

Metaphor in corpora: a corpus-driven analysis of Applied Linguistics dissertations¹



Tony Berber Sardinha²
Pontifícia Universidade Católica de São Paulo - PUC-SP

O presente estudo visa a desenvolver uma metodologia para identificar metáforas em corpora. O procedimento é baseado no desejo de que o computador possa fornecer um repertório de candidatos a metáfora no corpus sem que tivesse tido acesso a uma lista de metáforas possíveis naquele corpus. A metodologia trabalha com uma seleção inicial de palavras, partindo então para a detecção dos colocados em comum entre essas palavras e para o cálculo da distância semântica daqueles pares de palavras que possuam um número mínimo de colocados em comum. Os casos que satisfazem esses critérios são examinados cuidadosamente pelo pesquisador por meio de concordâncias. Essa metodologia foi aplicada a um corpus de dissertações de mestrado de Linguística Aplicada defendidas no Brasil. O trabalho enfatiza a importância do uso de metáforas nas dissertações de mestrado, como uma maneira de os novos pesquisadores demonstrarem pertencimento à Linguística Aplicada.

This study develops a methodology for finding metaphors in corpora. The procedure is based on the wish that, without a prior list of metaphors, the computer would provide a number of possible metaphor candidates. The methodology works by selecting an initial pool of word types in the corpus, finding shared collocates between pairs of those words and then computing a semantic distance measure for those word pairs which have a requisite number of mutual collocates. Cases which satisfy these criteria were then concordanced and interpreted. This methodology was applied to a corpus of MA dissertations in Applied Linguistics, completed in Brazil. The paper highlights the importance of the use of metaphors by novice Applied Linguistic researchers.

¹ I am indebted to my colleagues who read a previous version of this paper and to the two anonymous reviewers. I am particularly grateful to Doug Biber for his comments on an earlier draft.

² The author wishes to acknowledge the financial support provided by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brasília, Brazil) under grant 350455/2003-1, and by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasília, Brazil) under grant 397/04.



Introduction

The field of metaphor studies has attracted a growing number of researchers. This is partly due to the fact that there is a strong realization that metaphors are part and parcel of everyday life. Metaphors organize the way people think (LAKOFF; JOHNSON, 1980) and interact (CAMERON, 2003).

At the same time, in the last two decades there has been a major upturn in linguistics, as electronic corpora have been seen as indispensable element in language research (HUNSTON, 2002; BERBER SARDINHA, 2004, 2005).

In metaphor studies as well, corpus-based studies have become more common, for a number of reasons. One of them is that unlike investigations that rely on intuition, corpus-based studies can offer reliable information about the use of metaphors in language. Another is that corpora typically include large amounts of data, which can be searched to provide information about the frequency of known metaphorical expressions. Yet another is that genre or register-specific corpora can be explored to indicate metaphors that are typical of certain fields or subject areas.

There are several metaphor studies from a corpus perspective available in the literature (CAMERON; DEIGNAN, 2003; CHARTERIS-BLACK, 2004; MASON, 2004; DEIGNAN, 2005; STEFANOWITTSCH; GRIES, 2006). All of these studies provide a wealth of information on the patterning, frequency and distribution of metaphors in both general and specialized language, which would not be possible without the use of electronic corpora and computational techniques.

As electronic text becomes easier to collect, and large pre-compiled corpora become more easily available, metaphor analysts are faced with a challenge: how to find metaphors in a large body of data without being able to read the whole corpus?

Typically, metaphor analysts have relied on one or more of the following strategies for coping with large amounts of electronic texts (DEIGNAN, 2005, p. 92-93):

- based on previous literature or intuition (or both), draw a list of metaphorical expressions or of words that may be part of such expressions and search for those using a concordancer;
- read portions of the corpus and draw a list of expressions or words to search for in the corpus using a concordancer;

- search for cues, that is, words or expressions that may be found near metaphors, using a concordancer.

Although these strategies have enabled researchers to meet the goals of their respective research projects, I would argue there need to be other ways to detect metaphors in electronic corpora.

The rest of this paper describes one such method, which we may label 'corpus-driven' or 'data-driven', since it is based on the principle that the metaphors should emerge out of the data (TOGNINI-BONELLI, 2001; PETERS; WILKS, 2003), instead of being just searched for in the data.

The first thing to bear in mind is that at the current state of the art in computation, a fully automated method of finding metaphors is impossible to achieve. This would necessitate that computers ultimately understand human language and be able to think and interact like humans, which is of course beyond the capacity of our most sophisticated computers.

Realistically, then, what we should expect is that our method throws up a number of possible metaphorical candidates for the analyst to consider. At this early stage, we should not expect the method to be particularly accurate. We should expect that the method be used as tool that would help analysts do their job more efficiently, but not that it would do the whole job for them.

Procedures

In this section, I will detail the procedures involved in looking for metaphorical candidates in a corpus. The corpus is a collection of 36 Master's dissertations in Applied Linguistics, all completed in the Applied Linguistics and Language Studies Postgraduate Program at PUC-SP (the Pontifical Catholic University of São Paulo, Brazil). The dissertations were written in Portuguese. The corpus totals 502,438 tokens and 29,537 types.

The basic procedure followed in this study consisted in comparing the way pairs of words were used in the corpus. For example, suppose that 'time' and 'money' appeared in the corpus. Now, if we looked at 'time', we might find that it occurred with words such as 'waste', 'save', 'spend' and 'synchronize', among others, at a certain distance (say, three words on either side). And if we looked at 'money', we might find out that it occurred with words such as 'waste', 'save', 'spend' and 'receive', among others, within the same distance as our previous word (three words). We can call the two words in the pair that we are focusing on our 'focus words', and the set of words co-occurring with each one its 'collocate set'.

