Argumentation-Relevant Metaphors in Test-Taker Essays

Beata Beigman Klebanov and Michael Flor
Educational Testing Service
{bbeigmanklebanov,mflor}@ets.org

Abstract

This article discusses metaphor annotation in a corpus of argumentative essays written by test-takers during a standardized examination for graduate school admission. The quality of argumentation being the focus of the project, we developed a metaphor annotation protocol that targets metaphors that are relevant for the writer's arguments. The reliability of the protocol is κ=0.58, on a set of 116 essays (the total of about 30K content-word tokens). We found a moderate-to-strong correlation (r=0.51-0.57) between the percentage of metaphorically used words in an essay and the writing quality score. We also describe encouraging findings regarding the potential of metaphor identification to contribute to automated scoring of essays.

1 Introduction

The goal of our project is to automatically score the quality of argumentation in essays written for a standardized graduate school admission exam. Metaphors being important argumentative devices, we report on annotating data for potential training and testing of metaphor detection software that would eventually be used for automated scoring of essays.

Metaphors of various kinds can be relevant to argumentation. Some metaphors create vivid images and function as examples or as organizing ideas behind a series of examples. These are akin to pictures that are worth a thousand words, and are highly potent rhetorical devices. Metaphors of a less artistic crafting – more conventionalized ones, metaphors that we "live by" according to Lakoff and Johnson's (1980) famous tenet – subtly organize our thinking and language production in culturally coherent ways.

For an example of a vivid metaphor that helps organize the essay, consider an essay on the relationship between arts and government funding thereof (see example 1). The author's image of a piece of art as a slippery object that escapes its captor's grip as a parallel to the relationship between an artist and his or her patron/financier is a powerful image that provides a framework for the author's examples (in the preceding paragraph, Chaucer is discussed as a clever and subversive writer for his patron) and elaborations (means of "slippage", like *veiled imagery*, *multiple meanings*, etc).

(1) Great artistic productions, thus, tend to rise above the money that bought them, to bite, as it were, the hand that fed them. This is not always so, of course. But the point is that **great art is too slippery to be held in the grip of a governing power**. Through veiled imagery, multiple meanings, and carefully guarded language, a poem can both powerfully criticize a ruler and not blow its cover.

For an example of a conventional metaphor, consider the metaphor of construction/building. The connotation of *foundations* is something essential, old, solid, and lying deep, something that, once laid, remains available for new construction for a long period of time. It is often used to explain emergence

Proceedings of the First Workshop on Metaphor in NLP, pages 11–20,
Atlanta, Georgia, 13 June 2013.

of things – the existence of *foundations* (or support, or basis) is contrasted with the (presumed) idea of appearance out of nothing. Certain topics of discussion are particularly amenable for arguments from construction-upon-foundation. For example, consider an essay question "Originality does not mean thinking something that was never thought before; it means putting old ideas together in new ways," where an explanation of the emergence of something is required. Examples 2-6 show excerpts from essays answering this prompt that employ the foundation metaphor.

(2) The foundation of the United States was also based on a series of older ideas into which the fathers of our nation breathed new life.

(3) History is a progressive passing on of ideas, a process of building on the foundations laid by the previous generations. New ideas cannot stand if they are without support from the past.

(4) New discoveries and ideas are also original for some time, but eventually they become the older, accepted pieces that are the building blocks for originality.

(5) Original thinking can include old ideas which almost always are a basis for continued thought leading to new ideas.

(6) Humans are born of their ancestors, thrive from their intelligence, and are free to build on the intellectual foundations laid.

The two types of metaphors exemplified above have different argumentative roles. The first organizes a segment of an essay around it, firstly by imposing itself on the reader's mind (a property rhetoricians call *presence* (Perelman and Olbrechts-Tyteca, 1969; Gross and Dearin, 2003; Atkinson et al., 2008)), secondly by helping select supporting ideas or examples that are congruent with the parts of the target domain that are highlighted by the metaphor (this property is termed *framing* (Lakoff, 1991; Entman, 2003)), such as the idea of evasiveness purported by the ART AS A SLIPPERY OBJECT metaphor that is taken up both in the preceding Chaucer example and in an elaboration.

