Lexical Profiling Using the Shogakukan Language Toolbox

Takahiro Nakamura
Shogakukan, Inc.
2-3-1 Hitotsubashi Chiyoda-ku
Tokyo 101-0081, Japan
takahiro@shogakukan.co.jp

Yukio Tono
Meikai University
8 Akemi, Urayasu
Chiba 279-8550, Japan
y.tono@meikai.ac.jp

Abstract

The Shogakukan Language Toolbox is a natural language processing environment in which a suite of text processing tools and commands are made available via the web interface for the purpose of supporting corpus-based research or dictionary making.

One of the unique features of this toolbox is its integrated batch processing mode. This kind of mode is indispensable for lexicographic work. Ordinary windows-based concordancers are suitable for searching individual items in an exploratory way, but such an interactive, manual mode of searching is not efficient when it comes to extracting the patterns for large number of items at one time. No other web-based environment has this feature to our knowledge.

We will demonstrate the case of extracting from the British National Corpus a frequency list of verb subcategorization frames based on the COMLEX Syntax Lexicon 3.0.

1 Introduction

Recently, what are called "corpus-based" English bilingual dictionaries have come onto the market in Japan, for example, *Taishukan's Unabridged GENIUS English-Japanese Dictionary* (2001) and *Sanseido's WISDOM English-Japanese Dictionary* (2003). Shogakukan, a general publishing house in Japan, has been aware of the importance of corpus-based lexicography for some time. They developed an in-house corpus query system called the Shogakukan Corpus Query System (CQS) in 1999 (Tono et al. 2001), part of which has been commercially available for public access[1]. Despite such an improvement in the language resource environment, proper processing of corpus data for the purpose of revising a dictionary requires much expertise and time. Experience shows that the revision of even a single entry on the basis of corpus data can be time-consuming and can demand a lot of the lexicographers. Whilst this problem is not easy to solve, in this paper we report on one investigation into how much and what sort of summarization of the corpus data can be of help to a lexicographer in revising entries efficiently and accurately.

Normally, what happens is that if a lexicographer is given a list of entries for revision and is asked to use corpora for the task, he or she would have to search for each word one at a time, saving the results after each query. It is possible, however, to carry out an automatic search of all the entries on the list at once, and save the results in a desired format, such as a KWIC concordance or a collocation table. Lexicographers then have a much simpler job, as they can straightaway get to work and focus on their task of examining the output files and revising the entries, rather than spending time on the mechanics of searching and saving corpus data.

[1] Shogakukan Corpus Network (http://www.corpora.jp/).

For such a work scheme, we do not need a sophisticated graphical user interface. Although our CQS aims to provide a user-friendly graphical user interface, we have also produced a suite of commands intended for automatic batch processing of large corpora. This command-line corpus query environment is called the Shogakukan Language Toolbox (SLTB). Many processes, such as showing KWIC lines, sorting the concordance lines by specifying multiple keys from the node, are easily accessible and can be executed along with various options and switches. There is also a group of set commands for creating a superset or subset of the queries. Our toolbox is available in-house via remote access from Linux client machines, but we also provide a web-interface for authorized users from outside the company. Figure 1 shows the web interface of the SLTB.