Next, if we compared the two collocate sets, we would find that the two focus words share a number of words, such as 'waste', 'save' and 'spend'. Finally, we would extract concordance lines of our corpus for each of the focus words and decide whether a metaphor was underlying these uses of 'time' and 'money'.

By looking at the concordance lines, expressions such as 'save time', 'waste time' and 'spend time' on the one hand, and 'save money', 'waste money' and 'spend money' on the other, would suggest that the expressions formed with the focus word 'time' are metaphorical. We would call them metaphors because there is an implied mapping between two domains: the domain of 'time' and the domain of 'money'. 'Time' was conceptualized in terms of 'money': it can be saved, spent and wasted. The evidence provided by the corpus would point toward the existence of the well-known *TIME IS MONEY* conceptual metaphor. 'Time' is the target domain, and 'money' is the source domain. In the expressions, we will call a focus word such as 'time' the vehicle of the metaphor.

Note that in this fictitious example, there was no instance of 'time' being in the collocate set of 'money' or vice-versa, which means there would be no occurrence of the expression 'time is money' in the corpus. Note also that we had not anticipated the occurrence of the focus words, the collocate set, the shared collocates or the conceptual metaphor.

To recap, these are the terms that have been used so far to describe the analysis of the corpus:

- focus word: Any word from the corpus that is being considered at any particular point in the analysis. In concordancing, this would be called the 'node' word;
- focus word pair: A pair of focus words being compared at any point in the analysis;
- collocate: Any lexical word occurring in the vicinity of a focus word a certain number of times. The vicinity is a span of x words on either side of the focus word. A lexical word is any word that is either a noun, adjective, verb, adverb or numeral. Both the size of the span and the minimum frequency will be discussed later;
- collocate set: The set of collocates for a particular focus word;
- vehicle: A focus word that is used metaphorically in the corpus.

There are a number of decisions that had to be taken in order to implement these procedures.

The first concerned which words would be the focus words. Ideally, all words would be focus words, which means that each word would be paired with any other word (excluding itself). In our corpus, this would involve comparing 29,537 words types with each other (but excluding those pairs formed by identical words). The number of non-repeated combinations³ involved in 29,537 elements is given by $29537_C_2 = 29537! / 2!(29537-2)!$, or 436,202,416. That means there are over 436 million word pairs in the corpus. Since focus word pairs need to be interpreted, a number as high as that would render the interpretation impossible. Clearly, a selection procedure has to be put in place.

The decision was taken to select a sample of the word types in the corpus. There are several possible selection criteria that could be applied at this stage in the analysis. For example, we could select only nouns, as these are the prototypical vehicles in metaphor analysis. However, this would still leave us with thousands of focus words, resulting in over 40 million pairs. Or we could keep nouns and verbs, but this would only raise the resulting number of pairs.

Word frequency is another criterion for limiting initial focus word selection. The problem here is that there are no established guidelines for choosing a cut-off frequency point, above which words would be retained for pairing up.

The solution was to use frequency markedness as the initial selection criterion. For any one particular word, frequency markedness means the degree of difference between its frequency in the dissertation corpus as compared to its frequency in a different corpus, normally a larger corpus representing the range of registers in the language. We will use the terms 'study corpus' to refer to our dissertation corpus, and 'reference corpus' to refer to the corpus with which the frequencies were compared.

Our reference corpus was the Bank of Portuguese. This is a large open general corpus of Brazilian Portuguese, compiled as part of the DIRECT Project at PUC/SP (the Pontifical Catholic University of Sao Paulo).⁴ At the time of the comparison, the Bank held over 220 million words, taken from contemporary Brazilian speech and writing.

³ We mean combinations in which the order of the elements does not matter. Hence, comparing 'time' and 'money' is the same as comparing 'money' and 'time'.

⁴ <http://lael.pucsp.br/direct>.

Frequency markedness can be understood by referring to relative frequencies. If the word 'money' appeared in our study corpus 50 times (out of 502,438, in the whole corpus), it would have a relative frequency equal to .01%. If it appeared 500 times in the reference corpus (out of 200 million), it would be just .0002%. The difference between the two relative frequencies seems to be marked, with the frequency in the study corpus being 50 times larger than in the reference corpus. This appears to be large enough a difference to warrant the label 'marked frequency', but we do not know for sure until we apply a statistical test. This needs to be done for each one of the nearly 30 thousand words in the study corpus. Fortunately, a program such as KeyWords, which is part of WordSmith Tools (SCOTT, 1997) does this automatically. It takes as input two files, one being a frequency word list for the study corpus and another a word list for the reference corpus. It then outputs a list of 'key words', or words whose frequencies are marked, according to a statistical test. The frequency markedness test we applied in this study was the Log-Likelihood, which is default in KeyWords. We also adopted another default, the 500 key word list, ordered by degree of markedness (which is called 'keyness' in the program). This means that the top most word on this list has the most marked frequency of all; the second word has the second most marked frequency and so on.

At this stage the number of focus words has been drastically reduced (from nearly 30 thousand to 500). Nevertheless, 500 words would still generate 124,750 word pairs... A further reduction was necessary to bring the number of word pairs down to a manageable level. By manageable is meant a number not greater than one thousand.