By contrast, metaphors "we live by" without even noticing, such as TIME IS MONEY or IDEAS ARE BUILDINGS, are not usually accorded much reader attention; they are processed by using the conventional connotation of the word as if it were an additional sense of that word, without invoking a comparison between two domains (for processing by categorization see (Bowdle and Gentner, 2005; Glucksbeg and Haught, 2006)). Thus, the word *foundation* is unlikely to elicit an image of a construction site, but rather will directly invoke the concept of something essential and primary. It is unclear to what extent such highly conventionalized metaphors that are not deliberately construed as metaphors have the framing property beyond framing induced by any lexical choice – that of stressing the chosen over the un-chosen alternative (Billig, 1996). Therefore, the fact that an essay writer used a conventional metaphor is not in itself a mark of rhetorical sophistication; it is possible, however, that, if certain metaphorical source domains are particularly apt for the given target domain (as the domain of construction to discuss emergence), using the metaphor is akin to choosing a solid though not particularly original argument.

Our interest being in metaphors that play a role in argumentation, we attempted to devise an annotation protocol that would be specifically geared towards identification of such metaphors. In what follows, we review the literature on approaches to annotating metaphors in a given discourse (section 2), we describe the protocol and the annotation procedure (section 3), report inter-annotator agreement (section 4), quantify the relationship between metaphorical density (percentage of metaphorically used words in an essay) and essay quality as measured by essay score, as well as estimate the potential usefulness of metaphor detection for automated scoring of essays (section 5.2).

2 Related Work

Much of the contemporary work on metaphor in psychological and computational veins is inspired by Lakoff and Johnson's (1980) research on conceptual metaphor. Early work in this tradition concentrated on mapping the various conceptual metaphors in use in a particular culture (Lakoff and Johnson, 1980; Lakoff and Kövecses, 1987; Kövecses, 2002). Examples for various conceptual mappings are collected, resulting in the Master Metaphor List (Lakoff et al., 1991), showing common metaphorical mappings and their instances of use. For example, the LIFE IS A JOURNEY conceptual metaphor that maps the source domain of JOURNEY to the target domain of LIFE is used in expressions such as:

- He just sails through life.
- He is headed for great things.
- If this doesn't work, I'll just try a different route.
- She'll cross that bridge when she comes to it.
- We've come a long way.

While exemplifying the extent of metaphoricity of everyday English, such a list is not directly applicable to annotating metaphors in discourse, due to the limited coverage of the expressions pertaining to each conceptual metaphor, as well as of the conceptual metaphors themselves.

Studies of discourse metaphor conducted in the Critical Discourse Analysis tradition (Musolff, 2000; Charteris-Black, 2005) analyze a particular discourse for its employment of metaphors. For example, an extensive database of metaphors in British and German newspaper discourse on European integration in the 1990s was compiled by Musolff (2000); the author did not make it clear how materials for annotation were selected.

A systematic but not comprehensive approach to creating a metaphor-rich dataset is to pre-select materials using linguistic clues (Goatly, 1997) for the presence of metaphor, such as *utterly* or *so to speak*. Shutova and Teufel (2010) report precision statistics for using different clues to detect metaphoric sentences; expressions such as *literally*, *utterly*, and *completely* indicate a metaphorical context in more than 25% of cases of their use in the British National Corpus. Such cues can aid in pre-selecting data for annotation so as to increase the proportion of materials with metaphors beyond a random sample.

