A Multi-Cultural Repository of Automatically Discovered Linguistic and Conceptual Metaphors

Samira Shaikh[1], Tomek Strzalkowski[1,3], Ting Liu[1], George Aaron Broadwell[1], Boris Yamrom[1], Sarah Taylor[2], Laurie Feldman[1], Kit Cho[1], Umit Boz[1], Ignacio Cases[1], Yuliya Peshkova[1] and Ching-Sheng Lin[1]

[1]State University of New York – University at Albany, [2]Sarah M. Taylor Consulting LLC, [3]Polish Academy of Sciences

E-mail: samirashaikh@gmail.com, tomek@albany.edu

Abstract

In this article, we present details about our ongoing work towards building a repository of Linguistic and Conceptual Metaphors. This resource is being developed as part of our research effort into the large-scale detection of metaphors from unrestricted text. We have stored a large amount of automatically extracted metaphors in American English, Mexican Spanish, Russian and Iranian Farsi in a relational database, along with pertinent metadata associated with these metaphors. A substantial subset of the contents of our repository has been systematically validated via rigorous social science experiments. Using information stored in the repository, we are able to posit certain claims in a cross-cultural context about how peoples in these cultures (America, Mexico, Russia and Iran) view particular concepts related to Governance and Economic Inequality through the use of metaphor. Researchers in the field can use this resource as a reference of typical metaphors used across these cultures. In addition, it can be used to recognize metaphors of the same form or pattern, in other domains of research.

Keywords: metaphors, computational linguistics, resource

1. Introduction

Our repository consists of automatically extracted Linguistic and Conceptual Metaphors. Metaphors are mapping systems that allow the semantics of a familiar Source domain to be applied to a new Target domain so as to invite new frameworks for reasoning to emerge in the target domain. For example, the sentence

> *These qualities have helped him navigate the labyrinthine **federal bureaucracy**…*

is an example of a Linguistic Metaphor (LM). It maps the Source relation *navigate the labyrinthine* to the Target Concept of ***federal bureaucracy***. This LM likens something *labyrinthine* and that can be *navigated* to the concept of ***federal bureaucracy***. A set of LMs that map other relations in this Source domain to the concept of bureaucracy would allow us to advance the Conceptual Metaphor (CM) – BUREAUCRACY IS A MAZE.

Metaphors are pervasive in discourse, used to convey meanings indirectly. Thus, they provide critical insights into the preconceptions, assumptions and motivations of underlying discourse, especially valuable when studied across cultures. When metaphors are thoroughly understood within the context of a culture, we can gain substantial knowledge about cultural values. These insights can help better shape cross-cultural understanding and facilitate discussions and negotiations among different communities.

Our automated system is able to efficiently detect metaphors in large amounts of textual data in different languages. System output is stored in a relational database, which forms our metaphor repository. Metaphors stored in these tables can be compared and contrasted quickly using a query language that works on database tables. Since contents of our repository have been rigorously validated, comparisons made within and across languages allow us to gain knowledge about how peoples in these cultures view certain salient concepts such as BUREAUCRACY or GOVERNANCE. In addition, this repository can serve efforts for academics who wish to study metaphors in other domains of research.