The decision was then taken to select the top 100 words from the keyword list. From these, function words, proper names, non-Portuguese words and word classes other than nouns and verbs were removed. Ambiguous forms (such as 'lingüístico', which may be a noun or an adjective) were ignored. The result was a 53-word list. This would result in 1378 word pairs, a number still above the manageable limit set above. However, it was felt that imposing any more restrictions at this point would likely jeopardize the analysis, by eliminating words that might be vehicles. Further selection mechanisms would be put in place later on, based on criteria other than frequency markedness.

The resulting 53 keywords appear below (numbers in front of the words indicate ranking on the keyword list).

TABLE 1
Words with marked frequency selected for analysis

12 LINGUAGEM (language in use)	64 TEXTOS (texts)
15 ANÁLISE (analysis)	65 REFLEXÃO (reflection)
17 PROCESSO (process)	67 MENSAGENS (messages)
21 TRABALHO (work)	68 DADOS (data)
23 LÍNGUA (language)	69 PAPEL (role)
25 CONTEXTO (context)	70 CONSTRUÇÃO (construction)
26 FUNÇÃO (function)	71 SINAIS (signs)
28 AÇÃO (action)	72 LISTA (list)
29 CORPUS	73 ELEMENTOS (elements)
30 TEXTO (text)	74 MENSAGEM (message)
32 DISCURSO (discourse)	75 QUADRO (framework, table)
34 USO (use)	77 RESULTADOS (results)
35 PALAVRAS (words)	79 CARACTERÍSTICAS (characteristics)
36 RELAÇÃO (relationship, relation)	80 RELAÇÕES (relationship, relation)
39 PESQUISA (research)	83 ESTRUTURA (structure)
40 MODO (way, mode, fashion)	84 CONHECIMENTO (knowledge)
42 DESENVOLVIMENTO (development)	85 AULA (class)
43 PARTICIPANTES (participants)	88 ASPECTOS (aspects)
47 INFORMAÇÃO (information)	89 PRÁTICA (practice)
48 ESTUDO (study)	90 FORMAS (ways)
52 FALA (talk)	91 OBJETIVO (objective)
54 INTERAÇÃO (interaction)	95 CORRESPONDÊNCIA (correspondence)
55 FUNÇÕES (roles, capacities)	97 PROFESSORA (teacher)
56 COMUNICAÇÃO (communication)	98 PROCESSOS (processes)
60 PROFESSOR (teacher)	99 PROFESSORES (teachers)
61 GÊNERO (genre)	100 ALUNOS (students)
63 ENSINO (teaching)	

Note: words in brackets are their translations.

The next step was to compute the collocate sets of each keyword. They were extracted using a computer program written by the author. A collocate was defined as any lexical word occurring in the vicinity of a focus word a certain number of times. A lexical word is one that is either a noun,

a verb, an adjective, an adverb or a numeral. It was necessary to establish the extent of the vicinity as well as the lowest allowable frequency. The vicinity, or span, was set at two words intervening. This means that if 'save' occurs one, two or three words before or after 'time' it is within the allowed distance. On the other hand, if it occurs four or more words away from 'time', it is outside the allowed limits and is consequently dropped. As far as the lowest frequency, it was set at two, meaning that words that occurred once in the allowed distance were dropped.

After these limits were imposed on the data, there remained 1128 focus word pairs (and the same number of collocate set pairs).

The next step consisted of comparing the 1158 collocate set pairs. Remember, at this stage we wanted to determine which words 'time' and 'money', for example, had in common. We were interested in keeping words such as 'spend', 'save', and 'waste', since these are (a) shared and (b) common. The challenge at this point was to determine what is common between two collocate sets. This is a problem that has to do with setting a lowest possible number of shared collocates. By that is meant the number of collocates in common between two collocate sets. In our ongoing example, 'time' and 'money' have three shared collocates: 'spend', 'waste' and 'save'. Again, there are no clear guidelines for determining the ideal number of shared collocates. Whatever the number, it should be based on the idea that two sets of words have some potential meaning in common. In our example, 'time' and 'money' do have meaning in common. This sharing of meaning is the result of a metaphorical mapping in Western culture. Given the absence of an attested number for this purpose, we decided to use three as the minimum number of shared words between two collocate sets.⁵

A computer program was created by the author to calculate shared collocates between pairs of collocates sets. The program matched each pair of collocate sets and dropped those that did not have three or more collocates in common. This reduced the 1158 pairs to 907, which is below the 'manageable' level that we hoped to reach.

⁵Three is also the criterial number of three lexical links proposed by (HOEY, 1991) as an indication of meaning sharing between two sentences. Although we are not comparing sentences or studying cohesion, perhaps at some level there is a similarity between Hoey's criterion and ours, given that collocates occur originally in sentences.

However, there is a potential problem with the collocate sets with mutual collocates: they may refer to words which are closely related, such as 'money' and 'dollars', or 'money' and 'cash', and so on. Any of these pairs might share collocates such as 'save', 'spend' and 'invest'.

In order to address this problem we decided to apply a strategy that was based on semantic meaning, rather than frequency. Going back to our example: given that we would have established by now that 'time' and 'money' had at least three collocates in common ('save', 'waste', 'spend'), we would need to decide whether there was any possible metaphoricity between 'time' and 'money'. Suppose another pair of focus words had passed all the tests so far and had the same collocate sets: 'money' and 'dollars'. In this case, there is no metaphoricity between these two words, even though they have fulfilled all the criteria established so far. We needed a method which would at the same time weed out word pairs such as 'money' and 'dollars' and keep pairs such as 'time' and 'money'. It appeared to us that this was a problem of semantic distance. 'Dollar' and 'money' are semantically close, since 'dollar' is used as money. On the other hand, 'time' and 'money' are semantically distant. The similarity is actually conveyed metaphorically, by conceptualizing one in terms of the other. Such a method of assessing semantic distance would have to be automatic, since (a) a manual comparison would take too long, and (b) the comparison would have to be reliable.

