# Language-Independent Ensemble Approaches to Metaphor Identification

**Jonathan Dunn[1] Jon Beltran de Heredia[2] Maura Burke[1] Lisa Gandy[3] Sergey Kanareykin[4] Oren Kapah[5] Matthew Taylor[2] Dell Hines[2] Ophir Frieder[6] David Grossman[6] Newton Howard[7] Moshe Koppel[5] Scott Morris[1] Andrew Ortony[8] Shlomo Argamon[1]**

[1] Illinois Institute of Technology, Chicago, IL
[2] Behavioral Media Networks, Greenwich, CT
[3] Central Michigan University, Mt. Pleasant, MI
[4] Brain Sciences Foundation, Providence, RI
[5] Bar Ilan University, Ramat Gan, Israel
[6] Georgetown University, Washington, DC
[7] Massachusetts Institute of Technology, Cambridge, MA
[8] Northwestern University, Evanston, IL

## Abstract

True natural language understanding requires the ability to identify and understand metaphorical utterances, which are ubiquitous in human communication of all kinds. At present, however, even the problem of identifying metaphors in arbitrary text is very much an unsolved problem, let alone analyzing their meaning. Furthermore, no current methods can be transferred to new languages without the development of extensive language-specific knowledge bases and similar semantic resources. In this paper, we present a new language-independent ensemble-based approach to identifying linguistic metaphors in natural language text. The system's architecture runs multiple corpus-based metaphor identification algorithms in parallel and combines their results. The architecture allows easy integration of new metaphor identification schemes as they are developed. This new approach achieves state-of-the-art results over multiple languages and represents a significant improvement over existing methods for this problem.

Metaphor is a ubiquitous feature of human language and, as a result, true natural language understanding requires the ability to identify and understand metaphors. As shown by Black (1954), metaphor is far more than an ornamental figure of speech. Rather, it permeates many aspects of language and may serve a key role in structuring conceptual representation and inference (Carbonell 1980; Lakoff and Johnson 1980, 1999). A great deal of work in linguistics and cognitive science has been done over the last forty years to explicate the relationships between conceptual and linguistic metaphors (Lakoff 1993; for an overview, Steen 2007). From the computational perspective, however, even the foundational problem of identifying linguistic metaphorical expressions remains difficult and only partially solved, setting aside the more difficult problems of analyzing metaphorical meanings and finding conceptual metaphors. Furthermore, solutions that have been proposed typically rely on large amounts of manually curated knowledge representation and are generally specific to a single language.

This paper presents a new approach to identifying metaphorical expressions such as "bloated management," "raising the debt ceiling," or "falling into poverty." The system is fundamentally language agnostic and uses very small amounts of manually built semantic representations, in order to improve the application of the system to new languages.

Developing a system that can be easily and directly applied to new languages is key to the long-range goal of developing cognitive computing systems that can adapt to different languages and cultures. The reliable cross-linguistic identification and analysis of linguistic metaphors is an essential component of such a system because (i) metaphor permeates many aspects of language and (ii) metaphor varies widely across cultures and languages.

This paper presents a system for metaphor identification which gives (i) a significant improvement in accuracy for linguistic metaphor identification for English, (ii) improved results for Spanish, Russian, and Farsi metaphor identification, and (iii) a modular architecture that will more easily enable integration of multiple metaphor cues into a single framework, increasing accuracy in a language and domain independent fashion.

## Related Work

Two main strands of previous work need to be considered: linguistic metaphor identification, the focus of this paper, and the related task of conceptual metaphor identification. The first task is to find linguistic elements which have a metaphoric meaning; the second task is to divide these linguistic metaphors into groups of conceptual metaphors which reflect underlying patterns of thought. For example, the phrases "demolish his ideas", "rebuild the theory", and "support the hypothesis" are linguistic metaphoric expressions which can be said to reflect the single conceptual metaphor IDEAS ARE BUILDINGS. Shutova, et al. (2013) present a system which collapses both tasks into a single


Corresponding author: Shlomo Argamon, argamon@iit.edu

algorithm. First, a knowledge-base is created by clustering nouns and verbs according to their contexts of use, so that each noun or verb is a member of a given cluster (i.e., domain). Second, an input text is processed to find all grammatical relationships between a verb and its direct or indirect objects. The result is that grammatical relationships between nouns and verbs are treated as conceptual connections between the domains represented by the noun and verb. New instances are found by matching them with available seed instances.