Another approach is to decide on the source domains of interest in advance, use a dictionary or thesaurus to detect words belonging to the domain, and annotate them for metaphoricity (Stefanowitsch, 2006; Martin, 2006; Gedigan et al., 2006). Gedigan et al. (2006) found that more than 90% of verbs belonging to MOTION and CURE domains in a Wall Street Journal corpus were used metaphorically. Fixing the source domain is potentially appropriate if common metaphorically used domains in a given discourse have already been identified, as in (Koller et al., 2008; Beigman Klebanov et al., 2008).

A complementary approach is to fix the target domain, and do metaphor "harvesting" in a window around words belonging to the target domain. For example, Reining and Löneker-Rodman (2007) chose the lemma *Europe* to represent the target domain in the discourse on European integration. They extracted small windows around each occurrence of *Europe* in the corpus, and manually annotated them for metaphoricity. This is potentially applicable to analyzing essays, because the main target domain of the discourse is usually given in the prompt, such as *art*, *originality*. The strength of this method is its ability to focus on metaphors with argumentative potential, because the target domain, which is the topic of the essay, is directly involved. The weakness is the possibility of missing metaphors because they are not immediately adjacent to a string from the target domain.

The Metaphor Identification Procedure (MIP) is a protocol for exhaustive metaphoricity annotation proposed by the Pragglejaz group (Pragglejaz, 2007). The annotator classifies every word in a document (including prepositions) as metaphorical if it has "a more basic contemporary meaning" in other contexts than the one it has in the current context. Basic meanings are explained to be "more concrete, related to bodily action, more precise, and historically older." The authors – all highly qualified linguists who have a long history of research collaboration on the subject of metaphor – attained a kappa of 0.72 for 6 annotators for one text of 668 words and 0.62 for another text of 676 words. Shutova and Teufel (2010) used the protocol to annotate content verbs only, yielding kappa of 0.64 for 3 volunteer annotators with some linguistic background, on a set of sentences containing 142 verbs sampled from the British National Corpus. It is an open question how well educated lay people can agree on an exhaustive metaphor annotation of a text.

We note that the procedure is geared towards conceptual metaphors at large, not necessarily argumentative ones, in that the protocol does not consider the writer's purpose in using the metaphor. For example, the noun *forms* in "All one needs to use high-speed forms of communication is a computer or television and an internet cable" is a metaphor according to the MIP procedure, because the basic meaning "a shape of something" is more concrete/physical than the contextual meaning "a type of something," so a physical categorization by shape stands for a more abstract categorization into types. This metaphor could have an argumentative purport; for instance, if the types in question were actually very blurred and difficult to tell apart, by calling them forms (and, by implications, shapes), they are framed as being more clearly and easily separable than they actually are. However, since the ease of categorization of high-speed electronic communication into types is not part of the author's argument, the argumentative relevance of this metaphor is doubtful.

3 Annotation Protocol

In the present study, annotators were given the following guidelines:

> Generally speaking, a metaphor is a linguistic expression whereby something is compared to something else that it is clearly literally not, in order to make a point. Thus, in Tony Blair's famous "I haven't got a reverse gear," Tony Blair is compared to a car in order to stress his unwillingness/inability to retract his statements or actions. We would say in this case that a metaphor from a vehicle domain is used.
>
> . . . [more examples] . . .
>
> The first task in our study of metaphor in essays is to read essays and underline words you think are used metaphorically. Think about the point that is being made by the metaphor, and write it down. Note that a metaphor might be expressed by the author or attributed to someone else. Note also that the same metaphor can be taken up in multiple places in a text.