A possible solution was found in WordNet, an electronic dictionary of English (MILLER, 1990). The dictionary is organized as a database, with fields marked up to allow computational processing. In it, words are grouped into sets according to several sense relations, such as synonymy, antonymy, meronymy, etc. WordNet may be downloaded off the Internet and run on personal computers. It can be accessed through its own software or it can be searched by other custom built programs. WordNet was considered a possible solution to the challenges of weeding out words that were too close in meaning while at the same time keeping those that were distant in meaning because it can be used with a program called 'distance' (PEDERSEN; PATWARDHAN, 2002) that does exactly what we wanted: given a pair of words, it searches WordNet for all possible senses and calculates the difference in meaning between them, reporting the results in terms of a numerical score.⁶ The scores are computed by several statistical tests, which are described in its documentation. For instance, in the Leacock

Chodorow measure, a lower value represents a greater distance in meaning, while a higher value indicates closeness in meaning. To 'time and money', it gives a score of 1.51, and to 'dollar' and 'money', 1.86. This suggests that 'time and money' are more distant in meaning.

Again, there are no cut-off points to suggest possible values under which we should expect more metaphoricity in our focus word pairs. In order to set tentative cut-off points for the semantic distance scores, we decided to carry out a mini-study on the relationship between semantic distance scores, as reported by the 'distance' program, and metaphoricity, as expressed by pairs of words normally found in the literature as being part of metaphorical expressions (e.g. time and money). For the mini-study, three groups of words were set up. The first group (labelled 'H-r' in the table below) contained highly related words; some of these were taken from the *distance* help files (such as gem and jewel), while others were made up by the author. The expectation was that word pairs in this group would have high scores. The second group ('Unr') was made up of unrelated words (e.g. noon and string), and was created following the same principles. This group was expected to score low. The last group ('Met') contained word pairs which were part of well-known metaphors (e.g. time and money). This group was expected to score in between the high and low groups.

The pairs were submitted to the *distance* program and their scores were collected. The results appear in the table below.

⁶ An online version of 'distance' is available on the web at <http://lael.pucsp.br/corpora/similarity>.

TABLE 2
Mini-study: Semantic distance scores provided
by *distance* program

Group	Word pair		Score
H-r	gem	jewel	3.46
H-r	car	automobile	3.46
H-r	cup	container	2.77
H-r	violin	instrument	1.85
H-r	football	sport	2.36
H-r	teacher	worker	1.67
H-r	doll	toy	2.77
H-r	shirt	clothing	2.36
H-r	book	book	3.46
H-r	comb	brush	2.36
Unr	noon	string	0.98
Unr	elephant	pencil	0.75
Unr	computer	cloud	1.16
Unr	gem	automobile	1.26
Unr	violin	sport	0.98
Unr	toy	worker	1.51
Unr	shirt	book	1.38
Unr	football	comb	1.51
Unr	automobile	cup	1.51
Unr	computer	sport	1.85
Met	time	money	1.51
Met	love	journey	1.06
Met	love	war	1.38
Met	argument	war	1.38
Met	bad	down	1.26
Met	theory	building	1.26
Met	idea	plant	1.67
Met	idea	people	1.51
Met	idea	product	2.07
Met	life	container	1.51

The range of values for the scores of each group is the following:

TABLE 3
Mini-study: Range of semantic distance scores
for groups of words

Group	Min value	Max value
Highly Related	1.67	3.46
Unrelated	0.75	1.85
Metaphors	1.06	2.07

The ranges matched our expectations. Despite some overlap in scores, the value for the metaphor group stood roughly between 1 and 2, with unrelated words a little lower, and highly related words going from 1.67 up to the 3.47 ceiling (achieved with two identical words such as ‘book book’ or highly similar ones such as ‘gem jewel’).

Based on these numbers, it was decided that a cut-off point of 2 would be best suited. Word pairs with similarity values above this threshold were disregarded. A lower limit was not set, as there was no reason to suppose that seemingly unrelated words in the corpus could not be potential metaphors. The cut-off point of 2 would authorize a spurious pair like ‘money’ and ‘dollar’, though, which is undesirable, but we decided to err on the side of inclusion, not exclusion, as these scores are meant to be another filter for the data. As this stage is not the end of the analysis, any data that were retained here would still be processed further. In other words, any spurious pairs that were retained at this stage would not be automatically considered metaphors.

As a result of this last step, 737 word pairs remained as potential metaphor candidates (that is, they had a score of 2 or lower as reported by the *distance* program).

In order to recap the steps taken so far in filtering the data for manual analysis, we provide Tab. 4 below:

TABLE 4
Summary of data filtering at each stage of the research

		Resulting word pairs
Total word types in the corpus	29537	436,202,416
Words with marked frequency returned by WordSmith KeyWords	500	124,750
Words with marked frequency retained for analysis (sampled from top 100 words with marked frequency)	53	1,378
Marked frequency word pairs with non-empty lexical collocate sets ...		1128
... sharing 3 or more links ...		907
... with a <i>distance</i> score + 2		737

Of these 737 word pairs, I selected four cases to discuss in detail below, including instances of both metaphor and metonymy. The analysis of the data thrown up by the automatic procedures is highly interpretive, and proceeded as follows. Firstly, a concordance was run (using a Unix-based program developed by the author) for each word in the focus word pair. Secondly, the concordance was analyzed by the researcher for metaphorical expressions.