2. Related Research

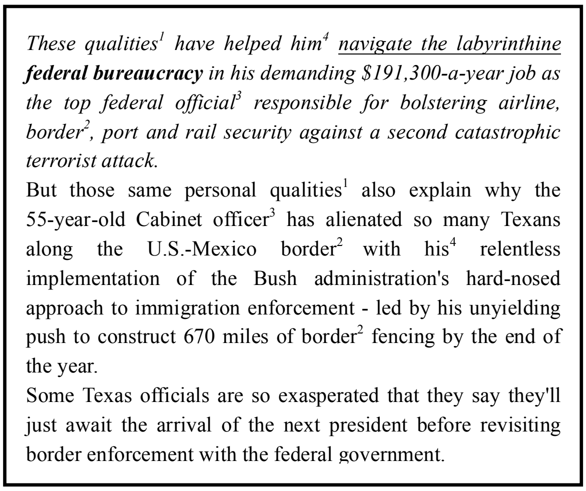
Approaches to metaphor detection are primarily based on semantic preferences, yielding limited scale, often hand designed systems (Wilks, 1975; Feldman & Narayan, 2004; Shutova & Teufel, 2010; Lakoff and Johnson, 1980). Knowledge-based approaches include MetaBank (Martin, 1998), a large knowledge base of metaphors empirically collected. Krishnakumaran and Zhu (2007) use WordNet (Felbaum, 1998) knowledge to differentiate between metaphors and literal usage. Gedigan et al (2006) identify a system that can recognize metaphor. However their approach is only shown to work in a narrow domain (e.g. Wall Street Journal). Such approaches are generally not robust and flexible enough to allow large scale extraction from unrestricted text sources and especially in languages that lack rich lexical resources. By contrast, our approach is fully automated to quickly populate a repository of metaphors, flexible enough to handle any domain of text, can be validated using empirical social science methods and can be utilized as a reference resource for a range of research fields.

The rest of this paper is organized as follows – in Section 3 we present briefly our approach to automatic metaphor detection, in Section 4 we explain how the system output is organized and stored in our repository. In Section 5, some exploratory cross-cultural comparisons are presented.

3. Our Approach

We have developed a data-driven computational approach to detect LMs that combines topical structure and imageability analysis to locate the candidate metaphorical expressions within text.

In Figure 1, we show some of the processing steps applied to an actual passage from our corpus.



These qualities^1 have helped him^4 navigate the labyrinthine ***federal bureaucracy*** *in his demanding $191,300-a-year job as the top federal official^3 responsible for bolstering airline, border^2, port and rail security against a second catastrophic terrorist attack.*

But those same personal qualities^1 also explain why the 55-year-old Cabinet officer^3 has alienated so many Texans along the U.S.-Mexico border^2 with his^4 relentless implementation of the Bush administration's hard-nosed approach to immigration enforcement - led by his unyielding push to construct 670 miles of border^2 fencing by the end of the year.

Some Texas officials are so exasperated that they say they'll just await the arrival of the next president before revisiting border enforcement with the federal government.

Figure 1: Excerpt from news article. Passage containing target concept highlighted in italics. The callouts ^{1}, ^{2} etc., indicate topic chains.

Given textual input, we first identify any sentence that contains references to Target concepts in a given Target Domain (Target concepts are elements that belong to a particular domain; for instance “government bureaucracy” is a Target concept in the “Governance” domain). We then extract a passage of length 2N+1, where N is the number of sentences preceding (or succeeding) the sentence with Target Concept.

Next, we employ dependency parsing to determine the syntactic structure of each input sentence. Topical structure and imageability analysis are then combined with dependency parsing output to locate the candidate metaphorical expressions within a sentence. For this step, we identify nouns and verbs in the passage (of length 2N+1) and link their occurrences – including repetitions, pronominal references, synonyms and hyponyms. This linking uncovers the topical structure that holds the narrative together. Our hypothesis is that metaphorically used terms are typically found outside the topical structure of the text. Any nouns or adjectives outside the main topical structure that also have high imageability scores and are dependency-linked in the parse structure to the Target Concept are identified as candidate *source relations*, i.e., expressions borrowed from a Source domain to describe the Target concept. In addition, any verbs that have a direct dependency on the Target Concept are considered as candidate relations. Our assertion is that any highly imageable word is more likely to be a metaphorical relation. We use the MRCPD (Coltheart 1981, Wilson 1988) expanded lexicon to look up the imageability scores of words not excluded via the topic chains. We have developed a method for expanding and creating a lexicon of imageability ratings, automatically, from existing resources (Liu et al., 2014).

