# Exploiting a Semantic Annotation Tool for Metaphor Analysis

Andrew Hardie,[1] Veronika Koller,[1]
Paul Rayson[2] and Elena Semino[1]

## 1. Introduction

In this paper, we describe the development and exploitation of a corpus-based tool for the identification of metaphorical patterns in large datasets.

The analysis of metaphor as a cognitive and cultural, rather than solely linguistic, phenomenon has become central as metaphor researchers working within 'Cognitive Metaphor Theory' have drawn attention to the presence of systematic and pervasive conventional metaphorical patterns in 'ordinary' language (e.g. *I'm at a crossroads in my life*). Cognitive Metaphor Theory suggests that these linguistic patterns reflect the existence of conventional *conceptual* metaphors, namely systematic cross-domain correspondences in conceptual structure (e.g. LIFE IS A JOURNEY). This theoretical approach, described further in section 2, has led to considerable advances in our understanding of metaphor both as a linguistic device and a cognitive model, and to our awareness of its role in many different genres and discourses. Although some recent research has incorporated corpus linguistic techniques into this framework for the analysis of metaphor, to date, such analyses have primarily involved the concordancing of pre-selected search strings (e.g. Deignan 2005). The method described in this paper represents an attempt to extend the limits of this form of analysis.

In our approach, we have applied an existing semantic field annotation tool (USAS) developed at Lancaster University to aid metaphor researchers in searching corpora. We are able to filter all possible candidate semantic fields proposed by USAS to assist in finding possible 'source' (e.g. JOURNEY) and 'target' (e.g. LIFE) domains, and we can then go on to consider the potential metaphoricity of the expressions included under each possible source domain. This method thus enables us to identify open-ended sets of metaphorical expressions, which are not limited to pre-determined search strings.

In section 3, we present this emerging methodology for the computer-assisted analysis of metaphorical patterns in discourse. The semantic fields automatically annotated by USAS can be seen as roughly corresponding to the domains of metaphor theory. We have used USAS in combination with key word and domain techniques in Wmatrix (Rayson, 2003) to replicate earlier manual analyses, e.g. machine metaphors in Ken Kesey's *One Flew Over the Cuckoo's Nest* (Semino and Swindlehurst, 1996) and war, machine and organism metaphors in business magazines (Koller, 2004a). These studies are described in section 4.

[1] Department of Linguistics and English Language, Lancaster University
*e-mail*: {a.hardie, v.koller, e.semino}@lancaster.ac.uk
[2] Computing Department, Lancaster University
*e-mail*: p.rayson@lancaster.ac.uk

## 2. Background

Cognitive Metaphor Theory (hereafter CMT), which is currently the dominant paradigm in metaphor research, claims that metaphor is a central and indispensable structure of both language and thought (e.g. Lakoff and Johnson 1980; Kövecses 2002). This claim is based on the observation that language use is characterised by the presence of systematic patterns of conventional metaphorical expressions, which are seen as realizations of conventional patterns of thought, or conceptual metaphors. For example, expressions such as *I've never won an argument with him* and *His criticisms were right on target* are regarded as linguistic realizations of the conventional conceptual metaphor ARGUMENT IS WAR (Lakoff and Johnson 1980: 4). Within this metaphor, WAR is the 'source' conceptual domain, and ARGUMENT the 'target' conceptual domain. This means that our understanding of arguments is systematically structured in terms of our knowledge of war: participants in arguments correspond to enemies, the outcomes of arguments correspond to winning or losing, the expression of one's ideas corresponds to attacking or defending, planning one's contributions corresponds to military strategy, and so on.