We show below examples of two metaphors and two non-metaphors found in the data.

Results

Metaphor: Trabalho e conhecimento (Work and knowledge)

‘Trabalho’ (work) and ‘conhecimento’ (knowledge) had a similarity score of 1.85. They shared 12 collocates, which are shown below. The list is arranged in frequency order, with the most frequent collocates at the top.

TABLE 5
Shared collocates for ‘trabalho’ (work)
and ‘conhecimento’ (knowledge)

Trabalho (work), conhecimento (knowledge):
desenvolvimento (development)
professor (teacher)
pedagógico (pedagogic)
contexto (context)
sala (room)
área (area)
tipo (type)
próprio (self)
social (social)
linguagem (language)
forma (way, form, shape)
alunos (students)

According to the collocates, ‘trabalho’ and ‘conhecimento’ have several mutual meaning mappings. Both are things that are ‘developed’ (desenvolvimento), both are construed as being related to the dealings of a ‘teacher’ (professor) and of ‘students’ (alunos), as being ‘pedagogic’ in nature, as being part of a ‘context’, as taking place in a ‘(class)room’ (sala), etc.

On the basis of this evidence, these concepts seem to be linked by means of an underlying metaphor which may be expressed as KNOWLEDGE AS WORK. This metaphor is compatible with the theoretical orientation of a large share of the investigations reported in the dissertations, which favor the view of knowledge being something that one works toward, or that is worked on by people, rather than something that is gained passively.

The concordance below illustrates some of the uses of these two words in the dissertations.

) exerce no desenvolvimento do conhecimento e na formação do e apesar do desenvolvimento do conhecimento da estrutura e do ribua para o desenvolvimento do conhecimento sobre o ser profes direta com o desenvolvimento do conhecimento pedagógico dos al damental no desenvolvimento do conhecimento do professor. A visadas e do desenvolvimento do trabalho em cada sala de aula.

segurança no desenvolvimento do trabalho pedagógico. 179 CAP. 4
ível o pleno desenvolvimento do trabalho de Antônio Carlos Jobim
desenvolver pesquisas sobre “o conhecimento do professor”. De
idáticas acabam por esvaziar o trabalho do professor, na medida
ências podem ser detectadas no trabalho do professor, na metodo
as formas de representação do conhecimento do professor. Ao e

The concordance shows the basic pattern into which the two words enter, which is N de/do N (N of N). The two words typically function as part of a noun group.

Metaphor: Ensino e construção (teaching and construction)

Another pair which is below the 2 point distance cut-off mark is ‘ensino’ and ‘construção’ (teaching and construction). These had a score of 1.85 on the distance test. Together, they have 7 mutual collocates, which are displayed below.

TABLE 6
Shared collocates for ‘ensino’ (teaching)
and ‘construção’ (construction)

ensino (teaching), construção (construction)
processo (process)
visão (vision, view)
relação (relation, relationship)
linguagem (language)
escrita (writing)
processos (processes)
conceitos (concepts)

The similarity between the two words, as indicated by their collocates, revolves around the fact that both are depicted as ‘processes’ (processo, processos) which people have particular ‘visions for’ or ‘views on’ (visão) and which stand in a particular ‘relation’ (relação) to other ‘concepts’ (conceitos) or things. Further, they are closely related to the workings of ‘language’ (linguagem), typically to the written variety (escrita). The underlying metaphor which seems to be working here could be described as ‘TEACHING AS CONSTRUCTION’. The theoretical emphasis in the dissertations

is on the constructive nature of teaching, with construction being clearly portrayed as a process-oriented rather than a product-oriented activity. This favors stating this as being ‘construction’ instead of ‘a construction’, as the latter would not imply the ongoing nature of teaching, but a finished product.

Some of the occurrences of the pair are illustrated in the concordance below:

línguas; c) objetivos e processos de ensino e aprendizagem. d) prepara-
 ções do professor nos processos de ensino e aprendizagem, e observo q
 bjetivos de ensino. 6- Processos de ensino e abordagens do processo.
 níficos, relativos aos processos de ensino-aprendizagem, como discutid
 para compreender os processos de ensino-aprendizagem, para desenha
 cial e educacional dos processos de ensino/aprendizagem de línguas. Ca
 cial e educacional dos processos de ensino/aprendizagem de línguas. Ca
 a compreensão dos processos em construção. Mas, aí, surgiu outro p
 que se fundamenta minha visão de construção de conhecimento e de le
 que realizei, entre a minha visão de construção do conhecimento e o pa
 teórica que envolve minha visão de construção do conhecimento e, port
 m O quadro 1 reflete uma visão de ensino centrada no aluno, em que a

Again, the basic pattern formed by each of the words is N de N (N of N).

Metonymy: reflexão e prática (reflection and practice)

A pair that is related to the above one is reflexão and prática (reflection and practice). They obtained a similarity score of 1.51. Their 18 common collocates appear below.

