Using a semantic annotation tool for the analysis of metaphor in discourse

Veronika Koller/Andrew Hardie/Paul Rayson/Elena Semino, Lancaster University
(Corresponding author: v.koller@lancs.ac.uk)

Abstract

This paper describes the application of semantic annotation software for analysing metaphor in corpora of different genres. In particular, we outline three projects analysing RELIGION and POLITICS metaphors in corporate mission statements, the WAR metaphor in business magazines, and MACHINE and LIVING ORGANISM metaphors in a novel and in a second collection of business magazine articles. This research was guided by the hypotheses that a) semantic tags allocated by the software can correspond to source domains of metaphoric expressions, and b) that more conventional metaphors feature a source domain tag as first choice in the type's semantic profile. The tagger was adapted to better serve the needs of metaphor research and automate to a greater extent the extraction of first choice and secondary semantic domains. Two of the three studies represent re-analyses of previous manual and/or lexical corpus-based investigations, and findings indicate that semantic annotation can yield more comprehensive results.

In diesem Artikel beschreiben wir die Anwendung eines semantischen Annotationsprogramms, um Metaphern in Korpora verschiedener Genres zu analysieren. Im Einzelnen skizzieren wir drei Projekte zu religiösen und politischen Meatphern in Firmenleitbildern, zur Kriegsmetapher in Wirtschaftsmagazinen und zu Maschinen- und Organismusmetaphern in einem Roman sowie einem zweiten Korpus von Artikel aus Wirtschaftsmagazinen. Unsere Forschung beruhte auf den Hypothesen, dass a) die von dem Programm zugewiesenen semantischen Kennzeichnungen den Quelldomänen metaphorischer Ausdrücke entsprechen können und b) konventionelle Metaphern eine Quelldomänen-Kennzeichnung an erster Stelle im semantischen Profil des entsprechenden Wortes zugewiesen bekommen. Das Programm wurde für die Bedürfnisse der Metaphernforschung adaptiert, in dem die Analyse von erstgereihten und sekundären Kennzeichnungen in einem größeren Ausmaß automatisiert wurde. Zwei der drei Studien stellen erneute Analysen früherer manueller und/oder lexikalischer korpus-basierter Untersuchungen dar, und die Ergebnisse zeigen, dass sich mit semantischer Annotation umfassendere Resultate erzielen lassen.

1. Introduction

In this paper, we investigate the use that can be made of semantic annotation in the analysis of metaphoric patterns across large bodies of text. Within the framework of conceptual metaphor theory, we will show that semantic domain tagging can be used to identify instances of metaphoric expressions in a lengthy text or corpus with a greater recall than other methods that have

previously been used. As a result, we have found it possible to expand on the results of earlier analyses of metaphor in discourse that were undertaken without the aid of a semantic tagging system.

The paper is structured as follows. In section 2, we outline the rationale for the use of corpus data in the investigation of conceptual metaphors, and discuss some previous work in this area. We will show that such work has typically relied on concordances of pre-selected search terms — and that by use of semantic tagging we can obviate the need for such a list of search terms. Section 3 gives an overview of the USAS tagger which we used to annotate our data, and the Wmatrix tool in which it is embedded. In section 4, we outline three studies of metaphor in discourse that we have undertaken to demonstrate the utility of the approach presented in this paper. Studies of religion and politics metaphors in corporate mission statements, war metaphors in business magazines, and machine and living organism metaphors in a different collection of business magazine articles and also in Ken Kesey's novel One Flew Over the Cuckoo's Nest, all demonstrate that it is possible to retrieve a more comprehensive set of metaphoric expressions via searches based on semantic annotation than by searching for a manually-selected set of lexical items. Finally, in our evaluation of our work to date we present some ideas for future analysis of metaphor in discourse using methods based on semantic annotation.

2. Previous work in corpus-based metaphor analysis and the case for semantic annotation

Critics of conceptual metaphor theory have argued that, while powerful theoretically, the framework has lacked an empirical basis (Steen 1999, Cameron 2003, Low 2003, Semino, Heywood and Short 2004, Deignan 2005). To remedy this situation, a number of researchers have harnessed corpus-based analysis to investigate metaphor in naturally occurring language. The decision for large-scale data sets rather than smaller data samples can be justified not only in terms of validity of results, but also along theoretical lines: If one accepts the claim that metaphor is at least partly constituted intertextually (Eubanks 2005), what is needed is an investigation into particular metaphors which are shared by, and distributed and reinforced among, members of a discourse community. Therefore, rather than restricting

the empirical application of conceptual metaphor theory to single data sets, corpus-based metaphor analysis seeks to ascertain metaphor usage in larger text collections that represent the various voices of a discourse community.

