Collocation Lists as Instruments for Metaphor Detection in Corpora*

(Listas de Colocações como Instrumentos para Detecção de Metáforas em Corpora)

Tony Berber Sardinha
*(Pontifícia Universidade Católica de São Paulo)***

Abstract: *This paper reports a study on the use of collocation lists as instruments for detecting metaphors in corpora. A collocation list contains the collocations for selected words in corpora together with concordances for those words. As corpora become more available to metaphor researchers, there is a growing need for developing ways to gain access to as much data as the corpus can offer. The research described here has hopefully come some way toward meeting the challenges of developing tools for metaphor corpus research. Results suggest that the collocation lists seem to be a good pre-processing instrument for corpus research of metaphor, despite accuracy problems.*
Key-words: *Corpus Linguistics; metaphor; collocations.*

Resumo: *Este trabalho apresenta uma pesquisa sobre o uso de listas de colocações como instrumentos para detecção de metáforas em corpora. Uma lista de colocação contém as colocações de palavras selecionadas de corpora juntamente com as concordâncias dessas palavras. Na medida que os corpora se tornam mais acessíveis aos pesquisadores de metáfora, começa a surgir uma necessidade de desenvolver maneiras de acessar a maior quantidade possível de dados que um corpus oferece. A pesquisa descrita aqui tentou enfrentar esse desafio, criando e testando ferramentas para pesquisa de metáfora baseada em corpus. Os resultados sugerem que as listas de colocações podem ser um instrumento eficaz de pré-processamento de corpus com vistas à análise humana de metáforas, a despeito de alguns problemas de precisão.*
Palavras-chave: *Lingüística de Corpus; metáfora; colocações.*

---

* The author is grateful to the two reviewers for their many valuable comments.
** Research supported by CNPq (Brasília, Brazil) under grant # 350455/03-1.

1. Introduction

In this paper, I wish to report a study on the feasibility of using lists of collocations as instruments for helping researchers spot metaphorical expressions in corpora. What sparked this investigation was the realization that at the moment we lack instruments in metaphor corpus analysis that will allow us to have as comprehensive a look as possible into a corpus and gather evidence for the presence of linguistic and conceptual metaphors. Arguably traditional tools such as wordlists and concordances do not fulfil this role, as each is restricted in its own way. A wordlist presents words in isolation and so does not allow a researcher to see that word in use, which is a pre-requisite for metaphor analysis. A concordance, on the other hand, requires that researchers have made up their mind about which word(s) to search, and since you can only find what you are looking for, a wrong decision may mean that certain metaphors will not be found.

A collocation list, by contrast, is a list of the collocates of each word in a corpus, as well as concordances for these words. A collocation is a co-occurrence of two or more words within a given span (distance from each other). For instance, 'stay' and 'course' collocate in a two word span forming a collocation such as 'stay the course'. In this case, 'stay' is a collocate of 'course', in the same way that 'course' is a collocate of 'stay'. In addition to collocates such as these, our collocation list also incorporates a concordance, which is a list of a given word in its original context. For instance, a concordance for 'stay' might bring up the following citation: 'President George W Bush has restated his policy that the US will stay the course in Iraq'.

Such a list of collocations is by no means a revolutionary tool, as it simply combines collocation lists and concordances. Nevertheless, commercial concordancing packages such as WordSmith Tools v.3 (Scott 1998) are not programmed to produce concordance indexes of all the words in a corpus or to generate collocate lists (see below) of each search word; rather, the user needs to specify one word or set of words as search words, produce the concordance and only then select the collocate table function for that single concordance. Because of this limitation, our collocation lists were generated by a script specially written for this research project.

The size of the collocation list will depend on the size of the corpus but it can be tailored to the needs of specific researchers, who can specify

constraints such as minimum frequency or a fixed number of collocations to be displayed. These lists are produced by an online application, available at http://lael.pucsp.br/corpora, and can work with corpora of any language and size (but this will depend on other factors such as Internet traffic and server availability). The program will scan the user corpus and detect all the collocates for each word, run concordances for each word and present the data in an organized way. The user can save or print out the results.

The layout of the lists is as follows:

```
==== node word =====
node word ==> collocate_1 collocate_n  (     count of collocates)

KWIC concordance
```

The main working concepts here are linguistic metaphor, conceptual metaphor and collocation. A linguistic metaphor is an expression in which we can find at least one term that has been used metaphorically, meaning that it was used to indicate a domain that is incongruous with the surrounding context (Cameron 2002: 674). Linguistic metaphors are also referred to as metaphoric expression, and they are considered to be grounded in (Gibbs & Franks 2002) or realizations of conceptual metaphors (Lakoff 1993: 203).

A conceptual metaphor in turn represents a cognitive mapping between a source domain and a target domain (Lakoff 1993: 203). For example, the conceptual metaphor 'THEORIES ARE BUILDINGS' represents a mapping between the source domain of buildings and the target domain of theories. As a result of this mapping, theories are conceptualized as buildings in linguistic expressions such as 'there is no foundation at all for this argument', where 'foundation' links to the domain of buildings and 'argument' to the domain of theories.

A collocation is 'the occurrence of two or more words within a short space of each other in a text' (Sinclair 1991:170), normally five words on either side of the node. This is the working definition adopted in this study. But collocation may also be defined in statistical terms, as the co-occurrence of two words at a higher frequency than could be expected by their individual frequencies. This has been adopted in previous lexical studies in Corpus Linguistics (Moon 1998; Stubbs 2002) as well as in corpus-based

metaphor studies (Berber Sardinha 2004; Deignan 1999b: 23), where it is not the norm: for instance, Deignan (1999b: 24) adopted a statistical criterion, whereas Berber Sardinha (2002) and Deignan and Potter (2004) did not.