During training, two annotators were instructed to apply the guidelines to 6 top-scoring essays answering a prompt about the role of art in society. After they finished, sessions were held where the annotators and one of the authors of this paper discussed the annotations, including explication of the role played by the metaphor in the essay. A summary document that presents a detailed consensus annotation of 3 of the essays was circulated to the annotators. An example of an annotation is shown below (metaphors are boldfaced in the text and explained underneath):

> F. Scott Fitzgerald wrote, "There is a **dark night** in every man's soul where it is always **2 o'clock in the morning**." His words are a profound statement of human nature. Within society, we operate under a variety of social **disguises**. Some of these **masks** become so second nature that we find ourselves unable to **take** them **off**.
>
> (1) Dark night, 2 o'clock in the morning: True emotions are not accessible (at 2 o'clock a person is usually asleep and unaware of what is going on) and frightening to handle on one's own (scary to walk at night alone); people need mediation to help accessibility, and also company to alleviate the fear. Art provides both. This metaphor puts forward the two main arguments: accessibility and sharing.
>
> (2) Masks, take off, disguises: could be referring to the domain of theater/performance. Makes the point that what people do in real life to themselves is superficially similar to what art (theater) does to performers – hiding their true identity. In the theater, the hiding is temporary and completely reversible at will, there is really no such thing as inability to take off the mask. The socially-inflicted hiding is not necessarily under the person's control, differently from a theatrical mask. Supports and extends the accessibility argument: not just lack of courage or will, but lack of control to access the true selves.

The actual annotation then commenced, on a sample of essays answering a different question (the data will be described in section 3.1). Annotators were instructed to mark metaphors in the text using a graphical interface that was specially developed for the project. The guidelines for the actual annotation are shown below:

> During training, you practiced careful reading while paying attention to non-literal language and saw how metaphors work in their context. At the annotation stage, you are not asked to explicitly interpret the metaphor and identify its argumentative contribution (or rather, its attempted argumentative contribution), only to mark metaphors, trusting your intuition that you *could* try to interpret the metaphor in context if needed.

Note that we have not provided formal definitions of what a literal sense is in order to not interfere with intuitive judgments of metaphoricity (differently from Pragglejaz (2007), for example, who provide definition of a basic sense). Neither have we set up an explicit classification task, whereby annotators are required to classify every single word in the text as a metaphor or a non-metaphor (again, differently from Pragglejaz (2007)); in our task, annotators were instructed to mark metaphors while they read. This is in the spirit of Steen's (2008) notion of deliberate metaphors – words and phrases that the writer actually meant to produce as a metaphor, as opposed to cases where the writer did not have a choice, such as using *in* for an expression like *in time*, due to the pervasiveness of the time-as-space metaphor. Note, however, that Steen's notion is writer-based; since we have no access to the writers of the essays, we side with an educated lay reader and his or her perception of a metaphorical use.

The annotators were instructed to give the author the benefit of the doubt and *not* to assume that a common metaphor is necessarily unintenional:

> When deciding whether to attribute to the author the intention of making a point using a metaphor, please be as liberal as you can and give the author the benefit of the doubt. Specifically, if something is a rather common metaphor that still happens to fit nicely into the argument the author is making, we assume that the author intended it that way.

To clarify what kinds of metaphors are excluded by our guidelines, we explained as follows:

> In contrast, consider cases where an expression might be perhaps formally classified as a metaphor, but the literal sense cannot be seen as relevant to the author's argument. For example, consider the following sentence from Essay 2 from our training material: "Seeing the beauty of nature or hearing a moving piece of music may drive one to perhaps try to replicate that beauty *in* a style of one's own." Look at the italicized word – the preposition *in*. According to some theories of metaphor, that would constitute a metaphorical use: Literally, *in* means inside some container; since style is not literally a container, the use of *in* here is non-literal. Suppose now that the non-literal interpretation invites the reader to see style as a container. A container might have more or less room, can be full or empty, can be rigid or flexible, can contain items of the same or different sorts – these are some potential images that go with viewing something as a container, yet none of them seems to be relevant to whatever the author is saying about style, that is, that it is unique (one's own) and yet the result is not quite original (replication).