To date, most of this work has taken a lexical approach of some description and analysed concordances of pre-defined search strings of single words or members of word fields. Although different in terms of research questions, methodological detail and incorporation of a critical perspective, most studies (Charteris-Black 2004, Deignan 2005, Koller 2004a, Musolff 2004, Stefanowitsch 2005, Skorczynska and Deignan 2006) have combined manual and computer-assisted analysis to investigate the use of metaphor in large data sets. For example, Deignan (2005: 174-183) has studied plant metaphors in English by concordancing a range of relevant expressions in the Bank of English corpus (e.g. plant, seed, harvest) and then manually analysing a section of each concordance in order to identify metaphoric instances of each expression. In several other studies researchers have identified metaphoric expressions in a subset of their data via detailed manual analysis, and then concordanced those expressions in the rest of the data (e.g. Semino 2002, Cameron and Deignan 2003, Charteris-Black 2004, Koller 2004a, Koller and Semino forthcoming, Semino and Koller forthcoming). An inevitable limitation that is shared by both approaches is the fact that the automatic analysis is restricted to pre-determined search strings. This means that further tokens of particular types can be automatically retrieved, but new metaphoric types cannot be identified, unless they happen to occur in close proximity to node expressions.

The semantic approach that we advocate in this paper also relies on some manual input and analysis. Firstly, our method is based on the categorisation of all lexical items in a corpus according to a lexicon that underlies the semantic annotation tool described in section 4. This lexicon was generated and is constantly being extended manually, leading to occasionally idiosyncratic results. Therefore, automatically generated results also need to be checked manually. While all corpus-based language analysis, including automated semantic annotation, helps to check the researcher's intuition, the underlying lexicon still relies on intuition to some extent. Nevertheless, we believe that a semantic annotation approach represents a valuable additional tool for researchers interested in analysing the use and function of metaphor in naturally occurring discourse on a large scale.

The string of projects that we detail in sections 4.1-4.3 is informed by the hypothesis that the semantic tags which the annotation software allocates to words in a corpus correspond to the source and/or targets domain of metaphorically used types. Crucially, our approach moves beyond pre-determined search strings and thus potentially allows for the investigation of open-ended sets of metaphoric expressions in large data collections.

3. Features of the USAS tagger

The software that we used in our work is called USAS, short for UCREL[1] Semantic Annotation System. It forms a component of Wmatrix, a web-based system for automated text annotation, which was developed by Paul Rayson.[2] It should be noted that Wmatrix was not at all designed with a view to metaphor analysis; rather it is a suite of programs for the quantitative analysis of text with the help of frequency lists, concordances and keyness indicators for words (including multi-word entities such as proper names, compound nouns and phrasal verbs), parts of speech and, most important for our purposes, semantic domains.[3] USAS assigns semantic domain tags, which are pre-defined in the underlying lexicon, to the types in a corpus; and Wmatrix compares the frequency of the tags thus allocated to a large external reference corpus — subsets of the British National Corpus (100 million words) — to ascertain tag keyness via statistical significance. Significance is measured using log likelihood, with a threshold value of 6.63 for $p<0.01$. While part-of speech tagging shows accuracy rates of 96-97 per cent, semantic tagging with USAS guarantees an accuracy of 91-92 per cent (Rayson et al. 2004). Our first hypothesis, and rationale for applying USAS, is that the semantic domains which are allocated can correspond to the source domains of metaphoric expressions.

Importantly, USAS allocates both first choice and secondary tags, using a hybrid approach to decide on the order of tags and thus to build a semantic

---

1 UCREL is in turn the acronym for the University Centre for Computer Corpus Research on Language, which is based at Lancaster University (UK).

2 Wmatrix, the web interface of which the USAS semantic tagger is a part, is available at http://ucrel.lancs.ac.uk/wmatrix/.