In the present study, though, I have opted for a non-statistical definition of collocation because what seems to be more important for this particular project is to preserve as much data on the list as possible thus presenting as much evidence as possible to researchers, rather than leaving out data because they do not meet statistical relevance. By this is not meant that statistical measures of association such as Mutual Information or the T and z scores should have no place in metaphor studies, but that for the purpose of this study, which is to explore an instrument to present corpus data in an organized and thorough way, statistical co-occurrence is not a priority. More research needs be done anyway to ascertain that word pairs having acceptable association scores are better predictors of metaphorical use than pairs having lower scores.

In the study of collocations, we normally distinguish the node word from its collocates. The node is that word which a researcher is focusing on; a collocate, in turn, is a word that co-occurs with the node within the span. So, in a sentence fragment like 'lay the foundation for an argument in support of human superiority', if our span were a space of five words around every word, and selected 'argument' as our node word, we would have all the words in this fragment as collocates.

Collocations are a useful means for metaphor identification because frequent collocates often signal an incongruity or tension between a word and its surrounding context (Deignan 2005), which in turn may indicate a metaphor. Going back to the previous example, if we had a list in front of us that showed the word 'foundation' collocating with 'argument', we might take this as initial evidence of metaphorical use. A prerequisite for this is thus that the researcher should be able to spot incongruous words in the collocating pair. Finally, we would need to go back to the actual texts or transcribed talk in our corpus to verify this in context, normally with the help of a concordancer. As said above, our collocation lists include concordances as a feature.

2. Previous literature on collocation and metaphor

There have been previous studies looking at lexis and collocation as metaphor signalling devices. White (2003) noticed that a range of adjectives collocated with metaphorical 'growth' in a corpus of Financial Times texts, indicating conventional metaphors more often than novel ones. Stubbs (Stubbs 1995: 384-385) observed that 'big' and 'small' occurred in metaphorical language, such as 'big business', 'big gun', 'small fry', and 'small beer'. Deignan and Potter (2004: 1237) noted that 'heart', in the metaphorical sense of 'seat of emotions', collocated with several recurring words, such as 'break', 'open', and 'win'. This led them to propose the more general principle that recurrent collocation may be a regular feature of metaphorical expressions: 'It may be that there is a tendency for the non-literal uses of frequent words to have regular collocates and syntactic patterns, as this helps to disambiguate them from literal uses.' (Deignan & Potter 2004: 1238).

The research presented here is a follow-up to my previous work on the relationship between collocations and metaphor use in Applied Linguistics dissertations. In Berber Sardinha (2002), I pursued the notion that word pairs with similar collocates (e.g. 'theory' and 'building', both having collocates such as 'foundation', 'buttress' and 'shaky') would be indicative of metaphors. In order to single out those word pairs, I applied three sampling criteria to the data: frequency markedness, number of shared collocates, and semantic distance. Firstly, frequency markedness provided a means of selecting genre-specific vocabulary, which may be desirable in metaphor studies (Charteris-Black 2004; Partington 1998); it also functioned as a sample reduction mechanism, cutting down the size of the initial sample to a more manageable 500 words. This was needed in view of the processing algorithm put in place there, which was designed to output word pairs, thus requiring that collocations for each word be cross-tabulated with those for every other word in the sample. A large word sample would be unwieldy for this, since it would involve too many pairwise comparisons. Secondly, number of shared collocates was a measure of strength between two words' collocations, suggesting that words above a certain minimum number of collocates (three) were more central to the discourse (Hoey 1991) and as a result should receive preferential treatment in the manual analysis phase of the investigation. Finally, semantic distance (Pedersen & Patwardhan 2002) enabled me to determine automatically

those word pairs in the sample that were distant enough in meaning to suggest domain incongruity (such as 'love' and 'journey'). This should stop synonyms ('research' and 'investigation') and other closely related words ('dissertation' and 'text') from further processing. This semantic distance calculation algorithm worked for English only, as it consulted the English WordNet database; as the data for the study were in Portuguese, word pairs had to be translated to that language before processing. A problem with this whole procedure was that it was entirely based on contrasting pairs of words with similar collocates, in the hope that these word pairs would include incongruous source and target elements. While a few of the resulting word pairs did have source and target lexis as predicted, the majority of them did not, including for example meronymic and hypernymic relationships, some of which indicated metonyms.

3. Procedures for corpus analysis

In this section, I describe in detail the procedures followed in this investigation, as well as the operational definitions of the main concepts involved in the study:

a) A metaphor Vehicle is the focus of a linguistic metaphor (Cameron 2003: 11). It is a word or group of words 'that can have an interpretation that is incongruous with the discourse context, or with the meaning created by the co-text' (Cameron, 2003, p.9), by linking to a different semantic field or conceptual domain from that elicited by the non-vehicle terms. Identifying the Vehicle terms is regarded as a necessary condition for the identification of a linguistic metaphor (Cameron 2003: 60).

b) A linguistic metaphor is an expression that has at least one Vehicle term and that can be interpreted as a mapping between two different domains.

c) A domain is a 'a sphere of knowledge, influence, or activity' (Merriam-Webster Online Dictionary). It is normally used in the metaphor literature 'to refer to the ideas or semantic field referred to by a lexical item'. (Cameron 2003: 9)

d) Here, collocation means simply the co-occurrence of two words, namely a node word and a collocate, at a certain distance from each other.