This approach represents the simplest operationalization of metaphor, combining linguistic and conceptual metaphors into a single structure, with linguistic metaphors directly and explicitly encoding conceptual metaphors. Mason (2004) similarly uses an explicit linguistic mapping to stand in for a conceptual mapping, using selectional preferences across texts from different domains to discover metaphoric mappings, an idea first employed by Wilks (1978).

Several linguistic metaphor identification systems train a classifier on feature vectors representing the input text, using as properties semantic similarity (Li and Sporleder 2010a,b), abstractness (Turney, et al. 2011), domain and event-status (Dunn 2013a,b), and a combination of abstractness, semantic category, and named entity features (Tsvetkov, et al. 2013). Hovey, et al. (2013) use vectors representing contextual similarity and combine them with several tree representations (e.g., POS tags) to classify using a forest of tree kernels. These approaches are more sophisticated than the first in that they posit more general properties which should distinguish metaphor from non-metaphor, but they do not attempt to identify conceptual metaphors.

The semantic signatures approach to identifying conceptual metaphors (Mohler, et al. 2013; Bracewell, et al. 2013) skips the search for linguistic metaphors and looks instead for sentences in which elements of the desired source and target domains are both present (e.g., GOVERNANCE and BUILDINGS). Thus, the system will find only instances of the desired conceptual metaphors, allowing a focused search. Semantic signatures are represented using clusters of related words extracted from WordNet, Wikipedia, and large corpora. A similar approach (Strzalkowski, et al. 2013) starts by searching for a given target concept, represented using lists of keywords. Given the passage within which the target concept occurs, source candidates are found by looking for text that both falls outside of the topic of the text and has a high imageability rating. This approach again uses the search for conceptual mappings to stand in for the search for linguistic metaphors.

A final approach (Gandy, et al. 2013; Neuman, et al. 2013; Assaf, et al. 2013) separates the tasks of identifying linguistic metaphors and extracting conceptual metaphors from these linguistic metaphors, thus achieving greater flexibility at both levels. Linguistic metaphors are identified using abstractness measures related to the approaches discussed above. Candidates for conceptual metaphors are generated from nominal analogies based on the linguistic metaphors and filtered by a set of constraints which ensure that there is a plausible relationship both between the conceptual and the linguistic metaphors and between the source and the target concepts.

The current work both includes and improves upon the work reviewed here in that it has the potential to allow all of these identification algorithms to be implemented side-by-side in a single framework and their results combined in order to determine which algorithm performs best on a given input. This is a significant improvement because linguistic metaphors have been shown to come in several different forms (Dunn 2013b), with individual algorithms performing well on a single form but failing on others. Thus, the approach presented below is designed to have the potential to incorporate the advantages of each algorithm while mitigating the accompanying failures of each.

## External Resources

Each of the identification algorithms described in this paper uses the same set of statistical, syntactic, and semantic resources. As noted above, a central goal is to minimize the use of manually constructed resources in order to ease the application of the methods to new languages and domains. We describe three sets of resources here: background corpora, syntactic parsing, and a lexical semantic database. The background corpora is simply baseline textual data for a given language without annotations. The syntactic parser and lexical semantic databases for each language, however, are manually designed (although the lexical semantic database has the same design across languages).

### Background Corpora

Background corpora are used to provide base estimates of the relative frequency of particular pairs of words. These are n-gram corpora (up to n=3 currently), with each word form stored with its associated lemma and part-of-speech. Our methods are mainly concerned with determining the frequency of lemma collocations within given windows. All corpus information is stored in a relational database, which responds to a normalized design in order to guarantee consistency. An extra denormalized table was added to improve performance because of the large size of the datasets. This table also directly includes lexical semantic information as described below, also to improve performance.