3 Although not the focus of this paper, it should be mentioned that we also investigated the (rather complex) link between part-of-speech and metaphoricity (see also Deignan 2005).

profile for each type. A combination of general likelihood ranking (derived from corpus and dictionary evidence), disambiguation by part-of-speech, participation in multiword expressions and topic information is employed in this disambiguation process. The resulting ordering of tags is designed to give the first choice tag as the most likely tag in a given context. Figure 1 shows the tag strings for lemmas of *campaign* that occur in a corpus of business magazine texts (see section 4.2). For reasons outlined below, we were particularly interested in the position of the G3 tag ("warfare") in the tag strings.

type	**PoS tag**	**semantic tag** (G3 "warfare")
campaign	NN1	X7+/Q2.2 I2.2/Q2.2 G1.2/Q2.2 G3
campaign	VV0	X7+/Q2.2 I2.2/Q2.2 G1.2/Q2.2 G3
campaignable	JJ	X7+/Q2.2 I2.2/Q2.2 G1.2/Q2.2
campaigned	VVD	X7+/Q2.2 I2.2/Q2.2 G1.2/Q2.2 G3
campaigner	NN1	X7+/Q2.2/S2mf I2.2/Q2.2/S2mf G1.2/Q2.2/S2mf
campaigners	NN2	X7+/Q2.2/S2mf I2.2/Q2.2/S2mf G1.2/Q2.2/S2mf
campaigning	This is listed neither under VVG nor under JJ	X7+/Q2.2 I2.2/Q2.2 G1.2/Q2.2 G3
campaigning	NN1	X7+/Q2.2 I2.2/Q2.2 G1.2/Q2.2 G3
cyber-campaigns	NN2	I2.2/Y2
e-campaign	NN1	G3/Y2

Figure 1: First and secondary tags for *campaign*

*** The verb form *attack* has a first tag Q2.2 (speech acts) rather than G3 (warfare), reflecting the ARGUMENT IS WAR metaphor.**

This feature of the program gives rise to our second main hypothesis, namely that secondary tagging indicates conventionality of metaphoric expression. To give an example, we started from the assumption that the war metaphor, which is a typical feature of business magazine texts (see Koller 2004a), manifests in linguistic expressions that are more or less conventional. Thus, marketing campaign would be highly conventional in contrast to coinages

such as efficiency jihad. If our hypothesis is correct, the source domain of war would feature as a first choice or even only tag in novel or less conventional metaphoric expressions, but as a secondary tag ranked behind a target domain tag in conventional expressions, where the metaphoric meaning is established as predominant. Indeed, as can be seen in Figure 1, campaign as both a verb and noun shows a first choice tag X7 ("wanting, planning, choosing"), which represents the target domain, while the source domain is represented by the last tag in the string, G3 ("warfare"). The extension e-campaign, on the other hand, has been allocated only the source domain tag, together with the secondary tag Y2 ("information technology and computing"). Figure 2 shows the relevant results for the whole of the business magazine corpus.

Expression	**G3 tag occurs in position x**	**Out of x tags**
battlefield	1	1
combat (noun)	1	1
cyberwars	1	1
e-campaign	1	1
gun	1	1
infantry	1	1
legion	1	1
legions	1	1
soldier	1	1
shotgun	1	1
warfare	1	1
warlike	1	1
warrior	1	1
warriors	1	1
war zone	1	1
weaponry	1	1
armed	1	2
army	1	2
barrage (noun)	1	2
battleground	1	2
blitz (noun)	1	2
bomb (noun)n	1	2
embattled	1	2
repel (verb)	1	2
spear (noun)	1	2

squads	1	2
veteran	1	2
war	1	2
warring	1	2
wars	1	2
weapon	1	2
armour	1	3
bomb (verb)	1	3
enlisted	1	3
gun (noun)	1	3
invade	1	3
invading	1	3
invasion	1	3
troops	1	3
bombard	1	4
bombarded	1	4
retreat	1	4
forces	1	6
shoot (verb)	1	6
assaulting	2	2
beleaguered	2	2
drafted	2	2
entrenched	2	2
fallout	2	2
front line	2	2
mobilized	2	2
mobilizing	2	2
recruits (noun)	2	2
strategic	2	2
trench	2	2
agents	2	3
battles	2	3
battling	2	3
conquer	2	3
conqueror	2	3
conquest	2	3
target (noun)	2	3
targets (noun)	2	3
battle	2	4

attacks*	2	5
foray	3	3
forays	3	3
maneuver	3	3
sign on	3	3
repel (verb)	3	4
campaign	4	4
campaigns	4	4
exploded	4	4
shoot (noun)	4	4
sights	4	4
field	5	6
barrage (verb)	-	1
blitz (verb)	-	1
blood	-	1
bombshell	-	1
brutal	-	1
brutality	-	1
combat (verb)	-	1
cut-throat	-	1
defeat	-	1
kill (noun)	-	1
radar	-	1
recruits (verb)	-	1
retrench	-	1
tactic	-	1
target (verb)	-	1
targets (verb)	-	1
victorious	-	1
backfire	-	2
bloody	-	2
bruise	-	2
enemy	-	2
spear (verb)	-	2
surrender	-	2
survival	-	2
survive	-	2
survivor	-	2
victory	-	2

bleed	-	3
campaigner	-	3
casualty	-	3
launch	-	3
fight	-	4
fighting	-	4
kill (verb)	-	4
killer	-	4
killing	-	4