A node word, as described above, is that word which one is focusing on. In our case, the node words are the possible Vehicles. A collocate is a lexical word (noun, verb, adjective, adverb, numeral) occurring within a span of five words to the left and right of a node word. The collocates of all node words were sorted in terms of their total number; all nodes that had two collocates were classified as 'level 2 collocate words', and so on, up to 256, which was the highest number of collocates obtained. This sorting was referred to as 'collocate levels'. Hence, when we speak of collocate level 2, we mean node words having two collocates, collocate level 3, three collocates, and so on.

The data analysis procedures were the following:

(1) Two corpora were collected consisting of texts from the domain of investment banking. For ease of reference, the corpora will be referred to as A and B. Corpus A contains investment bank conference calls. These conference calls, given in Portuguese, are an event where a group of executives sit in the bank's headquarters and speak remotely with investors, the press, government officials and the general public. The audience are mainly in Brazil, and they interact with the executives by telephone, by placing questions and comments. Corpus B is made up of press releases by a Brazilian bank, also in Portuguese. The texts are mainly written commentary and interviews with some of the bank's top executives. The material for each corpus was collected from the banks' websites. The material refers to the years between 2000 and 2005. In both cases, the preference for banking genres was because this study is part of a larger project that is aimed at describing business language[1] . The dimensions of the corpora are given in the table below.

Table 1: Size of corpora

Corpus	Tokens	Types	Texts
A	85,438	5,194	14
B	7,058	1,952	10

[1] The project is named DIRECT. For more information, visit the project's website at http://lael.pucsp.br/direct

Each corpus was manually annotated for metaphors, according to the operational definitions given above.

(2) Each corpus was processed by a script written in Unix Shell by the author. The script was run in Mac OS 10.3. The basic algorithm for the script is the following:

– For each corpus:

- Produce a word frequency list.
- Count tokens and types and output these counts to a list (for general corpus size statistics).

– For each word in each corpus:

- Detect if it has a tag for metaphor, placed during the manual annotation of the corpora. If it does, save it to a 'list of actual Vehicles', and carry on with the rest of the processing. If it does not, do not save it to that list and carry on with the processing.
- Find its collocates in a span of 5 words to its left and right.
- Apply the minimum allowed frequency criterion set for each corpus (frequency 2 for corpus A and frequency 1 for corpus B).
- Delete collocates that are numbers, loose letters, symbols, and words of the following word classes: articles, pronouns, prepositions and conjunctions.
- For each collocate level:
  - Display each word and its collocates, count them, produce a KWIC concordance of each word, and save this to a 'collocation list'.
  - Save the words to a list of 'possible Vehicles'.
  - Compare each word on the list of possible Vehicles to each word on a list of actual Vehicles from each corpus, and:
    - If a possible Vehicle matches an actual Vehicle, save it to a list of 'matched Vehicles';
    - If a possible Vehicle does not match an actual Vehicle, save it to a list of 'unmatched Vehicles'.

– Move to the next word, until the last word type in the corpus, ending the script.

(3) From the processing described above, four further analytical categories emerged, namely 'actual Vehicle', 'possible Vehicle' and 'matched Vehicle':

a) An actual Vehicle is a word that was marked up as a Vehicle during manual analysis.

b) A possible Vehicle is a word that had valid collocates.

c) A matched Vehicle is a possible Vehicle that matched an actual Vehicle.

d) An unmatched Vehicle is a possible Vehicle that did not match any actual Vehicle.

(4) From the categories above, an evaluation was carried out to measure the extent to which the collocation lists included the metaphors in the corpora. This was done by another Unix Shell script written by the author that calculated two statistics, namely recall and precision. In this study:

a) Recall refers to the share of vehicles found in the corpora. It is a measure of comprehensiveness. It is arrived at by dividing the total number of possible Vehicles by the total number of actual Vehicles. A recall of 100% means that all of the Vehicles were matched.

b) Precision has to do with the proportion of possible Vehicles that were matched Vehicles. It is a measure of accuracy. It is found by dividing the total number of matched Vehicles by the total number of possible Vehicles. A precision of 100% means that all of the possible Vehicles were actual Vehicles.

(5) Sample size was computed as the proportion of the total types in the corpora included in each collocate level list. For instance, if there were 1000 types in the corpus and 200 possible Vehicles on collocate list level 5, the sample size would be 20% (200 / 1000 * 100).

4. Performance results

Table 2 below shows the totals of actual Vehicles in corpora A and B:

Table 2: Actual Vehicles in each corpus

Corpus	Actual Vehicles
A	423
B	352

Tables Table 3 and Table 4 below give the totals of possible, matched and unmatched Vehicles in corpora A and B, as well as recall, precision and sample size, for each collocate level:

Table 3: Results for corpus A

Collocates >	Possible	Matched	Unmatched	Recall %	Precision %	Sample %
2	2887	266	157	62.88	9.21	55.56
3	2885	266	157	62.88	9.22	55.52
4	2871	265	158	62.65	9.23	55.25
5	2842	263	160	62.17	9.25	54.70
6	2776	261	162	61.70	9.40	53.43
7	2629	254	169	60.05	9.66	50.60
8	2472	238	185	56.26	9.63	47.58
9	2279	229	194	54.14	10.05	43.86
10	2098	206	217	48.70	9.82	40.38
11	1945	202	221	47.75	10.39	37.43
12	1827	193	230	45.63	10.56	35.16
13	1718	184	239	43.50	10.71	33.06
14	1593	174	249	41.13	10.92	30.66
15	1500	166	257	39.24	11.07	28.87
16	1424	157	266	37.12	11.03	27.41
17	1354	152	271	35.93	11.23	26.06
18	1286	147	276	34.75	11.43	24.75
19	1242	143	280	33.81	11.51	23.90
20	1181	137	286	32.39	11.60	22.73
21	1135	134	289	31.68	11.81	21.84
22	1086	130	293	30.73	11.97	20.90
23	1039	123	300	29.08	11.84	20.00
24	1004	121	302	28.61	12.05	19.32
25	964	117	306	27.66	12.14	18.55