The candidate relations identified in previous step are then used to compute and rank proto-sources. To determine this, we search for all uses of these relations in a balanced corpus and examine in which contexts the candidate relations occur. We search for their arguments in a balanced corpus, assumed to represent standard use of the language, and cluster the results. In the case of verb “navigate” we search a balanced corpus for the collocated words, that is, those that occur within a 4-word window following the verb, with high mutual information (>3) and occurring together in the corpus with a frequency at least 3. This search returns a list of words, mostly nouns in this case, that are the objects of the verb “navigate”, just as “federal bureaucracy” is the object in the given example. However, since the search occurs in a balanced corpus, given the parameters we search for, we discover words where the objects are literally navigated. Given these search parameters, the top results we get are generally literal uses of the word “navigate”. We cluster the resulting literal uses as semantically related words using WordNet and corpus statistics. Each such cluster is an emerging prototype source domain, or a proto-source, for the potential metaphor.

A ranked list of proto-sources from the previous step serves as evidence for the presence of a metaphor. If any Target domain elements are found in the top two ranked clusters, we consider the phrase being investigated to be literal. This eliminates examples where one of the most frequently encountered sources is within the target domain.

If neither of the top two most frequent clusters contains any elements from the target domain, we then compute the average imageability scores for each cluster from the mean imageability score of the cluster elements. If no cluster has a sufficiently high imageability score (experimentally determined to be >.50 in the current prototype), we again consider the given input to be literal. This step reinforces the claim that metaphors use highly imageable language to convey their meaning. If a proto-source cluster is found to meet both criteria, we consider the given phrase to be metaphorical.

In the current prototype system, we assign metaphors to one of three types of mappings. Propertive mappings – which state what the domain objects are and descriptive features; Agentive mappings – which describe what the domain elements do to other objects in the same or different domains; and Patientive mappings – which describe what is done to the objects in these domains. These are broad categories to which relations, with some exceptions, can be assigned at the linguistic metaphor level by the parse tag of the relation. Relations that take Target concepts as objects are usually Patientive relations. Similarly, relations that are Agentive take Target concepts as subjects. Propertive relations are usually determined by adjectival relations.

Affect of a metaphor may be positive, negative or neutral. Our affect estimation module computes an affect score taking into account the relation, Target concept and the subject or object of the relation based on the dependency between relation and Target concept. The expanded ANEW lexicon (Bradley and Lang, 2010) is used to look up affect scores of words. ANEW assigns scores from 0 (highly negative) to 9 (highly positive); 5 being neutral.

A sample LM with its associated proto-sources and other metadata stored in our repository is shown in Table 2.

Conceptual metaphors are posited based on groups of linguistic metaphors pointing to the same Source domain. We have realized conceptual source “spaces” for a number of conceptual source domains. These are created using a balanced corpus search for typical relations used with a high degree of frequency in a given Source Domain. Given that a source relation may invoke multiple source domains, we use a measure of inverted domain frequency to disambiguate between source domains. For example, the relation *kill* appears in the source domains of DISEASE, CRIME, ENEMY and MONSTER; however, using inverted frequencies from a balanced corpus search tells us that an LM with the relation *kill* should be associated with the source domain DISEASE with a greater probability than the other potential domains. In addition, the proto-sources extracted during LM detection are an additional source of evidence to disambiguate between potential source domains. A group of LMs invoking a CM is shown in Table 3.

We apply our algorithm to textual data in four languages – American English, Mexican Spanish, Russian Russian and Iranian Farsi. We detail in this paper the application of our approach to detection of metaphors using specific examples from the "Governance" and "Economic Inequality" domain. We have also collected data in "Democracy" domain in the aforementioned languages. However, our approach can be expanded to work on extracting metaphors in any domain, even unspecified ones. A detailed exposition on our algorithm and the novel techniques applied in its modules such as imageability analysis, have been published elsewhere (Strzalkowski et al., 2013, Broadwell et al., 2013).

4. Data in the Repository

We have organized data in the repository into 3 different layers – Data Layer, Linguistic Metaphor Layer and Conceptual Metaphor Layer. The organization allows us to write simple, efficient queries and enable comparisons across languages and cultures.