Figure 2: Conventionality of metaphoric expression and ranking of tags in business magazine corpus (G3 "warfare")

After outlining the basic mechanisms and affordances of the USAS tagger, we will in the following give an overview of three small-scale research projects in which we have used and adapted the software.

4. Using the USAS tagger for metaphor identification: projects and progress

Starting in 2005, we have to date used automated semantic annotation for metaphor analysis in three separate projects involving four different data sets. Based on the results, the software was adapted for metaphor analysis throughout.

4.1 RELIGION and POLITICS metaphors in mission statements

Given that Wmatrix and its components were not developed for metaphor analysis, we could at first use it only for basic searches of first choice tags and their associated word lists and concordances, i.e. ascertain all the types in a corpus that had a specific first choice tag and look at them in their co-text. At that early stage, secondary tags could only to be identified by going back to the lexicon. As part of a larger research project on religious and political metaphors in corporate discourse (Koller forthcoming), we worked with a 30,000-word corpus of corporate mission statements, which was built from the websites of the top and bottom 50 of the 2003 Fortune Global 500 companies. In the analysis, we first calculated which semantic domains were key when compared to the reference corpus (here, the one million word Written section of the BNC sampler). Given that all texts in the corpus represent a particular

genre that is linked to corporate discourse, even this first result can indicate the possible occurrence of metaphor by bringing up unexpected semantic domains. For instance, the domain L1 ("life and living things") is much less expected to be key in a collection of corporate texts than is the domain I1.1 ("money and pay"), which is a very strong key domain with a log likelihood of 201.78. The fact that "life and living things" does feature as a key domain (log likelihood 39.14) suggest that the associated lexemes may be used metaphorically. The type to get L1 as a first choice tag is organic, and the associated concordance line proves the assumption correct:

(1) The AXA Group is focused on two strategic priorities: Strengthen the Group's positions in the most developed or high potential markets in Western Europe, North America, and selected countries in Asia Pacific. Achieve operational excellence in each market by leveraging *organic growth*, technical and investment margins, quality and productivity.

In the first project, we decided to start from either the target or the source domain to ascertain metaphoric expressions. However, a search for the target domains "business" and "work and employment" yielded no valuable results. We therefore proceeded to look at the source domains we were interested in ("religion and the supernatural", "government and politics"). That part of the analysis was carried out on three levels. Since the relevant domains did not appear as key tags, we began by checking the semantic tag set for the whole corpus to see which words had a first choice tag for either of the domains. We also wanted to include words where the tag of interest was a secondary tag. We could not use the Wmatrix semantic tag frequency list to identify these, but rather had to manually examine the underlying annotated text to extract these items. Finally, to catch any words that had no relevant tag but might still be relevant metaphoric expressions, we went to the word list for the whole corpus, identified possible candidates for metaphoric usage and checked the associated concordance lines. Figure 3 lists all the types in the mission statement corpus that are associated with "religion and the supernatural" or with "government and politics", either by having been allocated a relevant tag somewhere in the tag string, or by having been identified as metaphoric expressions despite the lack of a source domain tag.

Figure 3 (below): Source domains and types in the mission statement corpus

NB: Numbers in brackets show frequency of tokens. Key: relevant domain as first tag, *relevant domain as secondary tag*, not tagged for relevant domain, item used metaphorically.