The two annotators who participated in the task hold BA degrees in Linguistics, but have no background in metaphor theory. They were surprised and bemused by an example like *in style*, commenting that it would never have occurred to them to mark it as a metaphor. In general, the thrust of this protocol is to identify metaphorical expressions that are noticeable and support the author's argumentative moves; yet, we targeted a reasonable timeline for completing the task, with about 30 minutes per text, therefore we did not require a detailed analysis of the marked metaphors as done during training.

3.1 Data

Annotation was performed on 116 essays written on the following topic: “High-speed electronic communications media, such as electronic mail and television, tend to prevent meaningful and thoughtful communication.” Test-takers are instructed to present their perspective on the issue, using relevant reasons and/or examples to support their views. Test-takers are given 45 minutes to compose an essay. The essays were sampled from the dataset analyzed in Attali et al. (2013), with oversampling of longer essays. In the Attali et al. (2013) study, each essay was scored for the overall quality of English argumentative composition; thus, to receive the maximum score, an essay should present a cogent, well-articulated analysis of the complexities of the issue and convey meaning skillfully. Each essay was scored by 16 professional raters on a scale of 1 to 6, allowing plus and minus scores as well, quantified as 0.33 – thus, a score of 4- is rendered as 3.67. This fine-grained scale resulted in a high mean pairwise inter-rater correlation (r=0.79). We use the average of 16 raters as the final score for each essay. This dataset provides a fine-grained ranking of the essays, with almost no two essays getting exactly the same score.

For the 116 essays, the mean length was 478 words (min: 159, max: 793, std: 142); mean score: 3.82 (min: 1.81, max: 5.77, std: 0.73). Table 1 shows the distribution of essay scores.

Score	Number of Essays	Proportion of Essays
2	4	0.034
3	33	0.284
4	59	0.509
5	19	0.164
6	1	0.009

Table 1: Score distribution in the essay data. The first column shows the rounded score. For the sake of presentation in this table, all scores were rounded to integer scores, so a score of 3.33 was counted as 3, and a score of 3.5 was counted as 4.

4 Inter-Annotator Agreement and Parts of Speech

The inter-annotator agreement on the total of 55,473 word tokens was κ=0.575. In this section, we investigate the relationship between part of speech and metaphor use, as well as part of speech and inter-annotator agreement.

For this discussion, words that appear in the prompt (essay topic) are excluded from all sets. Furthermore, we concentrate on content words only (as identified by the OpenNLP tagger[1]). Table 2 shows the split of the content-word annotations by part of speech, as well as the reliability figures. We report information for each of the two annotators separately, as well as for the union of their annotations. We report the union as we hypothesize that a substantial proportion of apparent disagreements between annotators are attention slips rather than substantive disagreements; this phenomenon was attested in a previous study (Beigman Klebanov et al., 2008).

POS	**Count**	**A1**	**A2**	**A1⋃A2**	κ
All	55,473	2,802	2,591	3,788	0.575
Cont.	29,207	2,380	2,251	3,211	0.580
Noun	12,119	1,033	869	1,305	0.596
Adj	4,181	253	239	356	0.525
Verb	9,561	1,007	1,039	1,422	0.563
Adv	3,346	87	104	128	0.650

Table 2: Reliability by part of speech. The column Count shows the total number of words in the given part of speech across the 116 essays. Columns A1 and A2 show the number of items marked as metaphors by annotators 1 and 2, respectively, while Column A1⋃A2 shows numbers of items in the union of the two annotations. The second row presents the overall figure for content words.

Nouns constitute 41.5% of all content words; they are 43.4% of all content-word metaphors for annotator 1, 38.6% for annotator 2, and 40.6% for the union of the two annotations. Nouns are therefore represented in the metaphor annotated data in their general distribution proportions. Of all nouns, 7%-8.5% are identified as metaphors by a single annotator, while 10.8% of the nouns are metaphors in the union annotation.