The Data Layer consists primarily of tables that store raw textual data, a corpus of documents and passages that have not been through any automatic processing except for being run for matches against a list of pre-determined keywords. This data is collected using our robust data acquisition process. We deploy automatic downloaders to search and download data from Internet, ensuring the copyright restrictions are met and terms of use are not violated. We aim to capture as much metadata as possible during the download process. Some metadata is readily available or intrinsic to the process – for instance: URL of download and date of download; some metadata capture requires additional processing of the downloaded content, date of publication of article being one of them.

The next layer in the repository stores linguistic metaphors extracted from the raw data and associated metadata. This is the Linguistic Metaphor layer in the repository. We have tables designed to store the linguistic metaphor instances, along with the clusters of terms that represent proto-sources, as well as the relations that give rise to these proto-sources. This layer contains all instances of passages that would be automatically identified as metaphorical as well as non-metaphorical by our system prototype. The utility of this layer is in its capability to allow searching through the relations associated with linguistic metaphors and their corresponding proto-source domains. Analysis of linguistic metaphors drives the conceptual metaphor determination process, which is efficiently supported by this layer.

The top-most layer in the repository is the Conceptual Metaphor layer. Conceptual metaphors are built on evidence from sets of linguistic metaphors. These are, therefore, modeled as a separate layer, stored in tables with their own associated metadata, such as correlations with other metaphors, sub-dimensions and links to their Source domains.

Data are stored separately in various layers of the metaphor repository for efficiency. Moreover, this design gives us the ability to track conceptual metaphors down to the actual data source (typically a web page), to examine the context in which particular instances of each metaphor occur i.e. the text containing a linguistic metaphor, the location where the content was published, what its genre is, and other associated metadata. In addition to various tables that contain textual data for processing and system output at various stages, we have stored the auxiliary data that is useful for system processing. These include lexicons such as ANEW (Bradley and Lang, 2010) for affective scores and MRC (Wilson, 1998) for imageability scores. Evaluation data collected through social science validation experiments for the project is also being stored in the repository. This allows us to easily access human assessments of specific examples and make comparisons to system output for analyses.

In Table 1, we show the amount of data in the repository for all four languages of interest. We are continually updating and inserting data into the repository, the numbers shown in Table 1 represent status of repository at the time of writing this paper. CMs are typically invoked using many LMs with distinct relations that belong to the same Source Domain; hence the number of CMs in our repository is in an emergent stage.

	English	Spanish	Russian	Farsi
Number of Documents	1,048,294	478,032	161,989	45,680
Number of Passages	6,624,100	6,031,328	9,958,345	328,781
Number of Processed Passages	189,862	17,261	14,873	23,503
Number of Linguistic Metaphors	99273	2939	1979	1543
Number of Conceptual Metaphors	49	80	4	12

Table 1: Amount of information stored in the repository for all four languages, including documents, passages and metaphors

We keep refining our algorithms as we collect and analyze validation data from human assessments. This leads to fluctuation in the number of linguistic and conceptual metaphors. Once our processing is complete, the repository will reflect the true extent of linguistic and conceptual metaphor data available across the documents and passages we have collected.

In Table 2, we show a sample row from our Linguistic Metaphor layer. For each passage that we process, we generate and insert a row in the Relation table of this layer. The top three proto-source clusters generated from balanced corpus search and their sample elements are shown, in addition to computed Affect (positive) and the type of relation (Patientive). The relation *navigate* indicates a way of dealing with or affecting the Target Concept (*federal bureaucracy*).

In Table 3, we show a group of four LMs and the potential conceptual source domains they point to. In Table 4, a sample CM invoked from LMs in Table 3 is shown. This sample group is presented illustratively; in practice a larger group of LMs (containing at least 10 or more distinct relations) may be required to invoke a CM. Affect of CM is calculated as the prevailing affect of the LMs comprising it.