Religion and the supernatural

acts (1)
belief/beliefs/believe/believed/believes (6/6/18/1/2)
call/calls (4/3)
celebrate (1)
cell/cells (1/1)
chapter/chapters (2/1)
communicate/communicates (6/1)
converted (1)
credo (5)
creed (3)
deliver/delivered/delivering/delivers/delivery (35/3/11/4/3)
Elf (1) (NB: proper name)
faith (2)
father (1)
font (2)
Fortune (2) (NB: proper name)
high/higher/highest (18/6/26)
host (3)
John (1)
living (2)
mass (1)
mediums (1)
mission (45)
new age (1)
office/offices (3/8)
passion (8)
possess (2)
presence (8)
religion/religious (2/1)
sacred (1)
save/saved/saving/saves (2/1/1/1)
service/services (45/103)
shrine (1)
soul (2)
spirit(s)/spiritually (21/3/1)
vision (35)
wizard (2) (NB: irrelevant technical term)
word (5)

Government/Politics

agencies/agency (4/1)
approval/approved (1/3)
assessment (1)
authority (1)
? *benefit(s)* (16/17)
blue (1)
budget (1)
bureaucratic/bureaucracy (1/1)
candidates (2)
capital (8)
care (6)
center(s)/centre(s) (6/5/1/1)

citizen(s)/citizenship (10/4/3) (NB: half of instances of 'citizens' literal)
civil (2)
commission (1)
constituents (2)
corporation(s) (21/5)
council (1)
country (7)
demonstrate/demonstrated/demonstrates/ demonstrating (2/5/3/3)
deregulation (1)
division(s) (1/4)
duties (2)
establishment (1)
executive(s) (3/3)
fiscal (5)
governance (5)
government(s) (6/3)
green (4) (NB: three instances of 'green' literal)
March (1) (NB: proper name)
municipal (2)
nation(s)/nationality (5/2/4)
non-government (1)
officer (3)
official(s) (2/1)
parties (2)
politics/political (1/1)
president(s) (2/1)
private (8)
public sector (1)
registration (1)
regulations (3)
return/returns (6/4)
revenue(s) (2/5)
rights (1)
run/running (2/1)
service (45)
stand/standing/stands (4/1/2)
states (5)
tap (1) (NB: POS-tagged as verb, domain assigned erroneously to noun)
welfare (2)

As shown in Figure 3, not all types that had been allocated a source domain tag were used metaphorically. Rather, words in all three groups — those with first choice, secondary or no relevant tags — needed to be linked back to their concordances to check for metaphoricity. Indeed, some words with a relevant first choice tag were used metaphorically while others were not (e.g. sacred and constituents vs. religion and council), and the same held true for words with relevant secondary tags (e.g. faith and officer vs. father and parties). Words without relevant tags, which were identified by checking the word list for the whole corpus, were all used metaphorically, but not for each token (e.g. only half of all instances of citizen occurred in the metaphoric collocation corporate citizen).

Obviously, this procedure was very time-consuming and cumbersome. In the follow-up project, we therefore adapted the USAS tagger to better meet the needs of metaphor analysis.

4.2 The WAR metaphor in business magazines

We proceeded with a small-scale project funded by Lancaster University Faculty of Arts and Social Sciences in which we re-analysed a corpus of 34 articles (ca. 41,000 words) on marketing and sales in Business Week magazine (see Koller 2004a: 64-113). Our focus was on realisations of the marketing is war metaphor. Based on insights gained in the first project, we did not expect a target domain search to yield any results and at the same time realised that a search for source domains would have to be automated to a greater extent than previously possible. To this end, we augmented the USAS tagger with what we called a domain push function, which makes it possible to find highly conventional metaphoric expressions tending to carry the source domain as a secondary tag. In order to capture all tags of the relevant domains, the domain push function changes the ranking of tags so as to list in first position tags for a particular domain that has been pre-specified as relevant. This eliminates the need to check the lexicon for the complete tag string for any given type.

The structure of the semantic tag set resulted in concordances that still contained many unwanted results which needed to be filtered out manually. Nevertheless, adding the domain push function yielded results that were not obvious at the start of the analysis. For example, we found that companies are frequently conceptualised as metaphorical plants, whereas brands are often conceptualised as race horses in business magazine articles on marketing and sales. This differentiated finding was not detected by the lexical corpus-based approach of the first analysis (Koller 2004a).

Despite this improvement, we were still no closer to identifying secondary tags at a glance. Our third project addressed that issue.

4.3 MACHINE and LIVING ORGANISM metaphors in a novel and in business magazines

The extension project, funded by the British Academy (SG-42813), was based on two data sets which had previously been analysed manually: Kesey's novel One Flew over the Cuckoo's Nest (ca. 150,000 words), split into two sub-corpora of one half each (see Semino and Swindlehurst 1996), and 40 articles (ca. 81,000 words) from business magazines portraying female executives (see Koller 2004b). We were particularly interested in machine and living organism metaphors, as these had been found to be of central importance in the novel.