Table 4: Results for corpus B

Collocates >	Possible	Matched	Unmatched	Recall %	Precision %	Sample %
2	1788	348	4	98.86	19.46	91.60
3	1753	342	10	97.16	19.51	89.81
4	1527	295	57	83.81	19.32	78.23
5	1084	208	144	59.09	19.19	55.53
6	746	149	203	42.33	19.97	38.22
7	608	128	224	36.36	21.05	31.15
8	537	116	236	32.95	21.60	27.51
9	471	97	255	27.56	20.59	24.13
10	398	77	275	21.88	19.35	20.39
11	359	71	281	20.17	19.78	18.39
12	322	66	286	18.75	20.50	16.50
13	286	54	298	15.34	18.88	14.65
14	253	49	303	13.92	19.37	12.96
15	223	44	308	12.50	19.73	11.42
16	201	38	314	10.80	18.91	10.30
17	184	34	318	9.66	18.48	9.43
18	172	32	320	9.09	18.60	8.81
19	158	30	322	8.52	18.99	8.09
20	145	26	326	7.39	17.93	7.43
21	132	22	330	6.25	16.67	6.76
22	121	21	331	5.97	17.36	6.20
23	112	16	336	4.55	14.29	5.74
24	103	12	340	3.41	11.65	5.28
25	97	12	340	3.41	12.37	4.97

Recall and precision are different across the two corpora. The degree of success in predicting Vehicles therefore seems to be dependent on the corpus. Results were better for corpus B, which is smaller. Perhaps this method produces better results with smaller corpora than with larger ones, comparatively. But this can only be demonstrated by applying the procedure to other corpora of different sizes.

Recall decreases continuously as the collocate levels rise. This is because at lower collocate levels there are more possible Vehicles, and so the odds of matching Vehicles are higher. Precision, on the other hand, tends to increase as collocate levels rise, but this increase is not steady across the two corpora. In corpus B, precision reaches its peak at level 9, then fluctuates and finally drops continuously.

From these results, we can work out the average and the best figures for each corpus, as shown Table 5 below:

Table 5: Average and best performance figures

	Corpus A	Corpus B
Best recall	62.88%	98.86%
Best precision	12.14%	20.59%
Average recall	44.60%	27.07%
Average precision	10.66%	18.48%

Across the two corpora, the average figures are 35.84% recall and 14.57% precision.

These numbers suggest that there is a clear contrast in performance between recall and precision. Recall is on average 5 to 6 times better than precision. This means that the procedure signals a lot of words, among which one can find a good share (from roughly a third to virtually all) of the Vehicles in the corpora. But there is equally a large number of wrong guesses in these (up to 90% of the words suggested as possible Vehicles may be wrong).

As far as sample sizes, they closely match recall. The correlations between recall and sample size for each corpus are very high, at $r^2=.99$ in both cases. Put another way, the number of metaphors retrieved by each list is proportional to the size of the list. Longer lists recall more Vehicles than shorter ones. Recall is on average 9% higher than sample size in corpus A and 2% in corpus B. These figures indicate that it should be easy for a researcher to predict the potential number of true Vehicles in a sample for a particular collocate level. The number of metaphors is about 2% to 9% higher than the size of the sample. Say that a researcher is faced with a collocate list containing 1,000 possible Vehicles out of a corpus with 10,000 word types. Since this represents 10% of the total types in the corpus, he/she might predict that the list carries about 10% of the total Vehicles in the corpus. Using the average precision figure of about 15%,

he/she might expect to find about 150 (15% of 1,000) actual Vehicles on that list. These predictions should be treated as mere speculation, though, since we can never know for certain how many metaphors there are in a corpus until we look at the data in detail.

On the face of them, the performance results may be regarded as disappointing, after all in general they are far from perfect, especially with respect to precision. This is not a fair assessment, though, since it is not reasonable to expect any computational linguistic procedure to be 100% right. In order to level up the playing field, we would need to compare these performance results with baseline measures of metaphor identification, which would in turn enable us to assess these results not in terms of what is ideal, but in terms of what is empirically possible. This is what is presented next.

5. Assessment

In order for us to have a better picture of how good the figures presented above are, a comparison was made between them and two other measures of metaphor identification: a random measure and a human identification measure.

(a) The random measure was obtained by having the computer pull out blindly a certain amount of words from corpus A (423, that is, the total of actual Vehicles in that corpus) and then match those words with the list of actual Vehicles for that corpus. After that, recall and precision scores were calculated. The results were 8.8% for both recall and precision.

(b) The human identification measure was arrived at by having a group of individuals point out possible Vehicles by looking at the word list for corpus A. The participants were MA students who had had a course on metaphor and Corpus Linguistics in Brazil. The list of each individual's possible Vehicles was matched to the list of actual Vehicles in the corpus and then recall and precision scores were calculated. The individual scores were then averaged out and overall recall and precision scores were obtained. Recall was 26.69%, while precision was 16.55%.

Chart 1 below summarizes these results.