LM ID	111
Source Relation Name	navigate
Target Concept	federal bureaucracy
Sentence	These qualities have helped him….
Top 3 candidate source clusters	1. [way, tools] 2. [terrain, patch] 3. [maze, labyrinth]
Affect	Positive
Type of Relation	Patientive

Table 2: An instance of Linguistic Metaphor. Proto-sources are clusters that are revealed by searching a balanced corpus for things that can literally be *navigated*. The affect computed for this metaphor is Positive and type of relation is Patientive.

LM ID	55
Source Relation Name	tangled in
Target Concept	federal bureaucracy
Sentence	His attorney described him as a family man who was lied to by a friend and who got tangled in *federal bureaucracy* he knew nothing about.
Affect	Positive
Type of Relation	Patientive
Potential Source Domains	MAZE
LM ID	**111**
Source Relation Name	maze of
Target Concept	federal bureaucracy
Sentence	The chart, composed of 207 boxes illustrates the maze of federal bureaucracy that would have been created by then-President Bill Clinton's relation health reform plan in the early 1990s.
Affect	Positive
Type of Relation	Propertive
Potential Source Domain	MAZE
LM ID	**110**
Source Relation Name	navigate
Target Concept	federal bureaucracy
Sentence	"Helping my constituents navigate the federal bureaucracy is one of the most important things I can do," said Owens.
Affect	Positive
Type of Relation	Patientive
Potential Source Domain	MAZE; SHIP
LM ID	**111**
Source Relation Name	navigate
Target Concept	federal bureaucracy
Sentence	These qualities have helped him navigate the labyrinthine federal bureaucracy in his demanding $191,300-a-year job as the top federal official[3] responsible for bolstering airline, border, port and rail security against a second catastrophic terrorist attack.
Affect	Positive
Type of Relation	Patientive
Potential Source Domain	MAZE; SHIP

Table 3: A group of Linguistic Metaphors pointing to potential source domain MAZE.

CM ID	Target Concept	Source Domain	Prevailing Affect	LM IDs used as evidence
3	BUREAUCRACY	MAZE	POSITIVE	[55, 232, 110, 111]

Table 4: A Conceptual Metaphor invoked from LMs shown in Table 3. Affect is calculated as the prevailing affect of LMs that form the CM.

5. Sample Comparisons

We have constructed validation tasks that are aimed at performing evaluation of metaphor extraction accuracy, and hence the integrity of the contents of the Repository. Native language experts recruited through Amazon Mechanical Turk (at least 30 subjects that meet a variety of filters such as grammar proficiency), as well as trained linguists undertake various validation tasks. The judgments so collected are tested for reliability and validity. Reliability among the raters is computed by measuring intra-class correlation (ICC) (McGraw & Wong, 1996; Shrout & Fleiss, 1979). ICC for our validation data is above 0.85 on average across all four languages, where a coefficient value above 0.7 indicates strong reliability. The accuracy on subset of data in the repository that has been validated is 75% on average across all four languages for LM detection; for CM detection the accuracy is 82%.

Using this empirically validated data, we are able to posit certain claims using multi-cultural metaphors in our repository. Figures 2 and 3 show some of the sample cross-cultural comparisons that can be made using various tables in the repository. In Figure 2, we show the source domain preferences for a subset of conceptual metaphors from our repository. We see that for some languages, certain source domains are preferred. Whilst

American English metaphors seem to be distributed evenly across the domains, source domains such as MAZE for Farsi and ENEMY for Spanish show up with a higher degree of frequency than other source domains. In Figure 3, we show the types of relations across languages based on mapping for certain Target Concepts in the domain of Governance. We classify (automatically) each relation according to whether it is Agentive (the way Target Concept acts or effects other things), Patientive (the way Target Concept is dealt with or affected) or Propertive (the way Target Concepts appears, feels, smells etc.). These mappings give additional insights about the cultural differences. For example, Farsi metaphors do not have as much evidence for the Patientive type of relation as do other languages; which may mean that the culture does not use metaphorical language to talk about ways to deal with the Target Concept of Governance. In Russian metaphors, on the other hand, the Propertive type of mapping is absent which may mean that preference to talk about Governance concepts in a propertive manner is proportionally lower in this culture.