More specifically, Semino and Swindlehurst (1996) conducted a traditional stylistic analysis of Kesey's novel, which is narrated in the first person by Bromden, a patient in a mental hospital. They argued that Bromden's peculiar mental functioning is partly conveyed by his overreliance on vocabulary to do with machinery, and particularly by his use of machinery metaphors to talk about people, institutions, etc.[4] Cumulatively, the use of these metaphors results in a mechanistic world view, which is closely connected with Bromden's mental illness. Semino and Swindlehurst also point out that machinery metaphors decrease in frequency in the second half of the novel, and argue that this is due to plot developments that gradually free Bromden from his original mind set. In addition, Semino and Swindlehurst noted a contrasting pattern in (literal and metaphorical) references to living organisms, which become more frequent as the novel progresses. This can also be related to Bromden's gradual liberation from the most extreme aspects of his mechanistic worldview. Likewise, the business magazine features had — next to the war metaphor used to conceptualise the portrayed female executives as warriors — used metaphors from the machinery and living organism domains to refer to companies and, by extension, their leaders (see Morgan 2006: 11-70). Our overall aim was to compare the use of metaphor source domains across sub-sets of data to ascertain change in metaphor usage within texts, as well as differences in usage between genres.

The choice of source domains presented us with the problem that not all domains manually identified and labelled by a researcher are actually reflected in the USAS tag set. Thus, while living organisms is captured by the tags L1, L2 and L3 ("life and living things", "living creatures: animals, birds, etc.", "plants"), there is no tag "machines". This meant that we had to find the closest match in the tag set, which in this case was the tags O2 and O3 ("objects generally", "electricity and electrical equipment"). In order to automate the searching for source and target domains, whether they were in first choice or secondary positions, we implemented a new feature in Wmatrix called broad sweep. In essence, the broad sweep search function searches the full list of possible semantic tags on each word in the text. An additional filter

[4] The dominance of the machinery domain in Bromden's worldview can be explained in autobiographical terms: He is familiar with machines thanks to his training as an electrician, but is also in awe of them as a result of an air raid in WW2 which precipitated his mental disturbances and resulted in his current hospitalisation.

for searching the resulting frequency list groups together portmanteau tags and tag-extensions. An example of the former is the high-level tag S1 ("social actions, states and processes"), which is differentiated into S1.2 ("personality traits") and further into S1.2.1 ("approachability and friendliness"). Tag extensions, on the other hand, include instances such as T3 ("time and age") and its special form T3- ("young"). By identifying such sub-categories, the broad sweep search does away with the need to go back to the tag set to see what else is included in the identified tag. This speeded up the analysis considerably.

The analysis of Kesey's novel involved three files, containing, respectively: the whole of the novel, the novel's first half, and the novel's second half.[5] The USAS facility in Wmatrix was used to semantically annotate each file, and to compare each file with the Imaginative writing section of the BNC sampler. Our discussion of the results focuses on the machinery source domain only, however, as the cut-off point that was chosen to divide the novel into two halves was appropriate for the analysis of this particular source domain, but not for the living organisms source domain.

A comparison of the novel as a whole with the reference corpus resulted in a list of 66 key semantic domains with a log likelihood threshold of 6.63 or above. The two most overused semantic domains were: "medicines and medical treatment" (log likelihood 893.67), and "objects generally" (log likelihood 547.45). The overuse of the former domain can be straightforwardly related to the fact that the novel is mostly set in a hospital. As a consequence, the expressions included under this domain are unlikely to be used metaphorically. The latter semantic domain, in contrast, was investigated in more detail.

Overall, the semantic tag "objects generally" was allocated as first choice tag to 1,352 tokens, corresponding to 326 types. These relate to a wide range of objects, including, for example, thing(s), bottles, bucket, etc. By examining the list of types, we were able to discard those items that were likely to be used to refer literally to elements of the text world (e.g. needles, glasses), but we also noticed the presence of several expressions that Semino and Swindlehurst had

[5] The cut-off point between the two parts corresponds to a particular plot development that, according to Semino and Swindlehurst (1996), is crucial in the development of Bromden's worldview and mental functioning.