CMT has led to considerable advances in our understanding of metaphor both as a linguistic and a cognitive phenomenon, and has influenced a number of important studies of the role of metaphor in politics (e.g. Chilton 1996), education (e.g. Cameron 2003), science (e.g. Brown 2003), business media discourse (e.g. Koller 2004b), etc. In recent years, however, many scholars have started to question the prevalent use of constructed examples in CMT (and metaphor research more generally), and have shown how the systematic analysis of metaphorical patterns in naturally occurring data raises further questions and leads to further insights into metaphor as a linguistic and cognitive phenomenon (e.g. Cameron 2003, Deignan 2005). Some researchers have also started to use electronic corpora for the investigation of metaphorical patterns (e.g. Charteris-Black 2004, Koller 2004b, Semino 2005). Deignan (2005), in particular, has used concordances to show how corpus data broadly supports the claims of CMT, but also reveals linguistic phenomena that cannot be fully predicted or explained by the theory in its current form (e.g. differences in the metaphorical use of different parts of speech). The exploitation of electronic corpora for metaphor research is restricted, however, by the fact that the identification of metaphorical expressions in texts has not been successfully automated, despite some promising attempts (e.g. Fass 1991, Mason 2004). Hence, existing studies have relied on the traditional 'manual' analysis of data or on the concordancing of selected metaphorical expressions in larger data sets, or on a combination of these two methods. For example, Skorczynska and Deignan (2006) have discovered some important differences in the metaphors used in scientific vs. popular business discourse by (a) carrying out a manual analysis of small data samples and (b) concordancing the metaphorical expressions thus identified in larger corpora. However, this methodology only allows researchers to find further instances of previously identified expressions.

## 3. Methodology

The basic method in our studies is to apply corpus annotation techniques to the problem of extracting candidate metaphorical expressions from the corpus. In particular, we aim to demonstrate the usefulness of an approach based on semantic

field annotation for an analysis of text based on Cognitive Metaphor Theory, as summarised above. The semantic annotation we employ is that produced by the USAS tool (Rayson et al. 2004), one of several linguistic annotation tools that have been embedded in the Wmatrix system (Rayson 2003, 2007). This section briefly explains the workings of these tools.

USAS (*UCREL Semantic Annotation System*)[3] is a framework for automatic semantic annotation of text that has been developed at Lancaster University over a period of some seventeen years. It consists of a semantic tagset embodying a general language ontology; a large semantic lexicon listing potential tags for words and/or multi-word expressions; and software that uses the lexicon to apply semantic tags to running text. The semantic tagset was originally loosely based on the Longman Lexicon of Contemporary English (McArthur 1981). It distinguishes 21 major semantic fields (see Table 1), and over 200 subdivisions, some examples of which are given here:

```
S1.2     Personality traits
S1.2.1   Approachability and Friendliness
S1.2.2   Avarice
S2       People
S2.1     People:- Female
S2.2     People:- Male
```

| | | | |
|---|---|---|---|
| **A** General and abstract terms | **B** The body and the individual | **C** Arts and crafts | **E** Emotion |
| **F** Food and farming | **G** Government and the public domain | **H** Architecture, buildings, houses and the home | **I** Money and commerce in industry |
| **K** Entertainment, sports and games | **L** Life and living things | **M** Movement, location, travel and transport | **N** Numbers and measurement |
| **O** Substances, materials, objects and equipment | **P** Education | **Q** Language and communication | **S** Social actions, states and processes |
| **T** Time | **W** The world and our environment | **X** Psychological actions, states and processes | **Y** Science and technology |
| **Z** Names and grammatical words | | | |

**Table 1**: Top-level semantic categories (domains) in the USAS tagset.

When USAS runs on a text, the output consists of each word with an associated string of semantic tags from this system, as in the example below:

```
now                    T1.1.2 Z4
Wales                  Z2 Z1mf
must                   S6+ A7+
do                     A1.1.1 G2.2- X9.2+ E3- N5+ G2.1% Z5
```

[3] http://www.comp.lancs.ac.uk/ucrel/usas/

```
precisely               A4.2+ A5.3+
the                     Z5
same                    A6.1+++
against                 Z5
New                     Z1mf[i24.2.1 Z3c[i24.2.1
Zealand                 Z1mf[i24.2.2 Z3c[i24.2.2 Z2
```

Of importance here is the ordering of the tags in each word's list. Unless USAS has good contextual evidence to change the order of the tags, the sequence is derived from that specified in the lexicon which reflects the likelihood of the candidate semantic fields in general English. The likelihood ordering in the USAS lexicon is manually derived from sources such as frequency-based dictionaries, and informed by past tagging experience and intuition. The first tag on each word's list is the one that is judged to be the main or most likely category for that word (the % sign, seen on *do* above, indicates a very unlikely analysis). Hence, Wmatrix selects the first tag from each word's list for frequency analysis and retrieval purposes. For more details of the disambiguation methods used in USAS, see Rayson et al. (2004).