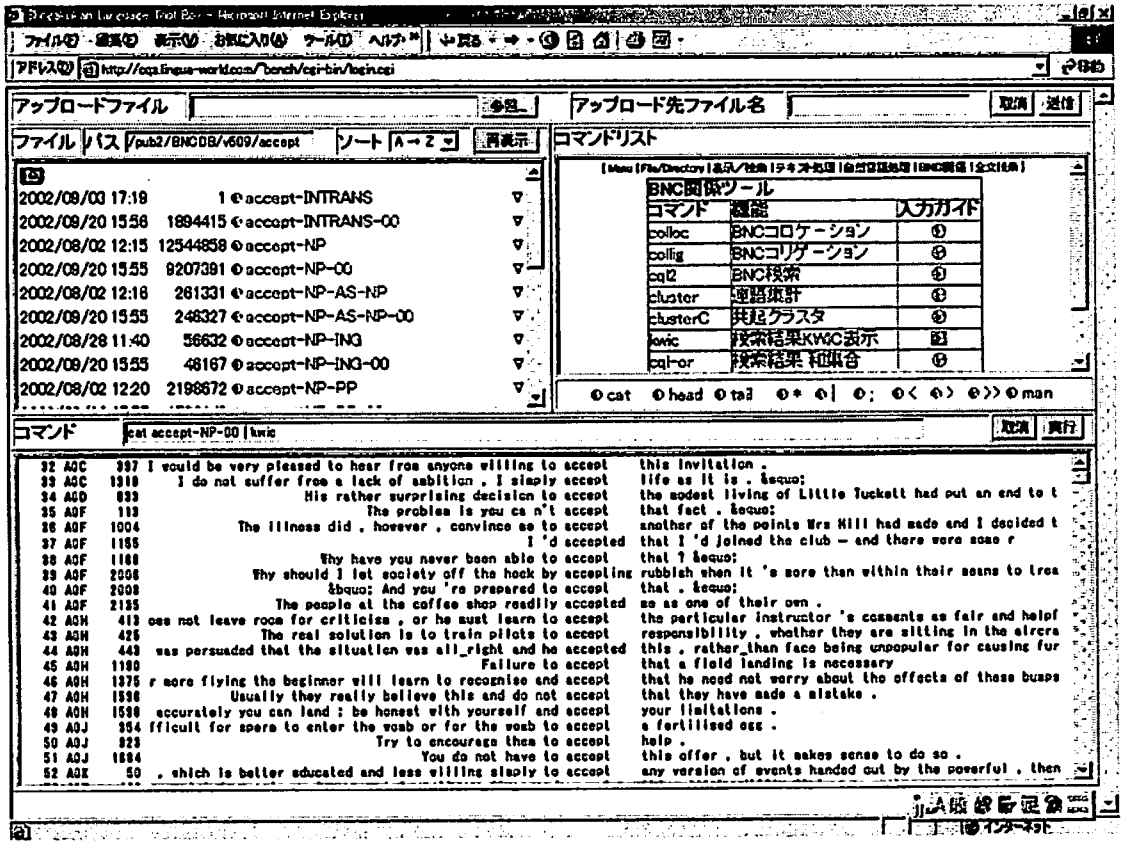

Figure 1: The Shogakukan Language Toolbox web interface

The web interface screen is divided into three main sections: the command-line window in the center, the directory and the command help window in the upper half and the results window in the lower half. The upper left part of the window shows the directory and file information, and is where users can upload files from a local client machine for processing on the server. The upper right part of the window shows the list of commands available for processing the files. This includes ordinary UNIX commands such as grep, sort, uniq, as well as specialized commands for corpus processing. It also provides the command-line interface for tagging the corpus data using free or commercial POS taggers for English or morphological analyzers for Japanese. In this paper, we will show an example of corpus data summarization using the Shogakukan Language Toolbox for the purpose of reducing the workload of lexicographers engaged in corpus analysis.

2 Extraction of verb subcategorization frame (SF) patterns

We set out to investigate how the Shogakukan Language Toolbox can be used to obtain verb subcategorization frequencies from large corpora. For this particular purpose, we decided to use the information con-

tained in the COMLEX Syntax Lexicon. The COMLEX Syntax project aimed to create a moderately broad-coverage lexicon recording the syntactic features of English words for the purpose of computational language analysis (Macleod et al. 1998). They produced a machine-readable lexicon called the COMLEX Syntax Lexicon, which is now available via the Linguistic Data Consortium. One of the unique features of the COMLEX Syntax Lexicon is the detailed subcategorization information contained in it for each verb and adjective. As shown in the example of the verb "accept" in (1), the section called SUBC specifies the list of SF patterns in LISP format. The pattern NP-PP, for instance, denotes that the verb can take a noun phrase followed by a prepositional phrase. Whenever possible, the lexicon shows the value of prepositions in PP (i.e. PVAL) (e.g. "from", "at", "for" in the case below).

```
(1) (VERB    :ORTH "accept"      :SUBC ((NP-PP :PVAL ("from" "at" "for"))
                                        (POSSING)
                                        (NP-ING-OC)
                                        (INTRANS)
                                        (THAT-S)
                                        (WH-S)
                                        (NP-AS-NP)
                                        (NP))
                                 :FEATURES ((VVERYVING :PRESPART T))
```

The current version of the lexicon (Version 3.0) has approximately 200 different SF patterns. About sixty of those patterns were used less than three times throughout the lexicon. In addition, though the number is very small, there are a few patterns which were described in the reference manual but realized by none of the words in the lexicon.