TABLE 7
Shared collocates for ‘reflexão’ (reflection)
and ‘prática’ (practice)

reflexão (reflection), prática (practice)
crítica (critical)
conceito (concept)
ações (actions)
tipo (type, kind)
voltada (geared to)
discussão (discussion)

oportunidades (opportunities)
 análise (analysis)
 alunos (students)
 professor (teacher)
 relação (relation, relationship)
 professores (teachers)
 perspectiva (perspective)
 forma (way, form, shape)
 fazer (do, make)
 atividades (activities)
 experiência (experience)

As the collocates reveal, these are ‘concepts’ (conceitos) which relate to ‘actions’ (ações), ‘activities’ (atividades) and ‘experiences’ (experiência) involving ‘teachers’ (professor/es) and ‘students’ (alunos). Further common characteristics of reflection and practice are that they occur at certain moments, if they are given the chance (oportunidade), and that they are categorized as being of particular ‘kinds’ (tipo, forma), most notably as ‘critical’ (crítica). The general meaning of ‘practice’ that comes across in the dissertations is that of habitual actions of a person in a professional or educational context. Reflection, in turn, refers to the act of thinking, on the part of a teacher, about what they do professionally, how and why they do it in the way they do it, and so on. The mutual meaning mappings between these two concepts suggest that reflection is (or should be, as the authors seem to be making a case for it) as much a part of the practice of a teacher as their other engagements. This also conceptualizes reflection as part of a teacher’s professional development, which implies a contiguous relationship, hence a metonymy.

The concordance below exemplifies some of the occurrences of ‘reflexão’ and ‘prática’.

rspectiva, Stake (1987) relaciona reflexão crítica ao conceito de tr
 ra 300 horas; (3) contextos para reflexão crítica através da observ
 ntarei minha própria definição de reflexão crítica com a qual trabal
 nglês, desenvolver o processo de reflexão crítica com alunos em fa
 r embasamento teórico e de uma prática crítica-reflexiva tive segur
 reflexivos e embasados em uma prática crítica. Atualmente, a parti

tituição de professores com uma prática crítica. Este estudo está ental. É do estudo das ações da prática que a teoria se alimenta seja feito um relato de ações da prática que ela considera relevant 87:148-149). 1.5.1 As ações da reflexão crítica Baseado em estu 3.3 A contribuição das ações da reflexão crítica Na discussão a ndamento em todas as ações da reflexão crítica geraram mais opo

The common patterns of use are N da N (N of N) and N ADJ.

Metonymy: ensino e reflexão (teaching and reflection)

A pair which is closely associated to the previous one is ‘ensino e reflexão’ (teaching and reflection). This received a similarity score of 1.38, which is slightly lower than that for reflection and practice. The 11 common collocates appear below.

TABLE 8
Shared collocates for ‘ensino’ (teaching)
and ‘reflexão’ (reflection)

ensino (teaching), reflexão (reflection)
processo (process)
visão (view)
relação (relation, relationship)
perspectiva (perspective)
prática (practice)
projeto (project)
professores (teachers)
forma (way, form, shape)
atividades (activities)
professor (teacher)
experiência (experience)

We begin to see links between this pair and previous ones. Here, teaching and reflection both are depicted as a ‘process’ relating to certain practices (prática) (typically that of ‘teachers’ (professor/es), on which particular ‘perspectives’ (perspectiva) are taken, and of which people hold certain ‘views’ (visão). These concepts are used in carrying out ‘projects’ (projeto) in the area of teacher development. The general view held in the

dissertations is that reflection is beneficial to teachers, as it helps them to develop professionally, to view themselves as educators rather than just as fountains of knowledge. Reflection is thus seen as being an integral part of contemporary teacher development. This seems to suggest another metonymy.

The concordance below illustrates some of the uses of the two words in the corpus:

ias denotando alguma visão de ensino-aprendizagem. Exemplo
o nos revela nenhuma visão de ensino-aprendizagem. Ser fund
seus postulados, uma visão de ensino/aprendizagem de língua
uagem de Bakhtin e a visão de ensino/desenvolvimento e apren
ociais mais amplos. A visão de reflexão abordada neste estudo
ntão que tive acesso à visão de reflexão crítica como proposta p
da está imbuída dessa visão de reflexão crítica que pressupõe
e aqui ressaltar que a visão de reflexão discutida por Zabalza (c
tegorizá-las a partir da visão de reflexão que a explica. Um dos
otados em relação ao processo de ensino/aprendizagem e ao papel d
alunos) em relação ao processo de ensino/aprendizagem e na dinami
-CONSTRUTIVISTA O processo de ensino/aprendizagem vem sendo há
consciente de que no processo de ensino/aprendizagem, não há gara
eram sempre no meu processo de ensino/aprendizagem. À Bete e
la teoriza sobre nosso processo de reflexão: 130. PR: E é assim. Isso
No entanto, vejo que o processo de reflexão através da narrativa de vid
rtunidade de viver um processo de reflexão bastante profundo, visitei m
Inglês, desenvolver o processo de reflexão crítica com alunos em fas
isa é, na verdade, um processo de reflexão crítica com momentos de
senvolver e analisar o processo de reflexão crítica com os alunos-mes
locados de lado. A perspectiva da reflexão crítica (Kemmis, 1987) col
m relação a como a perspectiva de ensino-aprendizagem foi abordada

The dominant phraseology here is again N de N (N of N).

The basic pattern: noun groups

All of the word pairs described above (and most of the others, arguably) entered into a common linguistic co-occurrence pattern: N prep N, that is, noun groups. Noun groups are a device for encoding linguistic

information into pre-packaged blocks. Noun groups are downranked elements, which are situated below the clause (HALLIDAY, 1994). The consequence of presenting most concepts in the form of noun groups is that the information comes across as assumed, rather than asserted (LOW, 1999).