Within the web-based Wmatrix system[4], a text can be automatically annotated by USAS and then be analysed in a number of ways. For example, it is possible to look at *key semantic categories*, also known as *key domains*. These are defined as an extension to keywords (see Scott, 2000), that is, they are semantic categories that occur significantly more frequently in the text being analysed than they do in some appropriate reference corpus. Wmatrix provides a list of key domains for each text it analyses, ordered by their statistical significance (calculated using log-likelihood). Wmatrix can also provide a frequency list for any particular semantic tag, to see what word forms within that category occur in the text. Finally, it is possible to run concordance searches in the text, not only for particular words, but also for all words that are tagged with a chosen semantic tag (in first position in the tag list). The Wmatrix interface integrates these functions, making it easy to move from key domain list to frequency list or concordance.

The easy accessibility of USAS annotation within Wmatrix made it a perfect tool to investigate the applicability of semantic annotation to metaphor analysis. However, some early experimentation suggested that the basic searches that Wmatrix allowed, which at that point only accessed the first semantic tag on each token's list, did not successfully capture all the items of potential interest. For example, *campaign* is a term from the source domain of WAR that is often used metaphorically (e.g. *political campaigns*, *advertising campaigns*). It is used metaphorically so often, in fact, that the top tag is a portmanteau that encodes 'Wanting/Speech acts', and the G3 tag ('Warfare') is at the end of the list[5]:

```
campaign                X7+/Q2.2 I2.2/Q2.2 G1.2/Q2.2 G3
```

Although this ordering reflects the observation that the 'Warfare' sense is less likely in general[6] it is not ideal for metaphor identification. When researching metaphor, at

[4] http://www.comp.lancs.ac.uk/ucrel/wmatrix/

[5] The meanings of tags in the lexicon entry for *campaign* are as follows: X7+ is 'Wanting', I2.2 is 'Business: selling', G1.2 is 'Politics', Q2.2 is 'Speech acts' and G3 is 'Warfare'.

[6] For example, in Collins English Dictionary (21st Century Edition) the "series of coordinated activities, such as public speaking and demonstrating, designed to achieve a social, political, or commercial goal" sense appears before the military sense.

least two domains are of interest for each linguistic item: the source domain and target domain. But only one of these can be the main tag assigned to a given token. We were at risk, then, of missing relevant lexical items. We dealt with this issue by building into Wmatrix an additional type of search, which we refer to as "broad-sweep" searching. This type of search takes into account the *full* set of tags on the list for each token in the text being analysed, without privileging the main semantic tag. Thus, for our analyses of the texts in the case studies that follow, we were able to use *broad lists* and *broad concordances* as well as the simple searches described above. The use of these tools is exemplified in the case studies that follow.

## 4. Case studies

In this paper, we use USAS to re-analyse two data sets for specific metaphors. The data are Ken Kesey's novel *One Flew Over the Cuckoo's Nest* on the one hand and a corpus of business magazine articles on the other. By looking at similar metaphors in both ─ MACHINE and ORGANISM metaphors for the novel, WAR, MACHINE and ORGANISM metaphors for the magazine articles ─, we aim to draw conclusions about how widely used metaphor source domains are employed in different genres. This comparison relates to frequency as much as it does to patterns such as word classes or intensification/attenuation.

### 4.1 Metaphors in Ken Kesey's *One Flew Over the Cuckoo's Nest*

This first case study is concerned with Ken Kesey's novel *One Flew Over the Cuckoo's Nest* (Kesey 1973). The novel was first published in 1962 and was later adapted for film. The story is told in the first person by 'Chief' Bromden, who is a patient in a mental hospital, and is described by most critics as a paranoid schizophrenic. Bromden, who is half white and half native American, is also a trained electrician and worked in the Army during WW2, until he had a mental breakdown during an air raid. In the novel he tells of his life in the hospital, where he lives in terror of the staff and pretends to be deaf and dumb in order to avoid interaction. However, the arrival of a new inmate, McMurphy, causes a dramatic change on the ward and in Bromden himself, who eventually decides to escape from the hospital.

In an earlier study, Semino and Swindlehurst (1996) discussed the use of metaphor in Bromden's narrative, and claimed that metaphorical patterns play a central role in projecting the narrator's idiosyncratic 'mind style', including his world view and cognitive habits. In particular, Semino and Swindlehurst claimed that Bromden makes frequent and striking use of imagery drawn from the source domain of MACHINERY, which is familiar and salient for him due to his professional and autobiographical experiences. For example, he describes the hospital as a *factory* where broken machines are fixed in order to be returned to society, and chronically ill patients as broken machines that cannot be repaired. More generally, Bromden regularly uses mechanical lexis to describe people (including, himself, other patients and staff), emotional reactions, life inside and outside the hospital, and the structure of society.