While the COMLEX Syntax Lexicon itself provides a class of "declarative" statements expressing verb SF patterns, the entries are not necessarily in the right format which allows for direct usage by search scripts or programs. We had to convert the statements into queries that our particular CQL (corpus query language) could process. Since the COMLEX Syntax Lexicon has a fine level of granularity in defining SF patterns and contains previously unseen or rare complement structures, mere conversion of the COMLEX subcategorization rules into pattern matching queries using part of speech (POS) tags such as the C5 tagset of CLAWS[2] did not produce good results. Specifically, two queries written for different COMLEX SF patterns often produced the same results because pattern matching queries using POS tag sequences alone could not precisely match the complex patterns described in COMLEX. We examined such cases in advance, and reduced the number of SF query patterns to approximately 110.

Previous studies on the extraction of SF patterns (e.g. Brent 1991, 1993; Manning 1993) mainly focus on a small number of SF patterns (6 types in Brent's study and 19 types in Manning). They claim a precision of between 90 and 96%.

3 Method

The procedure used to extract the frequency of SF patterns for the five hundred most frequent verbs in the BNC consists of two stages:

Stage 1: Extract the patterns from the corpus by using CQL query syntax corresponding to each pattern specified in the COMLEX Syntax Lexicon
Stage 2: Reduce unwanted cases by subtracting a subset of queries from the superset.

There is a good reason for making the process a two-step one. Whilst our CQL is flexible enough to specify the span of context words either preceding or following the key word, the results of the query could contain many unwanted strings due to the fact that CQL cannot accept the negative condition expressed as NOT. Thus it is necessary to subtract the subset obtained by the NOT condition from the overall search results. For example, in the COMLEX Syntax Lexicon, the pattern "the verb *accept* followed by noun phrases" is shown in (2)

[2] The C5 tagset is the tagset used in the distributed version of the BNC.

(2) accept-NP

This pattern can be converted to the following query syntax used for our CQL:

(3) '{L = "accept" P = "VV."} [0,2] {P = "N.*|PN.|DP."}'

This pattern in (3) reads "search for the lemma *accept*, whose POS tag is a lexical verb, followed by either nouns or pronouns after a span of zero to two words following the search word." Allowing the span of zero to two words between the search word *accept* and following nouns or pronouns helped to match determiners or adjectives modifying nouns, but at the same time also matched such unwanted strings as in (4).

(4) a. "accepted by him"
b. "accept for that"

We could have alternatively written queries for every possible pattern for NPs, but this was not feasible because the number of queries covering all possible permutations of NP turned out to be too many. Thus we decided to extract the general pattern first and subtract from it more specific subsets of patterns. The rationale we used was that if the COMLEX Syntax Lexicon covers all the possible SF patterns occurring in the corpus, then the noise or unwanted instances of any particular query result will be matched when some other SF patterns in the Lexicon are searched.

The two types of noise in our SF pattern extraction are shown in Table 1.

Noise type	Subcategorization	Possible noise SF patterns
1	NP	PART-NP, PP, PASSIVE
2	NP	NP-PP

Type 1 covers the elements which might occur in the indefinite context area specified as the span "[0, 2]" in our query. On the other hand, Type 2 indicates that a broad category like NP can entail more specific patterns such as NP-PP. In the case of Type 1, PASSIVE was a particular phenomenon which had to be specified in each query, because the COMLEX Syntax Lexicon does not take voice into account in their description of SF patterns. Some of the patterns could not be precisely specified due to the fact that surface POS sequences cannot unambiguously distinguish optional adverbial elements such as "on Sunday" or "to buy a drink" from required prepositional phrases or infinitival clauses.

We conducted an experiment to evaluate the overall efficiency of our batch program. The following section reports on the results of this experiment and discuss the potential of such corpus processing for supporting lexicographical work.

4 Results

Extracted SF patterns were evaluated in terms of modified *token recall-precision*[3]. This is recall-precision in terms of how many occurrences of verbs in the corpus are assigned correct SF patterns, compared with a gold standard established by manual analysis following the COMLEX Lexicon. What we mean by "modified" is that instead of examining each occurrence of the verb patterns, we assumed that if our method was correct, the sum of the results of each query for different verb SF patterns should be roughly the same as the total number of occurrences of the given verb. In other words, if our method worked well, all the instances of a particular verb should be properly classified into one of the subcategorization patterns specified in the COMLEX Syntax Lexicon and there should not be too much overlap.