The structure of the noun group in Portuguese takes a prepositional phrase as a post-modifier. Take, for instance, an occurrence from the previous concordance:

visão de ensino-aprendizagem

In this case, 'visão' is the Head, followed by 'de ensino-aprendizagem', a prepositional post-modifier. 'De', in turn, is a 'Minor Process' (Halliday, 1994), which reveals that there are participants involved there, even though they have been 'deleted' in the process of downranking. According to Halliday, this participant may originally be of several kinds, but for all purposes it is Range, which serves to 'specify the range or scope of the process', 'to define its co-ordinates' (HALLIDAY, 1994, p.146).

Schematically, this structure may be described as:

TABLE 9
Structure of nominal phrase 'visão de ensino-aprendizagem'
(vision for teaching and learning)

visão	de	ensino-aprendizagem
Head	Post-modifier	
α	β	
	Minor Process	Participant: Range

The actual process realized by the preposition cannot be clearly reinstated – its clear meaning has been lost in the process of downranking. In this particular case, it could be interpreted tentatively as 'is': 'ensino aprendizagem' is 'a visão'. By bringing back a potential process to replace the preposition, the nominal group assumes a configuration which resembles a metaphor of the 'X is Y' kind. The point here is that there is a meaning mapping here, albeit obscure, between the Head and the Post-Modifier which might be interpreted as metaphorical. If we assume that such a reading is possible, then we have a stronger case for seeing a metaphorical relationship among the nominal groups formed by the word pairs. If 'ensino' is 'visão', and 'reflexão' is 'visão', then by entailment 'ensino' is 'reflexão'.

Discussion

As far as the analysis is concerned, both the metaphors and the metonymies discussed in more detail point to a constructivist paradigm to teacher education. The figurative expressions are concepts used by educators and researchers in this paradigm.

This does not mean, however, that the way in which these metaphors and metonymies were actually used may be generalized to the field of teacher education as a whole. Perhaps their use reflects the particular flavor of teacher education that is typical of the local context of the Postgraduate program in which the research projects were completed. A methodology such as the one reported here, which does not start out with a pre-determined list of metaphors, may be at an advantage for picking up both widespread and 'local' metaphors.

There are correspondences between metaphors found here and those found in previous studies. For instance, Cortazzi and Jin (1999) discuss how scaffolding came to be used in education academic discourse to convey the metaphorical notion of learning as being supported by peers. The value imbued in this is that learning is not (or should not be) an individual endeavor. Similarly, the same authors point out that another associated metaphor, construction, is typical of education discourse. These views are characteristic of the constructivist paradigm of learning, which carries in itself a metaphorical statement. The fact that metaphorical representations typical of education are widely present in a different field, namely Applied Linguistics, reveals the transdisciplinary nature of Applied Linguistics.

Going back to the beginning of the analysis, we will notice that the particular sub-field of teacher development is dominant in the pool of words with marked frequency. The lexis of teacher education is thus predominant in the corpus, but that did not necessarily predict the metaphors and metonymies that were actually found.

The presence of figurative language in a corpus of texts written by students such as ours suggests that metaphors and metonymies are a means of entering and being accepted in the discourse community. The intertextual connections established between each dissertation and the others, and between them all and the academic discourse from which they borrow their concepts suggest students are constantly borrowing 'voices' from other's discourse as a means to legitimize their own (BAKHTIN, 1981). What is noteworthy about such borrowings in our case is that legitimisation involves a great deal of appropriating metaphors.

As far as the methodology is concerned, one of the questions that this study raises is to what extent these procedures add to the more subjective human process of identifying metaphors. One way in which they can aid the researcher is by providing a means to run a comb through a large corpus. How fine a comb will depend on the aims of the research. More stringent settings (lower p value for WordSmith key word extraction, tighter collocational span, higher number of links between sets of key words) are likely to throw up few but hopefully 'sure cases', whereas looser settings (higher p , wider span, fewer links) will return a larger but noisier set of cases to be interpreted. Given the exploratory stage of the whole set of procedures, it is perhaps wise to be less restrictive and allow for a greater level of 'noise' in the data.

Since this methodology is not driven by a previous list of metaphors, it may indicate unexpected metaphors, such as those that may have been 'hidden away' in the texts, or those which were too obvious to notice (because they were either part of the jargon or terminology of a certain discipline, or of the local discourse of a smaller community of speakers), or even those that were 'disguised' by nominal phrase constructions (such as the cases exemplified in this paper), among others. These may all escape the notice of a researcher who is reading a large set of textual data looking for metaphors.

A data-driven methodology such as ours has several other limitations. One of them is the amount of technical expertise that it requires, since the automatic part of the analysis was not done using one single piece of software. The researcher is likely to need a key word extractor (which by itself requires a reference corpus word list, something that may not be as easy to find as it seems, at least for languages other than English), a large-scale collocator (for which purpose more general concordance packages such as WordSmith Tools do not work, as they require that each word be concordanced on its own), a semantic similarity scorer, like *distance* (which in turn necessitates the installation of WordNet), not to mention several little programs to take the output from one program and feed it into the next. Another limitation is that even if the provision for the necessary software infrastructure is met, and all programs run successfully, there is no guarantee that the output contain any metaphors. This can only be confirmed by means of careful interpretation of the results. At this stage, the researcher may be faced with hundreds (if not thousands) of concordance lines to analyze and upon which to make subjective decisions about the presence or absence of

metaphors. This means that this methodology was not designed with the aim of speeding up metaphor analysis. Its goal is to allow the researcher to tackle the analysis of thousand of words without a previous list of metaphors. Finally, the filters imposed on the data may have excluded lots of metaphors from the final analysis. This is because there is no reason to suppose that there are no metaphors among the non-key words of the corpus, not least because the key words that one gets from WordSmith Tools in the first place are always relative to the features of the reference corpus chosen for comparison (a different reference corpus would mean a different set of marked frequency words). Cumulatively, this has a knock-on effect on the data, since each decision made along the way restricts the number of words that may be indicative of metaphors. The result is that there may be several metaphors expressed by words which were not picked up by the method.