In some cases, it is clear from the co-text that Bromden's use of such imagery is intended as metaphorical, especially when similes are used (e.g. *eyes glittering out . . . like the hard glitter of radio tubes*; Kesey 1973: 1). In other cases, Bromden seems

to believe in the presence and operation of machines, which, as readers, we conclude do not exist. This applies, for example, to the *fog machine*, which, according to Bromden, is located in the hospital's walls, and which he thinks is periodically used to flood the ward with fog. As readers, we are likely to conclude that Bromden has periodic attacks of panic and disorientation, which he explains as a result of the operation of the fog machine. In cases such as this, the reference to machinery is literal for Bromden, but can potentially function as an apt metaphor for readers, once they have become aware of Bromden's unreliability as narrator. Many other references to machinery are ambiguous as to whether Bromden uses them literally or figuratively, but all contribute to Kesey's metaphorical presentation of the hospital (and society) as a dehumanising environment, in which mental patients are regularly treated with electro-shock therapy.

Semino and Swindlehurst argued that Bromden's paranoid and mechanistic mind style is projected, in part, via frequent and creative realisation of conventional conceptual metaphors such as PEOPLE ARE MACHINES and INSTITUTIONS ARE MACHINES. In addition, they claimed that Bromden's mental and emotional development throughout the novel corresponds to a decrease in the frequency of MACHINERY images. More specifically, they identified a turning point in the plot about half way through the novel, and claimed that Bromden's use of MACHINE metaphor declines from that point:

> By our count, 82 machinery images may be found in the 112 pages that precede this episode and only 27 such images may be found in the remaining 143 pages. This count translates into an average of just over 7 instances of the machinery metaphor every 10 pages in the first part of the novel and just under 2 for every 10 pages in the second half. (Semino and Swindlehurst 1996: 163)

They also pointed out that, in the second part of the novel, Bromden's increasing sense of self-worth and freedom are often conveyed via descriptions of machines malfunctioning or breaking down around him in the hospital.

In addition, Semino and Swindlehurst noted that Bromden makes frequent use of the conventional conceptual metaphor POWERFUL IS BIG. He routinely refers to the nurse who runs the ward as *the Big Nurse*, and describes emotional outbursts as increases in physical size. Although other characters describe him as a very tall man (six feet eight inches in height), he claims that he has *shrunk*, and only feels himself becoming bigger again under McMurphy's influence. Finally, Semino and Swindlehurst argued that Bromden's development also corresponds with an increased ability to remember his childhood experiences on the Columbia river, resulting in an increase in vocabulary to do with nature. They also noted that McMurphy often uses ANIMAL metaphors for people (e.g. he describes the patients as a *flock of dirty chickens*; Kesey 1973: 50), and that Bromden himself uses similar metaphors towards the end of the novel.

In order to replicate Semino and Swindlehurst's (1996) study, we created an electronic version of the text of Kesey's novel, and used the Wmatrix software to search for key semantic domains. We selected as reference corpus the Imaginative Writing section of the BNC sampler[7]. A total of 66 semantic domains were found to be overused with a log likelihood threshold of 6.63 or above, corresponding to 99 per cent significance. The most overused domain was 'Medicines and Medical

[7] 222,541 words from BNC Sampler written imaginative corpus.

Treatment', which is not surprising given that the novel is mostly set in a hospital. What is more relevant for the purposes of this paper is that the top eight overused domains included: 'Objects generally' (log likelihood: 547.45) and 'Size: Big' (log likelihood: 139.86). We will now consider each domain in turn.

The tag 'Objects generally' was allocated to a wide variety of expressions referring to physical entities, such as *thing*, *rope* and *footbath*. More specifically, the list of types that were subsumed under this domain includes expressions that Semino and Swindlehurst had classified under the MACHINERY source domain, such as *machine*, *mechanical* and *wires*. Hence, we did not analyse all 1,352 tokens that received the 'Objects generally' tag, but focused on the following set of expressions which relate to machinery in particular:

- *machine* (24 tokens)
- *machinery* (21 tokens)
- *clock* (17 tokens)
- *wires* (17 tokens)
- *machines* (17 tokens)
- *wire* (15 tokens)
- *tubes* (13 tokens)
- *tube* (10 tokens)
- *wheel* (12 tokens)
- *mechanical* (9 tokens)
- *wheels* (7 tokens)
- *dials* (6 tokens)

Taken together, these expressions account for 168 tokens that were classified under the 'Objects generally' domain. Of these, approximately half were literal references to machinery in the text world, while the other half consisted either of references to non-existent machines, or of metaphorical references to machinery. This provides some empirical support for the findings of Semino and Swindlehurst's analysis, and suggests that the output of the USAS component of Wmatrix can be usefully exploited for the identification of metaphorical expressions within semantic domains that are likely to function as metaphorical sources.