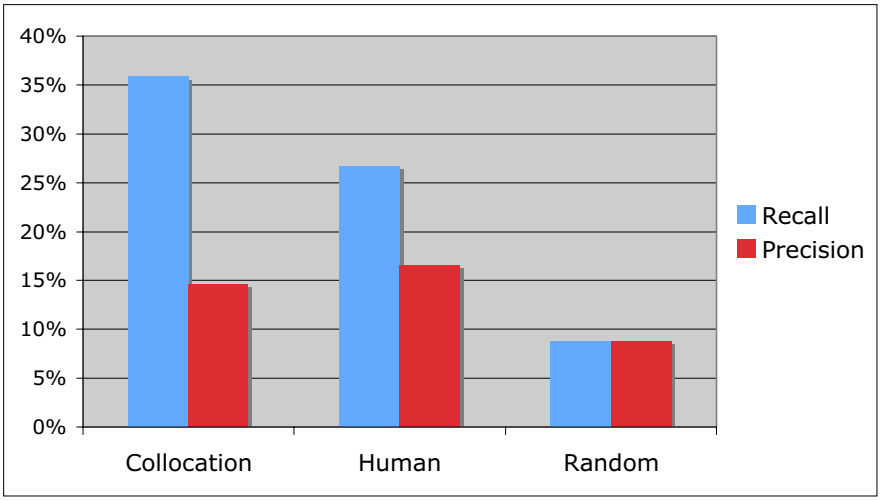

Chart 1: Comparison of collocation list with other methods of metaphor extraction

As the chart shows, collocation lists are best at recall, but are weaker at precision than humans.

The question that now arises is how to improve precision for the procedure presented here. The answer lies in going back to the original purpose of this procedure: it was intended as an instrument that presents researchers with a list of possible Vehicles, from which they should select the best candidates. Consequently, as said before, we should use the list of collocations as a starting point for analysis.

Bearing that in mind, the following mini-study was conducted. A list of possible Vehicles was selected and presented to a group of four individuals. They were asked to pick words from that list that they though might be Vehicles in the corpus (for ease of reference, each word that was picked was termed a 'suggested Vehicle'). These were the same instructions given to participants in the human identification task. The aim was to see whether people were better able to guess possible metaphor Vehicles with a collocation list than they were with a vocabulary list of the whole corpus. The pre-processed list was that for level 22 (corpus A) with 1086 words. It was picked at random out of the 24 lists. As far as identifying with the

plain vocabulary list, results from a previous study were used, instead of running another experiment; these were the recall and precision figures from the human identification task mentioned above. As such, they were obtained with a different group of individuals, but on the same corpus.

One of the participants was familiar with the corpus and its texts. He also had more experience in metaphor analysis by computer and with hand-coding corpora for metaphors. This was named the 'Expert' analyst. The other three participants had a similar background to those who took part in the previous human identification task mentioned above, that is, they were MA students doing Corpus Linguistics and/or Metaphor Studies projects, had not seen any of the texts in the corpus, and had only limited experience in hand-coding metaphors in corpora. These were termed the 'Beginning' analysts. The participants were divided in these groups because it was anticipated that there might be a difference in performance due to familiarity with the texts and with hand-tagging metaphor in corpora.

Recall and precision scores were calculated. Recall was obtained by dividing the total of suggested Vehicles by the total of actual Vehicles in corpus A (423) (times 100). Precision was obtained by dividing the total of matched Vehicles by the total of suggested Vehicles (times 100).

Table 6: Recall and Precision for human identification based on possible Vehicle list

Analyst	Suggested Vehicles	Matched Vehicles	Recall %	Precision %
Beginning 1	650	95	22.46	14.62
Beginning 2	892	118	27.90	13.23
Beginning 3	804	110	26.00	13.68
Expert	226	105	24.82	46.46

The figures in Table 6 suggest that, as far as recall, both kinds of analysts performed similarly, but insofar as precision, there is marked difference between them: the expert analyst achieved a precision score that was over three points higher than that of the beginners. Note that recall dropped a little from 30.73% to an average 25.3%. This happened because

the number of words chosen by the analysts (782) was much smaller than that on the original possible Vehicle list (1086). With the word sample being smaller, the odds of having any Vehicles in them decreases, as happened here.

In order to answer the initial question of whether looking at a collocation list increases the chances of a person accurately finding the actual Vehicles in a corpus, we must plot the precision figures with the previous human identification precision score, as the chart below shows.

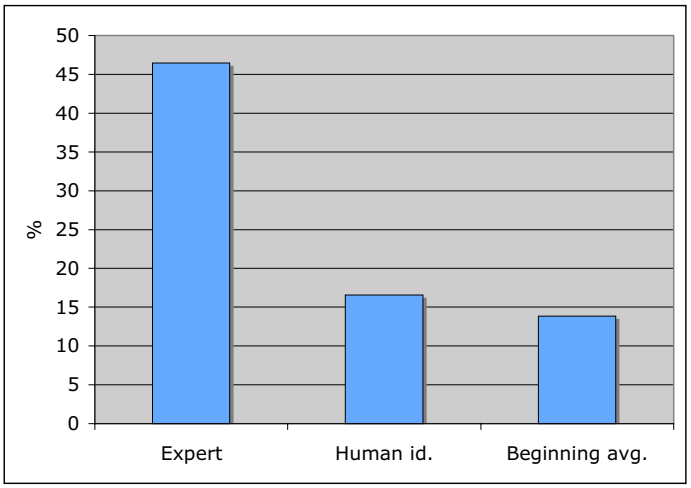

Chart 2: Precision scores for expert and beginning analysts working on possible Vehicle list and human analysts working on vocabulary list

The chart indicates that our collocation list will work as a good tool for spotting metaphors possible Vehicles if the analyst is experienced in metaphor coding and if he/she has good knowledge of the corpus that he/she is investigating. If, on the other hand, the analyst comes to the list before him/her with little or no previous knowledge of the corpus and with little experience in this sort of task, then the list does not seem to work to his/her benefit.

We can now move on to the more general question of whether a human-processed possible Vehicles list is a more precise instrument than the original computer-produced possible Vehicle list. In order to answer that, we must plot the expert and beginner precision rates against that for the procedure

alone (at level 5 for corpus A, as this was the collocation level list with which the coders worked). The table below displays the results of this comparison.