Overall, what these pros and cons seem to suggest is that automatic, computer-based techniques, and subjective, interpretive procedures should be seen as complementing, rather than as opposing, each other. It is hard to believe that a researcher faced with thousands (if not millions) of words to analyze as are increasingly made available nowadays would have nothing to gain by applying at least some principled computational techniques to his/her data in order to select a sample of words to analyze.

For Corpus Linguistics, one implication of this research is that while it has been recognized that collocational dissimilarity is an indicator of sense disambiguation (revealing how words acquire particular meanings), the flip side of this is that collocational similarity may be an indication of metaphorical meaning.

Concluding remarks

The study reported in this paper looked at a particular way of extracting metaphorical candidates from an electronic corpus, using computational tools. Its main goal was to devise a set of procedures for allowing a researcher to tackle the analysis of a large corpus without a set of metaphors selected beforehand.

The methodology was based on the idea that metaphors may be signaled by pairs words with shared lexis, as 'time' and 'money', which share collocates such as 'save', 'waste' and 'spend'. As word pairs proliferate (there are millions of word pairs in our corpus of under 30 thousand words!), a series of heuristics were devised to work as filters on the data. These filters

were meant to reduce the number of word pairs in the corpus to manageable levels (e.g. below one thousand). The final set of word pairs was meant to be hand coded by the analyst, who had the final say on whether there were any metaphors at all among the candidates thrown up the computer.

This method is best described as a tool for the analyst. It does not do the whole job for him/her, it simply tries to make the job possible to do. Without a tool such as this, the analyst is left with the usual tools of the trade for metaphor researchers doing corpus analysis, which typically are a pre-defined list of metaphors or a list of metaphors drawn by reading portions of the corpus, which are then searched for using a concordancer.

The methodology stresses the importance of subjective, interpretive human analysis, given that the computer can only at best suggest possible metaphors that need to be validated by careful examination of the data. Metaphor identification cannot be left to the computer alone. But data-driven techniques can certainly lend a helping hand to the researcher who needs to sift through hundreds of thousands of words of text. It is hoped that this paper has shown some benefits as well as pitfalls of computer-aided metaphor research, and that future research addresses the many challenges that lie ahead at the interface of Corpus Linguistics and Metaphor Studies.

References

- BAKHTIN, M. *The Dialogic Imagination*. Austin: University of Texas Press, 1981.
- BERBER SARDINHA, T. *Lingüística de Corpus*. São Paulo: Manole, 2004.
- _____. (Ed.). *A Língua Portuguesa no Computador*. Campinas, São Paulo: Mercado de Letras /FAPESP, 2005.
- CAMERON, L. *Metaphor in Educational Discourse*. London: Continuum, 2003.
- CAMERON, L.; DEIGNAN, A. Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse. *Metaphor and Symbol*, v.18, n.3, p.149-160, 2003.
- CHARTERIS-BLACK, J. *Corpus Approaches to Critical Metaphor Analysis*. Basingstoke: Palgrave Macmillan, 2004.
- CORTAZZI, M.; JIN, L. Bridges to learning: metaphors of teaching, learning and language. In: CAMERON, L.; LOW, G. (Ed.). *Researching and Applying Metaphor*. Cambridge: Cambridge University Press, 1999. p.149-176.

- DEIGNAN, A. *Metaphor and Corpus Linguistics*. Amsterdam; Philadelphia: John Benjamins, 2005.
- HALLIDAY, M. A. K. *An Introduction to Functional Grammar*. London: Edward Arnold, 1994.
- HOEY, M. *Patterns of Lexis in Text*. Oxford: Oxford University Press, 1991.
- HUNSTON, S. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press, 2002.
- LAKOFF, G.; JOHNSON, M. *Metaphors We Live By*. Chicago: University of Chicago Press, 1980.
- LOW, G. D. "This paper thinks...": Investigating the acceptability of the metaphor AN ESSAY IS A PERSON. In: CAMERON, L.; LOW, G. (Ed.). *Researching and Applying Metaphor*. Cambridge: Cambridge University Press, 1999. p. 221-248.
- MASON, Z. CorMet: a computational, *corpus*-based conventional metaphor extraction system. *Computational Linguistics*, v. 30, n. 1, p. 23-44, 2004.
- MILLER, G. WordNet: An on-line lexical database. *International Journal of Lexicography*, v. 3, n. 4, p. 235-312, 1990.
- PEDERSEN, T.; PATWARDHAN, S. distance Perl package: University of Minnesota, Duluth, 2002.
- PETERS, W.; WILKS, Y. Data-Driven detection of figurative language use in electronic language resources. *Metaphor and Symbol*, v. 18, n. 3, p. 161-173, 2003.
- SCOTT, M. *WordSmith Tools*. Version 3. Oxford: Oxford University Press, 1997.
- STEFANOWITSCH, A.; GRIES, S. T. *Corpus-based Approaches to Metaphor and Metonymy*. Berlin; New York: M. de Gruyter, 2006. vi, 319 p.
- TOGNINI-BONELLI, E. *Corpus Linguistics at Work*. Amsterdam; Atlanta, GA: John Benjamins, 2001.