In order to test Semino and Swindlehurst's claims further, we also created electronic versions of the first and second halves of the novel. The cut-off point was the stretch of narrative that Semino and Swindlehurst had identified as the pivotal moment in the plot. The two halves are roughly equivalent in length: the novel as a whole includes 114,029 word tokens; the first half contains 50,262 (44 per cent); and the second half contains 63,767 (56 per cent). When each half was separately compared with the Imaginative Writing section of the BNC sampler, the 'Objects generally' semantic domain turned out to be the second most overused domain in both cases. This may be an indication that Bromden's vocabulary is primarily concrete, since, as Semino and Swindlehurst suggested, he seems to lack appropriate vocabulary for mental, emotional and social experiences. However, the types of expressions that had the largest number of tokens for each half varied considerably. In both halves, the types *thing* and *things* had the largest number of tokens. However, in the first half, they were followed by expressions that had a high frequency of non-literal uses, such as *machine*, *machines*, *machinery*, and *wires*. In contrast, in the second half, they were followed by expressions that were primarily used literally,

such as *pole*, *tile*, *bottle* and *cup*. We looked particularly at the distribution and uses of the expressions listed above in the two halves of the novel. We found that approximately two thirds of the non-literal uses occurred in the first half of the novel (19 vs. 69 expressions). This proportion is similar to the result of Semino and Swindlehurst's manual analysis, even though the figures are not easily comparable: Semino and Swindlehurst considered a larger variety of expressions, and did not use the word as their unit of analysis, but rather any group of metaphorical expressions which were used within a single sentence and evoked the same source domain.

Let us now turn to the 'Size: Big' semantic domain, which was the eighth most overused domain in the novel as a whole. Out of 359 tokens that were subsumed under this domain, 311 are instances of the adjective *big*. The vast majority of these occurrences are non-literal (e.g. *Big Nurse*, *This puffs him up*), and reflect Bromden's association of size with mental and physical power. The same applies to some of the other expressions included under this domain, such as *puff up* and *grew*. All this strongly confirms Semino and Swindlehurst's claim concerning the dominance of the conceptual metaphor POWERFUL IS BIG in Bromden's narrative.

Our results were less conclusive in relation to references to nature, and organisms in particular. We considered two USAS domains: 'Living beings (animals)' and 'Living beings (plants)'. The latter was overused in both parts of the novel (the log likelihood was 23.94 for the first half and 28.04 for the second half). The former had low log likelihood values in both halves (respectively, 2.00 and 0.94). However, the cut-off point that was used to create the two halves of the novel is not appropriate to study differences in references to living beings, since McMurphy enters the plot well before this point, and is responsible for many metaphorical references to animals in particular. Further work will therefore be needed to investigate any significant changes in relation to nature imagery in the novel.

Finally, we will consider WAR as a potential source domain in the novel, since this domain features in our second case study. Although WAR metaphors are generally regarded to be highly conventional, Bromden makes little use of them in his narrative. The 'Warfare' semantic domain is in fact overused in the novel as a whole (with 142 tokens and a log likelihood of 52.87). However, most occurrences are literal references to Bromden's experiences in WW2, which loom large in his mental life. The few metaphorical expressions that we could identify under this particular domain were highly conventional, such as the use of the verbs *attack* and *defend* in relation to verbal arguments. Indeed, Semino and Swindlehurst had not included this domain in their discussion of metaphorical patterns in the novel.

Having concluded our first case study, in the next section we revisit a second study that was also originally carried out manually. We consider MACHINE and ORGANISM metaphors as we did for the novel, but in the second dataset we find that WAR metaphors play a more important part.