[1] http://opennlp.apache.org/index.html

Verbs are 32.7% of all content words; they are 42.3% of all content-word metaphors for annotator 1, 46.2% for annotator 2, and 44.3% in the union. Verbs are therefore over-represented in the metaphor annotated data relative to their general distribution proportions. Of all verbs, 10.5%-10.9% are identified as metaphors by a single annotator, while 14.9% are metaphors in the union annotation.

Adjectives are 14.3% of all content words; they are 10.6% of all content-word metaphors for annotator 1, 10.6% for annotator 2, and 11.1% in the union. Adjectives are therefore somewhat under-represented in the metaphor annotated data with respect to their general distribution. About 6% of adjectives are identified as metaphors in individual annotations, and 8.5% in the union annotation.

Adverbs are 11.5% of all content words; they are 3.7% of all content-word metaphors for annotator 1 and 4.6% for annotator 2, and 4% in the union. Adverbs are heavily under-represented in the metaphor annotated data with respect to their general distribution. Of all non-prompt adverbs, about 3-4% are identified as metaphors.

The data clearly points towards the propensity of verbs towards metaphoricity, relative to words from other parts of speech. This is in line with reports in the literature that identify verbs as central carriers of metaphorical vehicles: Cameron (2003) found that about 50% of metaphors in educational discourse are realized by verbs, beyond their distributional proportion; this finding prompted Shutova et al. (2013) to concentrate exclusively on verbs.

According to Goatly (1997), parts of speech differ in the kinds of metaphors they realize in terms of the recognizability of the metaphorical use as such. Nouns are more recognizable as metaphors than other word classes for the following two reasons: (1) Since nouns are referring expressions, they reveal very strongly the clashes between conventional and unconventional reference; (2) Since nouns often refer to vivid, imaginable entities, they are more easliy recognized than metaphors of other parts of speech. Moreover, morphological derivation away from nouns – for example, by affixation – leads to more lexicalized and less noticeable metaphors than the original nouns.

Goatly's predictions seem to be reflected in inter-annotator agreement figures for nouns versus adjectives and verbs, with nouns yielding higher reliability of identification than verbs and adjectives, with the latter two categories having more cases where only one but not both of the annotators noticed a metaphorical use. Since adverbs are the most distant from nouns in terms of processes of morphological derivation, one would expect them to be less easily noticeable, yet in our annotation adverbs are the most reliably classified category.

Inspecting the metaphorically used adverbs, we find that a small number of adverbs cover the bulk of the volume: *together* (11), *closer* (11), *away* (10), *back* (8) account for 46% of the adverbs marked by annotator 1 in our dataset. Almost all cases of *together* come from a use in the phrasal verb *bring together* (8 cases), in expressions like "bringing the world together into one cyberspace without borders" or "electronic mail could bring people closer together" or "bringing society together." In fact, 6 of the 11 cases of *closer* are part of the construction *bring closer together*, and the other cases have similar uses like "our conversations are more meaningful because we are closer through the internet."

Interestingly, the metaphorical uses of *away* also come from phrasal constructions that are used for arguing precisely the opposite point – that cyber-communications drive people away from each other: "email, instant messaging, and television support a shift away from throughful communication," "mass media and communications drive people away from one another," "by typing a message ... you can easily get away from the conversation."

It seems that the adverbs marked for metaphoricity in our data tend to be (a) part of phrasal constructions, and (b) part of a commonly made argument for or against electronic communication – that it (metaphorically) brings people together, or (metaphorically) drives them apart by making the actual togetherness (co-location) unnecessary for communication. The adverbs are therefore not of the derivationally complex kind Goatly has in mind, and their noticeability might be enhanced by being part of a common argumentative move in the examined materials, especially since the annotators were instructed to look out for metaphors that support the writer's argument.

5 Metaphor and Content Scoring

In order to assess the potential of metaphor detection to contribute to essay scoring, we performed two tests: correlation with essay scores and a regression analysis in order to check whether metaphor use contributes information that is beyond what is captured by a state-of-art essay scoring system.