Table 7: Improvement in performance

	Precision %	Growth %	Times
Expert	46.46	402.1	5.0
Beginning avg.	13.84	49.6	1.5
Procedure	9.25		

According to the results, precision was improved by the analysts to a great extent. The expert analyst was able to increase precision by 402%, a five-fold growth. The less experienced researchers did much less well, but still they improved performance by nearly 50%, or one and half times. This means that the analysts were able to locate promising Vehicles while at the same time ignoring false ones. Therefore, the answer to the question posed above is that manually post-processing a possible Vehicle list is a good strategy. People are able to throw away bad candidates and therefore return a more accurate list in terms of the Vehicles that it predicts.

6. Summary of findings

The main findings are:

(a) The lists work differently for different corpora, as results varied between the two corpora used in this study. Because of this, the remaining summary points below refer to the average between the two corpora.

(b) The collocation lists recall a fairly good number of actual Vehicles: about 1 in every 3.

(c) In doing so, their precision is low, though: only about 1 in every 7 possible Vehicles turns out to be an actual Vehicle.

(d) They recall more actual Vehicles than people do (people recall about 1 in 4 Vehicles).

(e) They do not predict Vehicles as accurately as people do (people accurately predict about 1 in every 6 Vehicles).

(f) They seem to work as a good starting point for experienced researchers. One such researcher became more accurate at finding Vehicles with them (about 1 in every 2 of his choices were actual Vehicles).

(g) They do not seem to make a difference for non-experienced researchers. These researchers were just as accurate working with or without the lists (about 1 in 10 words picked by the non-experienced researchers were actual Vehicles).

(h) Their accuracy can be improved greatly if people examine them and delete words that do not seem good Vehicle candidates. As a result, their precision rose up to 5 times.

(i) List size is proportional to recall: longer lists recall more Vehicles than shorter ones.

These results suggest that the collocation lists seem to be a good pre-processing instrument for corpus research of metaphor. Precision, which is their weak point, can be improved up to 5 times by carefully choosing words from the list.

7. Using collocation lists

In this section, I want to illustrate how we can use the collocation lists to make the leap from linguistic metaphor to conceptual metaphor. This is a personal account of the process, and it involves interpreting the data on the lists. As will be seen, this is not automatic and it presupposes making judgments about the lexical or conceptual fields represented in the data. Other researchers might draw different conclusions from the data or work through them in different ways[2] . The guiding principle, though, should remain the same for everyone: we should base our interpretation on the evidence presented by the lists.

[2] For more on the issue of variability in interpreting textual data in metaphor studies, see (Semino, Heywood, & Short, 2004).

The first example is of 'aberta' (open, [feminine]). Its entry on the collocation list is as follows:

```
==== aberta =====
aberta ==> aí, companhia, continua, empresa, específica, está, mercado,
nível, provisão, questão, receita, serviços, será, são, tarde (15).

[snippet of concordance:]
or de relações de mercado de uma empresa aberta eu tenho muitas limitações na
s apenas como ouvintes e mais tarde será aberta a sessão de perguntas e respos
 tela 11 nós temos a receita de serviços aberta no trimestre e no ano e enfim
quidados enfim é uma questão que está ai aberta mas acreditamos que juridicame
 tem direito a voto não são de companhia aberta e estão numa subsidiária no ex
```

For interpretation purposes, the main elements are the following:

(a) 'aberta': The potential Vehicle;

(b) 'aí … tarde': these are the collocates, which may include words expressing Topics, Vehicles, or none of them. A close translation of the collocates is a follows: 'there, company, continue/s, company, specific, is, market, level, provision, question, income, services, will be, are, late/afternoon.'

In the examples below, concordances have been shortened for reasons of space.

I begin with the potential Vehicle, 'aberta', which if used metaphorically may indicate a 'container' source domain. In order to find out if there is any indication of metaphorical use, I turn to list of collocates, where I notice nouns referring to company, question and income. These signal to me that this may indeed be a metaphorical use of 'aberta'. Next I turned to the concordance and confirmed this impression.

From the concordance, I worked my way toward conceptual labels for the metaphors, and laid out the lexical items in a table, as such:

Collocation (translation)	Source domain	Target domain
Empresa/companhia (company)	Container	Company's capital
Sessão de perguntas (question time)	Container	Question time
Questão (question)	Container	Question
Receita de serviços (services earnings)	Container	Budget

Based on the table, the following conceptual metaphors seem to present in the data:

- A Company is a container
- An event is a container
- A question is a container
- A Budget is a container

In this illustration, the collocates list was used to generate initial predictions, while the concordance helped me to confirm or reject these hypotheses. Although there is no predetermined sequence for using this list, it is arguably better to start with the collocates list, so that we can have a glimpse of the contents of concordance. In this particular case, the concordance was a key instrument in locating the actual linguistic metaphors (or parts of it) and in deciding what they meant in discourse. However, they did not help determine conceptual metaphors. As usual, it was up to the researcher to make the leap from linguistic data to conceptual metaphors, an extrapolation that is largely based on 'informed intuition' (Deignan 1999a: 180; Koller 2004a: 9) and subject to variability (Semino et al. 2004).

8. Discussion

It has been argued here that a sampling strategy is needed in corpus analysis of metaphor because, on the one hand, there is a large number of words in any corpus, and on the other, any word may be used metaphorically; as a result, any corpus may contain a large number of metaphors, with lots of occurrences each. Collocation lists were proposed as an instrument for viewing words and their co-texts, thus enabling researchers to sample words that seem to be used metaphorically in light of linguistic evidence.