We conducted an experiment with 509 verbs and examined how much overlap was observed between the total number of the query matches and the total occurrences of each verb. Figure 2 shows the results of the experiment.

[3] For various evaluation methods, see Matsumoto and Utsuro (2000: 571ff).

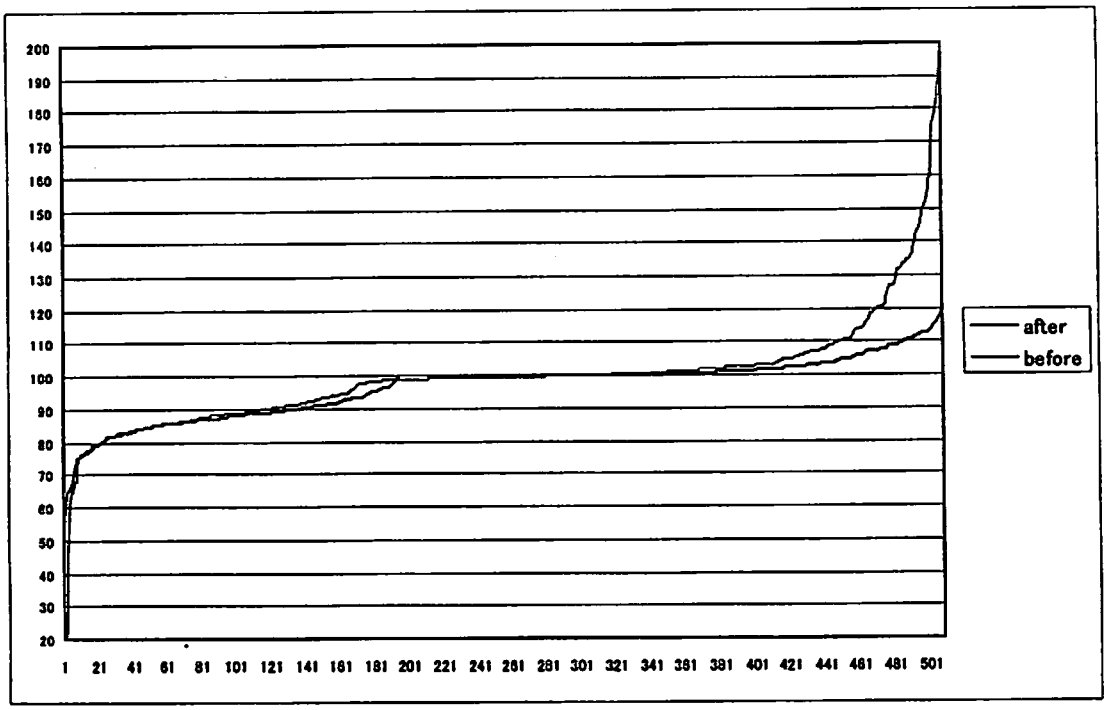

Figure 2: The results of the evaluation experiment

The horizontal axis shows the index number of the 509 verbs examined in this experiment while the vertical axis shows the degree of overlap in percentages. A score of 100 on the vertical axis means that there is a 100% overlap between the number of instances of SF patterns extracted by our queries and the total number of occurrences of that verb. A score of 90 shows that there were 10% fewer examples of SF patterns in comparison with the total number of occurrences. A score of 110, on the other hand, indicates that the queries matched the SF patterns in such a way that the proportion of the sum of instances of SF patterns and the total number of occurrences of the given verb is 110 to 100 (in other words, some of the queries overlapped and matched some corpus instances more than once). There are two curves labeled *before* and *after* which show the different results obtained after the query syntaxes were modified after the first experiment so as to increase the accuracy. This was done mainly by fixing some of the bugs in our operational commands and some errors we made in specifying some of the SF patterns.

If we define ±10% as an allowable range of errors, we can claim that approximately 360 verbs (ranks 120 to 480 in Figure 2) or 70% of the entire set of verbs, were adequately treated by our query algorithms. We also developed a web interface for checking the query results via concordance lines (see Figure 3). In the left frame, you can see the list of 509 verbs. When you click on one of these verbs, a summary of SF patterns will appear on the upper right frame. When you click on a particular SF pattern (in Figure 3, "accept-WH-S-00"), the instances of that pattern will appear in KWIC lines in a separate pop-up window. In this concordance window, you can sort, obtain word/POS n-gram statistics, and download the results. This could also serve as a lexicographer's workbench for revision tasks.