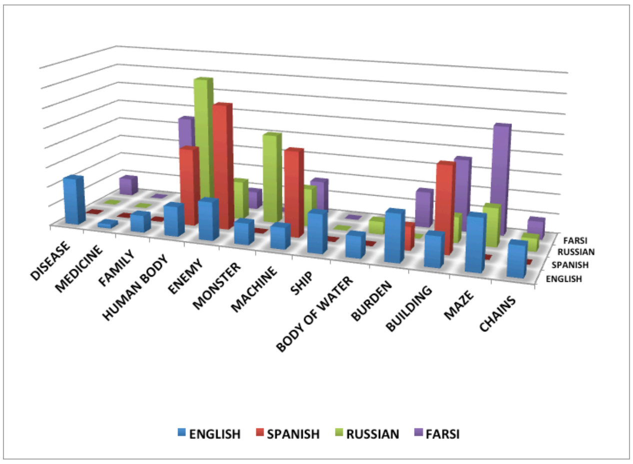
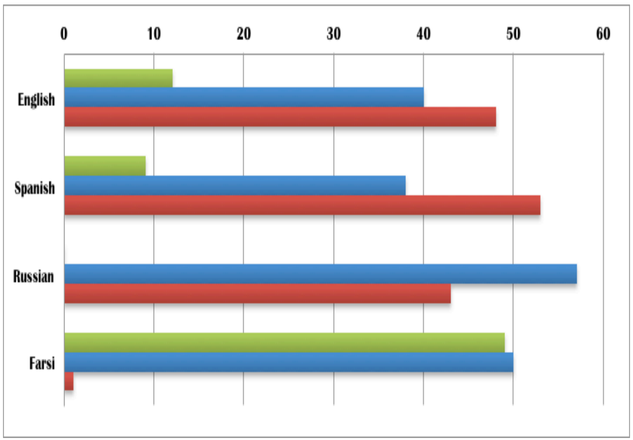


Figure 2: Source Domain preferences across cultures. Y-axis represents the proportion of preference (%) for a particular source domain for a language. For example, ENEMY is the preferred Source Domain in Spanish, MAZE in Farsi.

In Figure 4, we show the source domain preferences for the target concept of Poverty across three languages – English, Spanish and Russian. These data show whether there are Source Domains (SD) that may be preferred for a Target Concept in certain cultures over other cultures. The Y-axis represents Source Domain preferences in each language combined as a stacked column. Hence, the column will not add up to 100% for each SD. Also note that some SDs that had less % of metaphors are absent from the graph, for ease of presentation. We can see that POVERTY is a PHYSICAL BURDEN is quite prevalent in Spanish, and not in other languages. On the other hand, POVERTY is a DISEASE metaphors are present in all three languages under consideration.



- ***Propertive*** **: the way Target appears: *looks, smells, sounds, feels*, etc.**
- ***Agentive*** **: the way Target acts or affects other things: *kills, chokes*, etc.**
- ***Patientive*** **: the way to deal with it or to affect it: *tame it, protect from*, etc.**

Figure 3: Source mapping across cultures. X-axis represents the proportion of preference (%) for a particular type of mapping. For example, in Farsi, the Patientive type of mapping is quite low in proportion, whereas it is the dominant type of mapping for Spanish.