analysed as machinery metaphors. These included some of the types with the highest number of tokens, such as machine (24 tokens), machinery (21 tokens), clock (17 tokens), wires (17 tokens), and machines (17 tokens). All types that could be related to the machinery source domain were further investigated via the concordance facility included in the Wmatrix environment. This enabled us to view all instances of each type in context, and to identify the presence of any metaphoric tokens. This analysis revealed that approximately half of the occurrences of machinery-related lexis in the novel involved non-literal expressions. This confirmed Semino and Swindlehurst's general claim about the presence of machinery metaphors in the novel. Crucially, we were also able to identify an open-ended set of machinery metaphors, something which is not possible when using concordances alone.[6]

We then proceeded to investigate Semino and Swindlehurst's claim that machinery metaphors decrease in frequency as the novel progresses. A direct comparison of the first half of the novel with the second half yielded relatively few overused semantic domains, and did not reveal any relevant patterns. We therefore adopted the same procedure for each of the two halves of the novel that we have described for the whole novel, i.e. we compared each half with the Imaginative writing section of the BNC sampler. In both cases, the "objects generally" semantic domain turned out to be the second most overused semantic domain (log likelihood 362.06 for the first half of the novel and 329.24 for the second half of the novel). However, there is a difference between the two halves of the novel in terms of the types that were included under this semantic domain. The machinery-related expressions we mentioned above mostly occur in the first half of the novel. In contrast, the types included under "objects generally" in the second half of the novel mostly correspond to objects such as pole and tile, which literally "exist" in the fictional world. This provides some support for Semino and Swindlehurst's manual analysis of the differences between the two halves of the novel.

The other dataset, consisting of articles from business magazines describing female executives, had also previously been observed to contain machine and

[6] We used the same approach to investigate Semino and Swindlehurst's claim that Bromden makes particularly frequent use of the conceptual metaphor powerful is big. We found that the 'Size: Big' semantic domain is the eighth most overused domain in the novel (log likelihood 139.86), and that the expressions included under this domain are mostly used metaphorically.

living organism metaphors on the basis of a prior, manual analysis (Koller 2004b). As before, we compared the dataset to a reference corpus (here, the Written section of the BNC sampler) and established a list of 152 key semantic domains. Among these, "objects generally" was the 57th most key (log likelihood 35.42). However, only 5.9 per cent of the 459 tokens in this category are actually used metaphorically – unsurprising given the general nature of the category as noted above. "Living creatures: animals" and "plants" also featured on the list of key domains (in 18th and 63rd place respectively, log likelihood 117.0 and 32.4), together accounting for 163 tokens, of which 14.7 per cent proved to be metaphoric. The number of instances of metaphoric usage within these two fields was substantially higher than the number identified in the original manual analysis (27 tokens as compared to eight in the manual analysis for "objects generally", and 34 as compared to 17 for "living creatures: animals" and "plants" taken together). While extensive manual analysis of the items retrieved via the semantic annotation was required, this proved no more time-consuming than a fully manual analysis, while at the same time generating more comprehensive results.

A trend that was observable with both "objects generally" and the two living organism categories was that the metaphors tended to be used to conceptualise the company, rather than the female executive who is the subject of the article. In the case of the living organism metaphors, the company is conceptualised as an animal or plant and the executive as a carer, nurturer, or more specifically gardener. Interestingly, for the majority of tokens in these categories, the semantic tag in first position was the source domain. The exception was the highly conventional metaphorical usage of growth and branch(es) (relating to companies). This further corroborates our hypothesis that types representing metaphoric expressions tend to be allocated the source domain as first choice tag.

5. Evaluation and outlook

Through the research carried out in the three projects described above, we have uncovered new requirements for the USAS tagger and the Wmatrix software. With the addition of the two features described above (domain push and broad sweep), we have gone some way to addressing the needs of metaphor researchers. These are the first steps that we have highlighted in

order to support approaches using a corpus-based methodology for the investigation of metaphor in large-scale data sets. The techniques are intended to complement searches based on pre-determined search strings and move towards the scalable investigation of open-ended sets of metaphoric expressions.

To carry the project forward, we envisage a large-scale study of metaphor use in different genres. This will be based on a corpus of one million words, selected from the 100 million words of the BNC. In particular, the corpus will be divided into ten sub-sections based on genre, nine written and one spoken, all of which are comparable in terms of subject matter. Because of the difficulty in working out contextual meaning in conversational data, our spoken subsection will contain what in the BNC is termed "context-governed" speech, i.e. public meetings, etc. Of the written subsections, one is planned to consist of fiction, and one of institutional documents such as government leaflets. The other eight will represent different topic domains (arts, sciences) within the broader text types of academic writing, and newspaper text. This design will allow us to examine metaphor use across a range of potentially significant dimensions of text-type variation.