Traditionally, researchers have resorted to a number of sampling strategies, such as using intuition, drawing on previous literature or choosing metaphor keywords based on reading selected parts of the corpus (Charteris-Black 2004; Deignan & Potter 2004; Koller 2004b). Few have gone through a whole corpus, identifying each metaphor in the data (e.g. Cameron 2003: 58; Gibbs & Franks 2002).

Without a principled means of selecting words for closer analysis, we face the risk of under-representing the metaphors in our corpus. We may not find important metaphors or we may stick to familiar or celebrated ones. Deignan and Potter (2004: 1234) refer to both of these problems:

> 'A difficulty for any corpus investigation is the wealth of data that could be examined, and a decision about which lexemes to focus on has to be made. There is a risk that patterns of potential interest may be missed, a risk which is difficult to avoid completely, because without a vast team of researchers and unlimited time, it is not possible to examine every linguistic realization of a whole semantic field. In trying to minimize this risk when we selected the lexemes to focus on in this study, we drew on an earlier overview of the source domain that had been undertaken using a corpus approach.'

We may have a better grasp of the potential problem of under-representation in current corpus analysis of metaphor by comparing data from studies that have been based on a partial selection of metaphor keywords for corpus analysis and those that have carried out a thorough identification of metaphors. If the former reveal a lower metaphor density rate than the latter, then that might be attributed to techniques that retrieve fewer metaphors than they could. Metaphor density (Cameron 2003) is simply the number of metaphors occurring every n words in a corpus. Cameron (2003) manually coded a 26,613 word corpus and found 711 occurrences of metaphors, thus giving a density of 27 metaphors per 1000 words. Likewise, in this study, our corpora have been fully coded by hand; the two corpora add to 92,507 words, with 667 instances of metaphor, which yields a density of 242 metaphors per 1000 words.

By contrast, studies that have used sampling procedures for corpus analysis include Koller (2004b) and Charteris-Black (2004). Koller (2004b) focused on a corpus of business texts from British and American press of 323,894 words. She predefined a set of 400 metaphorical keywords that she then searched for in the corpus (p.48), retrieving 1604 occurrences of metaphor. This yields a density of 5 metaphors per thousand words. Charteris-Black (2004) looked at a number of text types, and selected metaphor keywords (p.89) by extensively reading the texts in the corpus (p.117). In a 17,618 word corpus of New Labour Manifestos, he found 76 metaphors, which amounts to a thousand-word density of 4. In a much larger corpus of British newspaper sports reports (152.2 million words),

the density was 2, and in a 16-million word corpus of financial reports from a British magazine, the density was 3.

Although there is no way of knowing what the real metaphor density of any of these corpora were, the lower densities suggest that maybe a good deal of metaphor instances were missing from the analysis, and that this may have been a result of the sampling methods employed, which included pre-defining a pool of metaphor keywords beforehand based on 'previous knowledge … and anectodal evidence from individual media texts' (Koller, 2004b, p.48), and searching the corpus 'using some of the metaphor vehicles identified in the sample' (Koller 2004b: 178).

The instrument presented here certainly cannot solve the problem of under-representation, but at least it can present researchers with an organized and thorough snapshot of their data in a way that other instruments cannot, such as concordances (because they require a definition of which words to search) and wordlists (because they present words in isolation).

9. Conclusion

Metaphor analysis depends on human interpretation. If the data are of a manageable size, then one can analyze the whole data set by reading all of the texts. Under these circumstances, there would be no need for sampling a corpus for an initial list of Vehicles.

However, when data are much bigger than what we can process by reading each text in full, then the issue of sampling comes up. The problem with corpus linguistic investigations of metaphor is that the amount of data a corpus can hold is so large that it typically requires a sampling method of some form. Different sampling methods have been used in the literature, such as starting with examples of metaphor, coding a small corpus in full, reading a section of a large corpus, and working with lexicographers who are describing large numbers of word forms (Deignan 2005: 93).

Therefore, finding one's way into an electronic corpus is crucial, as Koller's (2004a: 8) testimony reveals, with respect to her analysis of a 135 thousand word corpus:

> 'As I thus lacked a distinct set of search words to be run through a concordance program, the search in this case was a non-electronic one.'

Current sampling methods such as reading parts of the corpus for search terms or beginning with a set of examples drawn from intuition of from the literature may under-represent[3] the data. This is because such methods may be leaving behind potentially interesting metaphors, as Cameron and Deignan (2003: 151) report:

> 'If he or she has not identified a particular form as worthy of study, it may not emerge from the data during the analysis, and an important metaphorical use may be missed. This reflects an important difficulty in researching linguistic metaphors through a corpus: We are trying to trace patterns of meaning but can only begin our analysis by looking at forms.'

Researchers seem to have recognized this as a problem. The solution (or part of it) may come from corpus linguistics itself, for example by relying on a list of collocations for the whole corpus. A researcher working through such a list can have a general picture of the patterns in the corpus and can then make inferences about the possible Vehicles in the data.

In view of the results presented here, the following questions must be raised: What real benefit do these lists bring to researchers? Are they more beneficial to some researchers than to others?