As a metaphor-derived feature, we calculated **metaphorical density**, that is, the percentage of metaphorically used words in an essay: All words marked as metaphors in an essay were counted (content or other), and the total was divided by essay length.

5.1 E-rater

As a reference system, we use e-rater (Attali and Burstein, 2006), a state-of-art essay scoring system developed at Educational Testing Service.[2] E-rater computes more than 100 micro-features, which are aggregated into macro-features aligned with specific aspects of the writing construct. The system incorporates macro-features measuring grammar, usage, mechanics, style, organization and development, lexical complexity, and vocabulary usage. Table 3 gives examples of micro-features covered by the different macro-features.

Macro-Feature	Example Micro-Features
Grammar, Usage, and Mechanics	agreement errors verb formation errors missing punctuation
Style	passive, very long or very short sentences, excessive repetition
Organization and Development	use of discourse elements: thesis, support, conclusion
Lexical Complexity	average word frequency average word length
Vocabulary	similarity to vocabulary in high- vs low-scoring essays

Table 3: Features used in e-rater (Attali and Burstein, 2006).

E-rater models are built using linear regression on large samples of test-taker essays. We use an e-rater model built at Educational Testing Service using a large number of essays across different prompts, with no connection to the current project and its authors. This model obtains Pearson correlations of r=0.935 with the human scores. The excellent performance of the system leaves little room for improvement; yet, none of the features in e-rater specifically targets the use of figurative language, so it is interesting to see the extent to which metaphor use could help explain additional variance.

5.2 Results

We found that metaphorical density attains correlation of r=0.507 with essay score using annotations of annotator 1, r=0.556 for annotator 2, and r=0.570 using the union of the two annotators. It is clearly the case that better essays tend to have higher proportions of metaphors.

We ran a regression analysis with essay score as the dependent variable and e-rater raw score and metaphor density in the union annotation as two independent variables. The correlation with essay score improved from 0.935 using e-rater alone to 0.937 using the regression equation (the adjusted R^2 of the model improved from 0.874 to 0.876). While the contribution of metaphor feature is not statistically significant for the size of our dataset (n=116, p=0.07), we are cautiously optimistic that metaphor detection can make a contribution to essay scoring when the process is automated and a larger-scale evaluation can be performed.

6 Conclusion

This article discusses annotation of metaphors in a corpus of argumentative essays written by test-takers during a standardized examination for graduate school admission. The quality of argumentation being the focus of the project, we developed a metaphor annotation protocol that targets metaphors that are relevant for the writer's arguments. The reliability of the protocol is κ=0.58, on a set of 116 essays (a total of about 30K content word tokens).

We found a moderate-to-strong correlation (r=0.51-0.57) between the density of metaphors in an essay (percentage of metaphorically used words) and the writing quality score as provided by professional essay raters.

As the annotation protocol is operationally effi-

[2] http://www.ets.org/erater/about/

cient (30 minutes per essay of about 500 words), moderately reliable (κ=0.58), and uses annotators that do not possess specialized knowledge and training in metaphor theory, we believe it is feasible to annotate a large set of essays for the purpose of building a supervised machine learning system for detection of metaphors in test-taker essays. The observed correlations of metaphor use with essay score, as well as the fact that metaphor use is not captured by state-of-art essay scoring systems, point towards the potential usefulness of a metaphor detection system for essay scoring.

References

Nathan Atkinson, David Kaufer, and Suguru Ishizaki. 2008. Presence and Global Presence in Genres of Self-Presentation: A Framework for Comparative Analysis. *Rhetoric Society Quarterly*, 38(3):1–27.

Yigal Attali and Jill Burstein. 2006. Automated Essay Scoring With e-rater®V.2. *Journal of Technology, Learning, and Assessment*, 4(3).