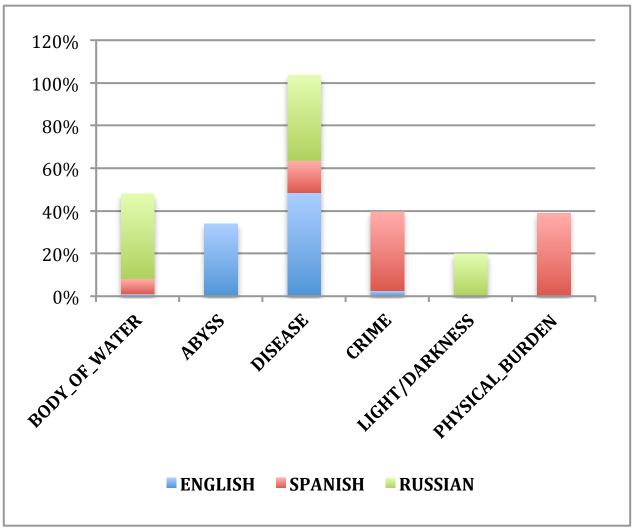


Figure 4: Source Domain preferences for Target Concept of POVERTY in English, Spanish and Russian metaphors with high imageability

6. Conclusion

In this article, we described a resource that allows academic activity in the field of metaphor research. Our primary goal in creating this repository is to derive cross cultural comparisons via the use of metaphors in language amongst the four cultures we are interested in – America, Mexico, Russia and Iran. Other research efforts can also benefit from this resource, for instance, researchers who wish to recognize metaphors of the same kind or pattern. We continue to update our repository as part of our main project.

7. Acknowledgements

This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0024. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

8. References

Bradley, M.M. & Lang, P.J. 2010. *Affective Norms for English Words (ANEW): Instruction manual and affective ratings.* Technical Report C-2. University of Florida, Gainesville, FL.

Broadwell, George A., Umit Boz, Ignacio Cases, Tomek Strzalkowski, Laurie Feldman, Sarah Taylor, Samira Shaikh, Ting Liu, Kit Cho, aand Nick Webb. 2013. Using imageability and topic chaining to locate metaphors in linguistic corpora. In Ariel M. Greenberg, William G. Kennedy, Nathan D. Bos and Stephen Marcus, eds. *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction SBP 2013.*

Fellbaum, C. editor. 1998. WordNet: An Electronic Lexical Database (ISBN: 0-262-06197-X). MIT Press, first edition.

Feldman, J. and S. Narayanan. 2004. Embodied meaning in a neural theory of language. *Brain and Language*, 89(2):385–392.

Gedigian, M., Bryant, J., Narayanan, S., & Ciric, B. (2006). Catching Metaphors. *Proceedings of the Third Workshop on Scalable Natural Language Understanding ScaNaLU 06* (pp. 41-48). Association for Computational Linguistics.

Krishnakumaran, S. and X. Zhu. 2007. Hunting elusive metaphors using lexical resources. In Proceedings of the Workshop on Computational Approaches to Figurative Language, pages 13–20, Rochester, NY.

Lakoff, George and Johnson, Mark. 1980. *Metaphors We Live By*. University Of Chicago Press.

Liu, Ting, Kit Cho, George Aaron Broadwell, Samira Shaikh, Tomek Strzalkowski, John Lien, Sarah Taylor, Laurie Feldman, Boris Yamrom, Nick Webb, Umit Boz and Ignacio Cases. 2014. Submitted to LREC 2014.

Martin, James. 1988. A Computational Theory of Metaphor. *PH.D. Dissertation*

McGraw, K. O., & Wong, S. P. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46.

Shrout, P. E., & Fleiss, J. L. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86* (2), 420-428.

Shutova, E. and S. Teufel. 2010. Metaphor corpus annotated for source - target domain mappings. *In Proceedings of LREC 2010, Malta.*

Strzalkowski, Tomek., George Aaron Broadwell, Sarah Taylor, Laurie Feldman, Samira Shaikh, Ting Liu, Boris Yamrom, Kit Cho, Umit Boz, Ignacio Cases and Kyle Elliot. 2013. Robust Extraction of Metaphor from Novel Data. In *Proceedings of the First Workshop on Metaphor in NLP at the North American Association of Computational Linguistics Conference (NAACL-2013) Atlanta, USA*.

Wilks, Yorick. 1975. Preference semantics. *Formal Semantics of Natural Language*, E. L. Keenan, Ed. Cambridge University Press, Cambridge, U.K., 329--348.

Wilson, M.D. 1988. The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers, 20(1), 6-11.*