### 4.2 War, machine and organism metaphors in the business magazine corpus

The second study uses as its empirical base a corpus of 40 different articles (81,378 words) on 20 different businesswomen, published between 1996 and 2001 in the business magazines and papers *Business Week* (US), *Economist* (UK), *Financial Times* (UK), *Forbes* (US) and *Fortune* (US). In the original study (Koller 2004a), this corpus was analysed manually, using a combination of quantitative and qualitative

methods to compare the metaphoric descriptions of businesswomen to those used for their male counterparts. The aim was to identify all metaphors used by journalists to describe businesswomen, and thus no particular word field relating to any particular metaphor was established in advance. As a consequence, the corpus was searched manually for relevant metaphoric expressions, including negations (e.g. *she is no shrinking violet*), similes (e.g. *this will be like turning a battleship around*) and phrases explicitly referring to metaphor on a meta-level (e.g. *she doesn't inspire lamb metaphors*). Deducing the underlying conceptual metaphor from the linguistic expressions was largely based on "informed intuition" (Deignan 1999: 180), although some parameters point towards the central nature of the metaphors in the discourse of business magazines (Low 1999: 164): there are creative extensions of the metaphor by the speaker/writer, as well as extensions of metaphors used by others. Moreover, we find explicit metaphoric statements of the A IS B variety, e.g. *she is a war machine*, while the aforementioned *she doesn't inspire lamb metaphors* explicitly discusses the semantic features that are (not) mapped. Finally, text producers challenge the metaphor use of others; thus, the explicit statement *she is a war machine*, made by a colleague of the executive in question, is questioned by the writer of the article, who modifies it to *If Brink's a machine, she's a machine that likes to live well*. After identifying central conceptual metaphors in this way, these were compared by looking at the number of metaphoric expressions (tokens) that each gives rise to. This step helps ascertain the conceptual weight of each metaphor.

The manual analysis yielded a total of 186 metaphoric expressions, realising 23 different conceptual metaphors. By far the most prominent among these is the WARRIOR/FIGHTER metaphor, being realised in 89 instances accounting for almost half of all metaphoric expression (47.85 per cent). The sheer dominance of this metaphor warrants its inclusion in the computer-assisted re-analysis of the corpus. Further down, we find businesswomen being conceptualised as (hard) objects (seven instances) or machines (one instance). In order to compare the results of the USAS-based analysis to those for the fictional texts (see previous section), it seems reasonable to also include any metaphors that relate to non-human living organisms as a source domain. This group is rather diverse, including (wild) animals (nine instances), flowers (three instances) and birds (two instances). A related metaphor is that of the manager as gardener, an entailment of the COMPANIES ARE LIVING ORGANISMS metaphor. In the manual analysis, this particular conceptualisation occurred three times. On the whole then, results of the USAS analysis were compared against the following findings: WAR metaphor 89 metaphoric tokens, MACHINE metaphor a combined eight tokens and (non-human) LIVING ORGANISM metaphor a total of 17 tokens.

In the current paper, we focus on the metaphoric references to businesswomen only, and limit our analysis to the three prominent metaphors of WAR, MACHINES and (NON-HUMAN) LIVING ORGANISMS, to allow for better comparisons with the literary genre analysed above. Replicating parts of the business magazine study involved rendering the data electronically readable and searching it for key semantic domains, where the BNC Sampler CG Business[8] served as the reference corpus within Wmatrix. With the log likelihood threshold set at 6.63 for 99 per cent significance, the search yielded 152 semantic domains as overused to a statistically significant degree in the magazine corpus. This finding points towards a thematic variety in the corpus

[8] 141,143 words from the British National Corpus Sampler Context Governed Business sub-corpus consisting of company and trades union talks or interviews; business meetings; sales demonstrations etc.

that is perhaps surprising in texts taken from special-interest publications. However, one has to not only remember that the articles are portraits of particular individuals who work for companies in very different industries but also that the portraits elaborate on the women's personal life and biography.

Obviously, 'Personal names' and 'People: female' are among the five most overused domains in relative terms. Crucially though, the domain of 'Warfare' ranks high up in the top ten percent at 15th (log likelihood 145.25), which corroborates the finding from the original study that businesswomen are described as metaphorical warriors and fighters proportionately more often than their male colleagues (for findings and discussion see Koller 2004a: 12). The WAR metaphor is itself a hybrid that combines the semantic component of physical aggression with a more attenuated constituent centring on strategy, the latter being reflected in the USAS semantic tag for 'Warfare' being a sub-domain of government. To account for the aspect of physical aggression, the semantic domain of 'Violent/aggressive' was also taken into account. This ranks in 24th place, at a log likelihood of 100.12. Together, the two domains show 289 tokens.

