or weaken the feeling they are expressing, or to make an expression more indirect. 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.

Use one of the words *speak*, *tell*, *say*, *talk* to complete these sentences

1. I got to _____ to him a bit and he was such a friendly guy.
2. First of all I can not _____ for him, Love was one subject matter that he knew a lot about, the good and the bad parts of it.
3. I like this guy but I can not _____ if he likes me back.
4. I am sorry to _____ I don't find that very helpful

Figure 4 The fill-in-blanks exercise of *speak*, *talk*, *say*, *tell*

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, idiosyncrasies, 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.

Put the verb <i>get</i> in its proper place
1. I a chance to talk to Bun B today and do a little interview got .
2. I a kick out get .
3. Perhaps when I finally around to watching Alias, it will change my point of view get .
4. I always something from my mom, despite my age get .

Figure 5 The Rearrange exercise of *get*

Use one of these word pairs to complete the sentence: <i>proud and glad, willing and able, small and despised, confident and sure, and fine and dandy.</i>
1. I am _____ of myself, or so I seem.
2. I am _____ to be a Singaporean.
3. I am _____ to work with the school children's needs.
4. I am _____.
5. I am _____, but I do not forget your firm advice.

Figure 7 The Double Gapping exercise of adjectives

Complete the sentence in a way that makes it true for you:
1. I get upset when I
2. I get very impatient if

Figure 6 The Sentence Completion exercise of *get***Table 9 The top three most commonly used five-grams of *speak, talk, say, tell***

speak	I can not speak for I did not speak out I am speaking of the
talk	I am not talking about I am talking to you I got to talk to
say	I can not say I I am sorry to say I am not saying it
tell	I can not tell you I can not tell if I am here to tell

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

- [1] Coxhead, A. (1998) *An academic word List*. Occasional Publication Number 18, LALS, Victoria University of Wellington, New Zealand.
- [2] Moskowitz, G. (1978) *Caring and sharing in the Foreign Language class*. Heinle & Heinle.
- [3] Lewis, M. (1997) Implementing the lexical approach: putting theory into practice. *Language Teaching Publication*.
- [4] Lewis, M. (1993) *The lexical approach*. *Language Teaching Publication*, England.
- [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/corpora_ling/content/corpora/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