5 Discussion

In this section, we discuss the implications of using such a corpus summarization technique for revising an actual dictionary entry. One obvious advantage of having such frequency information on verb SF patterns is that a lexicographer can then judge how much space should be allocated to the description of those patterns under the entry. For example, Table 1 shows the comparison of the results of our analysis of the SF patterns of the verb *help* with the space assigned to each pattern in the *Shogakukan Progressive English-Japanese Dictionary*.

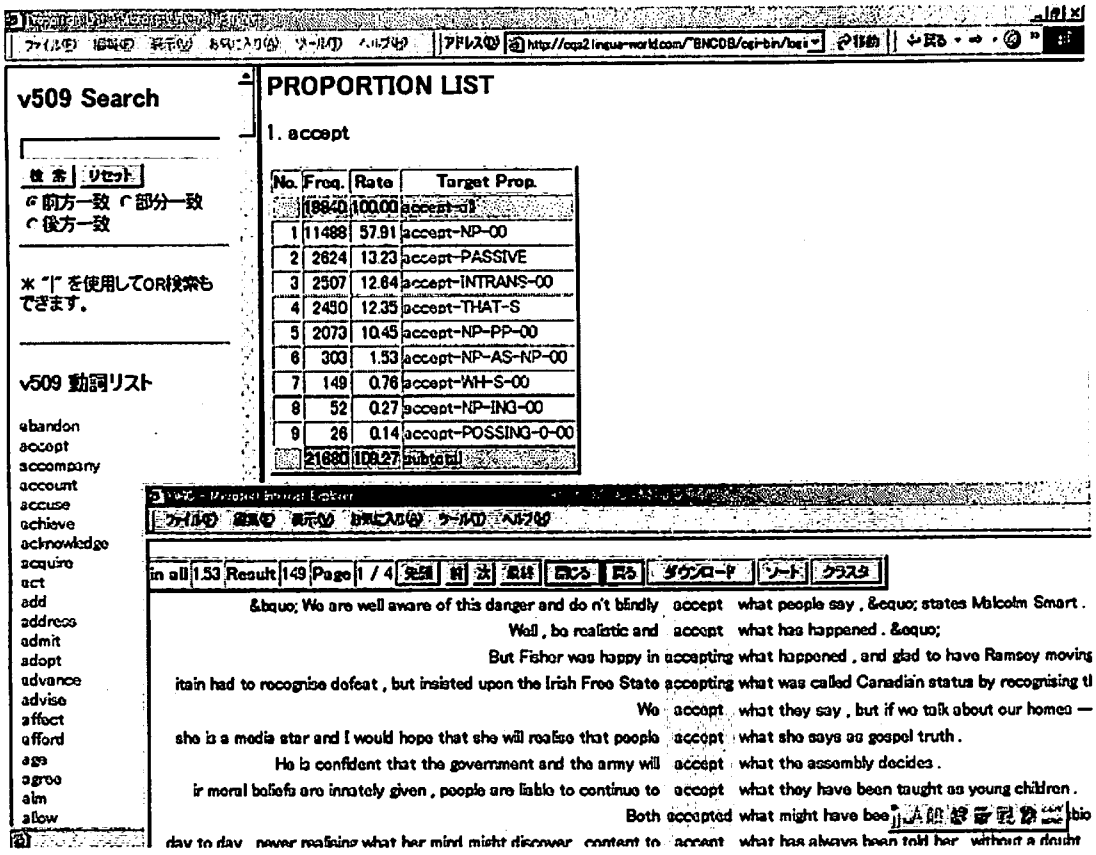

Figure 3: The web-interface of checking query results

No.	Freq.	Rate	Target	sense no.	lines
all	40484	100	help-all		
1	13475	33.29	help-NP	1-(1),2,3,5,6	20
2	5194	12.83	help-TO-INF	1-(1)	1
3	4556	11.26	help-INTRANS		
4	4477	11.06	help-INF	1-(1)	
5	3698	9.14	help-NP-INF	1-(1)	6
6	3300	8.15	help-NP-PP	1-(2)	7
7	3260	8.05	help-NP-TO-INF	1-(1)	
8	2031	5.02	help-PP		
9	1007	2.49	help-PASSIVE		
10	525	1.3	help-ING	4	16
total	41523	102.59	subtotal		

Table 1: Frequency of SF patterns for the verb *help* and the description of each pattern in the *Shogakukan Progressive English-Japanese Dictionary*.