As the USAS tag set does not feature a domain 'Machines', the relevant semantic domain was identified as 'Objects', especially since the manual analysis also ascertained a metaphor source domain (HARD) OBJECTS. In the search results, the 'Objects' domain features at 57 (out of 152; log likelihood 35.42) and thus at a much lower level of significant overuse than the 'Warfare' and 'Violent/aggressive' domains. It does show a high number of tokens though, 459, which is undoubtedly due to the general nature of the category. Finally, the potential source domain of (NON-HUMAN) LIVING ORGANISMS is in the USAS tag set spread across 'Living creatures: animals' and 'Plants'. The former is to be found in 18th place (log likelihood 117.03), while the former ranks 63rd (log likelihood 32.4); together they account for 163 tokens.

In the next step, the search was extended to find all types that had a relevant tag somewhere in their list of tags, i.e. utilise the so-called broad-sweep search function of the version of Wmatrix adapted for metaphor analysis. In the resulting five word lists (for 'Warfare', 'Violent/aggressive', 'Objects', 'Living creatures: animals' and 'Plants'), the types were checked against their concordances to ascertain how often they were used metaphorically. This part of the analysis obviously and necessarily repeats the manual analysis, and is based on the same criteria for metaphor identification (see above). This final step limited the number of relevant tokens, in some cases considerably so.

Obviously, particular domains are better candidates for metaphor source domains than others, depending on genre and topic of the texts under investigation. Further, the rationale for using semantic tagging for metaphor analysis in the first place is the assumption that the source domains of highly conventional metaphors will probably be allocated a target domain tag, while words realising more novel metaphors are likely to receive a source domain tag. Both considerations come into play when looking at the results for broad-sweep searches. In business magazines, 'Warfare' and 'Violent/aggressive' are good candidates for metaphor source domains, so it comes as no surprise that two thirds of all tokens with one or both of the two tags are actually metaphorically used. Also, the most common metaphoric expressions of war, which can be said to have achieved the status of technical terms, are tagged for their target domain, with the source domain as a secondary or further tag. Examples include *target(s)*, *strategic*, *flak* and *forays*, all of which are used metaphorically only. However, concordance searches also reveal that quite a few types that are allocated a

source domain tag and could therefore be assumed to be unconventional in the genre, are actually used only, or predominantly, metaphorically as well, e.g. *war/-s, -ring*, *weapon*, *swords*, *bomb*s and *guns*. This means that the language to describe female executives in business magazines is even more belligerent than the USAS tag set gives them credit for.

Given the general nature of the category 'Objects', it is unsurprising to find the concordances reveal that only 5.88 per cent of its 459 overall tokens are actually used metaphorically. In further contrast to 'Warfare' and 'Violent/aggressive', most of the metaphorically used tokens are allocated a source domain tag 'Objects' or the related tag 'Material'. This indicates that expressions from these semantic domains have not entered business magazine discourse as technical terms to the extent that expressions relating to war and physical aggression have. This throws an interesting light on the main mental models underlying business (magazine) discourse and possibly social practices and behaviour in the related social field. Looking at 'Living creatures: animals' and 'Plants' corroborates and intensifies a trend already observed with 'Objects', namely, that it is often the company rather than its female executive who is metaphorised in this way. By way of entailment, the executive then becomes a carer, nurturer or, more specifically, gardener, and indeed these source domains were identified as central in the manual analysis. (Incidentally, they are as gendered as the WARRIOR/FIGHTER metaphor, only more conventionally so.) With the exception of the highly conventional 'growth' and 'branch(es)' – relating to companies – all types used only or mostly metaphorically, were allocated the source domain tags. The conversion rate from all tokens to metaphorically used tokens was 14.72 per cent. Table 2 summarises the findings in comparison to the manual analysis.

| **domain** | **all tokens** | **metaphoric tokens (% of all tokens)** | **tokens manual analysis** |
|---|---|---|---|
| 'War', 'Violent/aggressive' | 289 | 191 (66.01%) | 89 |
| 'Objects' | 459 | 27 (5.88%) | 8 |
| 'Living organisms: animals', 'Plants' | 163 | 34 (14.72%) | 17 |

**Table 2**: Comparison metaphoric tokens USAS and manual analysis

The above shows that utilising a semantic tagger for metaphor analysis yields two to three times more results than manual analysis. While the computer-assisted method still requires extensive manual reworking, this is no more time-consuming than an exclusively manual analysis and it generates more comprehensive results.