Following on from this third project, we are particularly interested in the genre perspective, because as mentioned above, the interaction between genre and topic in text types such as medical or business magazine articles or the industrial novel is of special importance for ascertaining metaphor. By contrast, we would anticipate more broadly-defined text types, such as academic or fictional prose in general, to be more varied in both the topics that are discussed and, therefore, equally varied in the conceptual metaphors and related metaphoric expressions that are employed. In terms of genre differences, then, we wish to find out both what different metaphors are attracted by different genres, and what genre-specific uses there are of "ubiquitous" metaphors. We believe that the three projects outlined in this paper have laid the ground for a larger-scale analysis of metaphor usage in different genres, and we hope to have provided the metaphor research community with an additional tool for analysing metaphor in discourse.

References

Cameron, Lynne (2003): *Metaphor in Educational Discourse*, London.

Cameron, Lynne/Deignan, Alice. (2003): "Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse", in: *Metaphor and Symbol* 18, 149-160.

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

Deignan, Alice (2005): *Metaphor and Corpus Linguistics*, Amsterdam.

Eubanks, Philip (2005): "Globalization, 'corporate rule', and blended worlds. A conceptual-rhetorical analysis of metaphor, metonymy, and conceptual blending", in: *Metaphor and Symbol* 20, 173-197.

Koller, Veronika (2004a): *Metaphor and Gender in Business Media Discourse*, Basingstoke.

Koller, Veronika (2004b): "Businesswomen and war metaphors: 'possessive, jealous and pugnacious'?", in: *Journal of Sociolinguistics* 8, 3-22.

Koller, Veronika (forthcoming) "Missions and empires: religious and political metaphors in corporate discourse", in: Musolff, Andreas/Zinken, Jörg (edd.): *Metaphor and Discourse*, Basingstoke.

Koller, Veronika/Semino, Elena (forthcoming): "Metaphor, politics and gender: a case study from Germany", in: Ahrens, Kathleen (ed.): *Politics, Gender and Conceptual Metaphor*, Basingstoke.

Low, Graham (2003): "Validating metaphoric models in Applied Linguistics", *Metaphor and Symbol* 18, 239-254.

Morgan, Gareth (2006): Images of Organization (updated ed.), *Thousand Oaks*, CA.

Musolff, Andreas (2004): *Metaphor and Political Discourse. Analogical Reasoning about Europe*, Basingstoke.

Piao, Scott/Rayson, Paul/Archer, Dawn/McEnery, Tony (2004): "Evaluating lexical resources for a semantic tagger", in: *Proceedings of 4th International Conference on Language Resources and Evaluation* (LREC 2004), 499-502.

Rayson, Paul/Archer, Dawn/Piao, Scott/McEnery, Tony (2004): "The UCREL semantic analysis system", in: *Proceedings of the Workshop "Beyond Named Entity Recognition. Semantic Labelling for NLP Tasks" in association with the 4th International Conference on Language Resources and Evaluation* (LREC 2004), 7-12.

Semino, Elena (2002): "A sturdy baby or a derailing train? Metaphorical representations of the euro in British and Italian newspapers", in: *Text* 22, 107-139.

Semino, Elena/Koller, Veronika (forthcoming): "Metaphor, politics and gender: a case study from Italy", in: Ahrens, Kathleen (ed.): *Politics, Gender and Conceptual Metaphor*, Basingstoke.

Semino, Elena/Swindlehurst, Kate (1996): "Metaphor and mind style in Ken Kesey's One Flew Over the Cuckoo's Nest", in: *Style* 30, 143-166.

Semino, Elena/Heywood, John/Short, Mick (2004): "Methodological problems in the analysis of a corpus of conversations about cancer", in: *Journal of Pragmatics* 36, 1271-1294.

Skorczynska, Hanna/Deignan, Alice (206): "Readership and purpose in the choice of economics metaphors", in: *Metaphor and Symbol* 21, 87-104.

Steen, Gerard (1999): "From linguistic to conceptual metaphor in five steps", in: Gibbs, Raymond W. Jr./Steen, Gerard (edd.): *Metaphor in Cognitive Linguistics*, Amsterdam, 57-77.

Stefanowitsch, Anatol (2005): "The function of metaphor: developing a corpus-based perspective", in: *International Journal of Corpus Linguistics* 10, 161-198.