The verb *help* had 40,484 occurrences in the entire BNC. The total of our queries based on the COMLEX Syntax Lexicon matched 41,523 cases, which was slightly (approximately 3%) over the actual number of

occurrences[4]. It is fair to say, however, that results with this level of matching are good enough for them to be made available as databases for lexicographers to consult for the purpose of improving the description of the dictionary entries. Let us consider how the data summarized in Table 1 can be exploited to revise the entry for the verb *help*.

The most frequent SF pattern, "help-NP", for instance, is covered under the senses No. 1, 2, 3, 5 and 6 in the dictionary (see Table 1), which makes up 20 out of 50 lines (40%). This seems to be a fairly good proportion of space to be assigned to this pattern, but on closer inspection, we can see that this is not the case. Under sense 1, different SF patterns (No. 1, 2, 4, 5, 7) are treated and 7 out of 15 lines for sense 1 is allocated to the pattern "help-NP-PP", which is ranked No.6 (8.15%) in our analysis. In contrast, SF patterns such as "help-TO-INF", "help-INF", "help-NP-INF", which are more common than the pattern "help-NP-PP" (altogether they make up 41.08%), are treated less favorably in coverage despite the fact that these infinitival complements make up very high percentages of the overall SF patterns.

Another example of such discrepancy between the frequency of corpus SF patterns and the description of SF patterns in dictionary entries is the treatment of the pattern "help-ING". Whilst it was ranked No. 10 (1.3%) in terms of corpus occurrences, as many as 16 lines (32%) were allocated to the explanation of this pattern. This is mainly due to the fact that the lexicographer used up a lot of space in writing about a popular idiomatic phrase "*cannot help doing*". Relevance to examination questions is certainly one of the possible reasons for such imbalances in treatment. We believe, however, that lexicographers should have more bases for making judgments on how much space to allocate to any particular pattern or sense.

If lexicographers were provided with more objective frequency data on SF patterns, entries for verbs such as *help* would look quite different from what they are now. We have shown that, at the very least, there would be an improvement in the description of the SF patterns in which the verb *help* is followed by infinitival clauses, as well as when it is used as an intransitive. The description of each verb in the dictionary will thus become more systematic when this type of frequency information is available to lexicographers.

6 Conclusion

We hope that we have been able to show that our system is both effective and time-saving, and can prove very useful in helping lexicographers revise a dictionary using corpus data. One of the strengths of our approach is that our lexicon database is not static, but dynamic in the sense that one can query the database via the web interface and that the pattern matching queries can be applied to new sets of data such as the forthcoming American National Corpus or PERC[5] corpus. This type of data mining technique based on corpora will surely become mainstream in computational lexicography in the future. We hope that we can develop a similar "executable lexicon" for other major lexical categories in the near future.

References

Brent, M.R. 1991 Automatic acquisition of subcategorization frames from untagged text. *29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA, pp. 209-14.

Brent, M.R. 1993 From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics* 19: 243-62.

Macleod, C., Grishman, R. and Meyers, A. 1998 *COMLEX Syntax Reference Manual Ver.3.0.* Proteus Project, NYU.

Manning, C.D. 1993 Automatic acquisition of a large subcategorization dictionary from corpora. *31st Annual Meeting of the Association of Computational Linguistics*, Columbus OH, pp. 235-42.

Tono, Y., H. Iwasaki, T. Nakamura, M. Suzuki and E. Egawa, 2001 Shogakukan Corpus Query System in Collaboration with the American National Corpus Project. In Lee, S. (ed.) *ASIALEX 2001 Proceedings: Asian Bilingualism and the Dictionary*. The Second Asialex International Congress, August 8-10, 2001, Yonsei University, Korea, pp. 231-235.

[4] While the COMLEX Lexicon had several other SF patterns for *help* (e.g. help-PART, help-P-POSSING), they were omitted from this analysis due to their negligible frequencies (less than 0.6% each).

[5] PERC Website (http://www.perc21.org/)