## 5. Conclusion

In this paper, we have described the use of an existing semantic annotation tool (USAS) and corpus retrieval software (Wmatrix) for metaphor analysis. We have discussed two earlier manual analyses and the application of the tools to the original data sets. We were able to replicate the manual analyses and we also showed that the

automatic techniques highlighted further concordance examples for manual checking. Our analyses also begin to suggest that metaphorical source domains such as MACHINES, ORGANISMS and WAR differ in how and to what extent they are exploited in different texts and genres. For example, WAR metaphors are much more prominent in the business magazine data than in Kesey's novel. There is also variation in terms of what exactly is being metaphorically constructed via mappings from particular source domains, as in the case, for example, of ORGANISM metaphors in our two data sets. However, more comparative work of this kind is needed to corroborate these tendencies.

Using a combination of automatic semantic analysis, key word and domain techniques, we have begun to develop a new methodology for the computer-assisted analysis of metaphorical patterns in discourse. Although the existing tools were not developed with the metaphor researcher in mind we have shown how they might be used to assist in the extension of metaphor analysis in terms of scalability and coverage for larger datasets. In the next stage of our research, we intend to apply our emerging methodology to explore patterns of conventional metaphorical expressions in a much larger corpus extracted from the BNC. We intend to use manual analysis of metaphorical patterns in discourse to train the tools to provide large-scale computer-assisted analysis. Our research would have implications for automatic semantic analysis and word sense disambiguation in the area of computational linguistics. In addition, we hope to be able to explore the differences and similarities that can be identified in dominant metaphorical patters across the genres included in such a corpus, and therefore explore implications of these findings for existing claims in metaphor theory.

## Acknowledgements

Our research has been funded by grants from Lancaster University's Faculty of Social Sciences and the British Academy (Small Research Grant SG-42813).

## References

Brown, T. L. (2003). *Making truth: metaphor in science.* Urbana, Ill.: Urbana and Chicago: University of Illinois Press.

Cameron, L. (2003) *Metaphor in Educational Discourse*. London: Continuum International Publishing.

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

Chilton, P. (1996) *Security Metaphors: From Containment to Common House*. New York: Peter Lang.

Deignan, A. (2005) *Metaphor and Corpus Linguistics*. Amsterdam: John Benjamins.

Fass, D. (1991). met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17, 1, 49-90.

Kesey, K. (1973) *One Flew Over the Cuckoo's Nest*. London: Pan.

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

Koller, V. (2004b). *Metaphor and Gender in Business Media Discourse: a Critical Cognitive Study*. Basingstoke: Palgrave.

Kövecses, Z. (2002). *Metaphor: A Practical Introduction*. Oxford and New York: Oxford University Press.

Lakoff, G. and Johnson, M. (1980) *Metaphors We Live By*. Chicago: Chicago University Press.

McArthur, T. (1981). *Longman Lexicon of Contemporary English*. London: Longman.

Mason, Z. (2004) CorMet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30, 1, 23-44.

Rayson, P. (2003). Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. *Ph.D. thesis*, Lancaster University.

Rayson, P. (2007) *Wmatrix: a web-based corpus processing environment*. Computing Department, Lancaster University.

Rayson, P., Archer, D., Piao, S. L., McEnery, T. (2004). The UCREL semantic analysis system. In proceedings of the workshop on *Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with LREC 2004*, Lisbon, Portugal, pp. 7-12.

Scott, M. (2000), Focusing on the text and its key words, in L. Burnard and T. McEnery (eds), *Rethinking Language Pedagogy from a Corpus Perspective, Volume 2*. Frankfurt: Peter Lang., pp. 103-122.

Semino, E. (2005) The metaphorical construction of complex domains: the case of speech activity in English, *Metaphor and Symbol*, 20, 1, 35-69.

Semino, E. and Swindlehurst, K. (1996) Metaphor and mind style in Ken Kesey's *One Flew Over the Cuckoo's Nest*, *Style*, 30, 1, 143-166.

Skorczynska, H. and Deignan, A. (2006) Readership and purpose in the choice of economics metaphors, *Metaphor and Symbol*, 21, 2, 87-104.