The answers lie in who is using the lists. If a researcher has similar characteristics to our expert analyst, then the answer is more likely to be that the lists seem to be good tools. By similar characteristics is meant having a fairly good idea of the contents and genre of his/her corpus, having experience in coding linguistic metaphors, and in general being

---

[3] By under-representation is meant to lack members of a population. In our case, the population is all the word types in a corpus. Our goal should be to have a sampling frame, i.e. 'a list of all members of a population' (Minke & Haynes, 2002: 72), that is representative, from which we could draw as many metaphors as possible or desired. Underrepresentation is a useful concept in other areas where good sampling techniques are critical, such as welfare policy: 'in studies that include only administrative data collection, persons or families for whom no information is included in the administrative list [...] have no chance of being included in the sample, and thus will not be represented in the results of the study. [...] Similarly, surveys that collect data from sampled persons also are subject to underrepresentation due to sampling from incomplete or outdated lists.' (Mohadjer & Hussain Choudhry, 2001: 132).

willing to go through the list making careful judgment about which words to pick from the list as possible Vehicles. A post-processed collocation list can return up to a quarter of the Vehicles in a corpus, with an accuracy bordering 50%. This means that a good share of the metaphors might be accounted for, in less time than would be necessary to investigate every word in the corpus. If, on the other hand, researchers do not fit this profile, then these lists do not seem to help. This perhaps indicates that corpus linguistic methods take some getting used to, and that people tend to get better at selecting viable candidate metaphors from a collocation list as they become more experienced.

Despite the fact that underrepresentation in a sample may be a problem, as discussed above, it must be stressed that this is not an 'either-or' situation: machine sampling should be combined with human sampling. A sample offered by the machine may be added to a hand prepared one, with redundancies being eliminated. In the end, it is always up to the researcher to choose the sampling method that best fits his/her objectives, skills, corpus size, and theoretical background, and therefore using reading and intuition to select words may be perfectly acceptable under some circumstances, for example, when the corpus is small enough to be fully read, or when a researcher is not interested in locating all or even the majority of the metaphors in the corpus.

We need to continue to explore new ways to gain access to as much data as our corpus can offer, develop new instruments to meet the demand of the use of larger corpora in metaphor research and test these tools with actual researchers with different skills and purposes. These lists may be useful to students as well. In a teaching environment, teachers may prepare collocation lists and take them to the classroom to help students who are learning about metaphor to come to grips with corpus data. I hope the research described here has come some way toward meeting the challenges of developing tools for metaphor corpus research.

Recebido em dezembro de 2005
Aprovado em abril de 2006
E-mail: tony4@uol.com.br

References

Berber Sardinha, T. 2002. Metaphor in early applied linguistics writing: A corpus-based analysis of lexis in dissertations. *I Conference on Metaphor in Language and Thought*. Catholic University of São Paulo, Brazil.

_____. 2004. Metáfora e Corpora: Teleconferências. *III Encontro Internacional de Tradutores*. Fortaleza, CE, Brazil.

Cameron, L. 2003. Metaphors in the learning of science: A discourse focus. *British Educational Research Journal, 28*(5)

_____. 2003. *Metaphor in Educational Discourse*. London: Continuum.

Cameron, L., & Deignan, A. 2003. Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse. *Metaphor and Symbol, 18*(3), 149-160.

Charteris-Black, J. 2004. *Corpus Approaches to Critical Metaphor Analysis*. Basingstoke: Palgrave Macmillan.

Deignan, A. 1999a. Corpus-based research into metaphor. In L. Cameron & G. Low (Eds.) *Researching and Applying Metaphor* (pp. 203-220). Cambridge: Cambridge University Press.

_____. 1999b. Linguistic metaphor and collocation in nonliterary corpus data. *Metaphor and Symbol, 14*(1): 19-36.

_____. 2005. *Metaphor and Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins.

Deignan, A., & Potter, L. 2004. A corpus study of metaphors and metonyms in English and Italian. *Journal of Pragmatics, 36*(7): 1231-1252.

Gibbs, R. W., & Franks, H. 2002. Embodied metaphor in women's narratives about their experiences with cancer. *Health Communication, 14*(2): 139-165.

Hoey, M. 1991. *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Koller, V. 2004a. Businesswomen and war metaphors: 'Possessive, jealous and pugnacious'? *Journal of Sociolinguistics, 8*(1): 3-22.

_____. 2004b. *Metaphor and Gender in Business Media Discourse: A Critical Cognitive Study*. Basingstoke: Palgrave Macmillan.

Lakoff, G. 1993. The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and Thought* (pp. 202-251). Cambridge: Cambridge University Press.

Minke, K. A., & Haynes, S. N. 2002. Sampling issues. In J. C. Thomas & M. Hersen (Eds.) *Understanding Research in Clinical and Counseling Psychology* (pp. 69-96). Hillsdale, NJ: Lawrence Earlbau.

MOHADJER, L., & HUSSAIN CHOUDHRY, G. 2001. Adjusting for missing data in low-income surveys. In M. Van der Ploeg & R. A. Moffitt & C. F. Citro (Eds.) *Studies of Welfare Populations: Data Collection and Research Issues* (pp. 129-156). Washington, DC: National Academy Press.

MOON, R. 1998. *Fixed Expressions and Idioms in English - A Corpus-Based Approach*. Oxford: Clarendon Press.

PARTINGTON, A. 1998. *Patterns and Meanings - Using Corpora for English Language Research and Teaching*. Amsterdam/Philadelphia: John Benjamins.

PEDERSEN, T., & PATWARDHAN, S. 2002. distance Perl package (Version 0.1): University of Minnesota, Duluth.

SCOTT, M. 1998. WordSmith Tools (Version 3). Oxford: Oxford University Press.

SEMINO, E., HEYWOOD, J., & SHORT, M. 2004. Methodological problems in the analysis of metaphors in a corpus of conversations about cancer. *Journal of Pragmatics, 36*: 1271-1294.

SINCLAIR, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

STUBBS, M. 1995. Collocations and Cultural Connotations of Common Words. *Linguistics and Education,* 7: 379-390.

_____. 2002. Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics,* 7(2): 215-244.

WHITE, M. 2003. Metaphor and economics: The case of 'growth'. *English for Specific Purposes, 22*(3): 131-151.