Yigal Attali, Will Lewis, and Michael Steier. 2013. Scoring with the computer: Alternative procedures for improving reliability of holistic essay scoring. *Language Testing*, 30(1):125–141.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. 2008. Analyzing disagreements. In *COLING 2008 workshop on Human Judgments in Computational Linguistics*, pages 2–7, Manchester, UK.

Michael Billig. 1996. *Arguing and Thinking: A Rhetorical Approach to Social Psychology*. Cambridge University Press, Cambridge.

Brian Bowdle and Dedre Gentner. 2005. The career of metaphor. *Psychological Review*, 112(1):193–216.

Lynne Cameron. 2003. *Metaphor in Educational Discourse*. Continuum, London.

Jonathan Charteris-Black. 2005. *Politicians and rhetoric: The persuasive power of metaphors*. Palgrave MacMillan, Houndmills, UK and New York.

Robert Entman. 2003. Cascading activation: Contesting the white houses frame after 9/11. *Political Communication*, 20:415–432.

Matt Gedigan, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *PProceedings of the 3rd Workshop on Scalable Natural Language Understanding*, pages 41–48, New York.

Sam Glucksbeg and Catrinel Haught. 2006. On the relation between metaphor and simile: When comparison fails. *Mind and Language*, 21(3):360–378.

Andrew Goatly. 1997. *The Language of Metaphors*. Routledge, London.

Alan Gross and Ray Dearin. 2003. *Chaim Perelman*. Albany: SUNY Press.

Zoltan Kövecses. 2002. *Metaphor: A Practical Introduction*. Oxford University Press.

Veronika Koller, Andrew Hardie, Paul Rayson, and Elena Semino. 2008. Using a semantic annotation tool for the analysis of metaphor in discourse. *Metaphorik.de*, 15:141–160.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press, Chicago.

George Lakoff and Zoltan Kövecses. 1987. Metaphors of anger in japanese. In D. Holland and N. Quinn, editors, *Cultural Models in Language and Thought*. Cambridge: Cambridge University Press.

George Lakoff, Jane Espenson, Adele Goldberg, and Alan Schwartz. 1991. Master Metaphor List, Second Draft Copy. Cognitive Linguisics Group, Univeristy of California, Berkeley: http://araw.mede.uic.edu/˜alansz/metaphor/METAPHORLIST.pdf.

George Lakoff. 1991. Metaphor and war: The metaphor system used to justify war in the gulf. *Peace Research*, 23:25–32.

James Martin. 2006. A corpus-based analysis of context effects on metaphor comprehension. In Anatol Stefanowitsch and Stefan Gries, editors, *Corpus-Based Approaches to Metaphor and Metonymy*. Berlin: Mouton de Gruyter.

Andreas Musolff. 2000. *Mirror images of Europe: Metaphors in the public debate about Europe in Britain and Germany*. München: Iudicium. Annotated data is available at http://www.dur.ac.uk/andreas.musolff/Arcindex.htm.

Chaïm Perelman and Lucie Olbrechts-Tyteca. 1969. *The New Rhetoric: A Treatise on Argumentation*. Notre Dame, Indiana: University of Notre Dame Press. Translated by John Wilkinson and Purcell Weaver from French original published in 1958.

Group Pragglejaz. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1):1–39.

Astrid Reining and Birte Löneker-Rodman. 2007. Corpus-driven metaphor harvesting. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 5–12, Rochester, New York.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor Corpus Annotated for Source-Target Domain Mappings. In *Proceedings of LREC*, Valetta, Malta.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(1).

Gerard Steen. 2008. The Paradox of Metaphor: Why We Need a Three-Dimensional Model of Metaphor. *Metaphor and Symbol*, 23(4):213–241.

Anatol Stefanowitsch. 2006. Corpus-based approaches to metaphor and metonymy. In Anatol Stefanowitsch and Stefan Gries, editors, *Corpus-Based Approaches to Metaphor and Metonymy*. Berlin: Mouton de Gruyter.