Table 1: Size of Background Corpora by number of unique unigrams, bigrams, and trigrams.

| Language | Unigrams | Bigrams | Trigrams |
|---|---|---|---|
| English | 9.2 million | 85.3 million | 255.3 million |
| Spanish | 3.4 million | 18.4 million | 78.9 million |
| Russian | 6.4 million | 57.7 million | 67.0 million |
| Farsi | 8.0 million | 92.7 million | 329.0 million |

Four databases were compiled holding this information, one for each language: English, Spanish, Russian and Farsi. The English, Spanish, and Russian databases are based on Google N-Grams (post-1970), and the Farsi database is derived from a collection of scraped Farsi blogs together with the pre-existing Hamshahri and Dadegan tagged corpora.

The sizes for the corpora built for each language are shown in Table 1.

### Syntactic Parsing

A preprocessing step for all texts in which metaphors are to be identified is to tokenize the text and produce a syntactic dependency parse of each sentence with part-of-speech (POS) tags. We map POS tags to a small standard generic set of tags (noun, verb, adjective, etc.) to ensure consistency in processing across languages. For English and Spanish we use OpenNLP for POS tagging and MaltParser for parsing, for Russian we use Freeling and MaltParser, and for Farsi we use the Stanford POS tagger and MaltParser trained on the Dadegan corpus.

### Lexical Semantic Resources

We use several basic resources to provide lexical semantic information for nouns. First, values from the MRC Psycholinguistic Database (Wilson 1988) are used for estimates of Concreteness, Familiarity, and Imageability of common nouns; these are extended to nouns not in the MRC by the statistical association method of (Turney et al. 2011), thus creating a robust database of lexical semantic information for nouns in English. Values for Concreteness, Familiarity, and Imageability for non-English languages were computed by using Bing's machine translation service to find the value for the corresponding English noun. Second, we identify a set of 128 basic semantic categories from WordNet, as those including at least 100 nouns among the 5,000 most common nouns in the Corpus of Contemporary American English. Similar sets of semantic categories were constructed for the other target languages, though there were many fewer categories, as the WordNets in those languages are less well-developed. As a result, the system also depends upon a manually constructed set of semantic categories for each language; WordNet provides this for many languages, but an alternative would be needed for languages which lack a well-developed WordNet equivalent.

## Heterogeneous Metaphor Identification

Our ensemble approach seeks to accommodate a multiplicity of different metaphor identification methods while optimizing overall accuracy. The basic design concept is that metaphor identification is performed by a network of identification modules, connected together in a reconfigurable graph. There are two main benefits to this design: first, composing and configuring modules allows maximum flexibility to configure the system to perform specific tasks; second, the system can be easily improved by implementing new modules as long as they conform to the existing interfaces.

### Candidate Extraction

Once the text is processed and parsed, the system needs to extract pairs of words which are syntactically related in forms that may imply a metaphoric meaning. The simplest forms are adjective-noun pairs, verb-noun pairs, or noun-noun pairs, but the system allows multi-word phrasal components to be extracted as candidate pair items, thus resulting in complex lexicalizations of metaphor. Rather than just relying on proximity in the input text, the full parse tree is used to allow extraction of pairs of terms that are related due to language-specific constructs. For example, Russian often features an elided verb and Farsi has the mosnad dependency, both of which may imply a specific relation between a noun and an adjective which should be tested for possible metaphoric meaning.

The system uses custom-designed rules to extract appropriate pairs of candidate terms from text in supported languages. These rules match subtree-patterns in syntactic dependency trees. Thus, part of the candidate extraction process can be automatically transferred across languages (e.g., adjective-noun pairs within a certain number of words), but another part of the candidate extraction process must be manually constructed for each new language depending on the grammatical structure of that language.

### Modular LMID Classifier

The core identification module links together specific metaphor classifiers in a directed acyclic graph (DAG) model. This design gives flexibility in synthesizing the best practices in metaphor identification into one system while rapidly developing and integrating improvements over time. Each node in the graph contains a routine that provides judgments about candidate pairs being tested. The network can be easily configured to contain different combinations of two node types, Classifiers and Combiners. Classifiers return a decision about each candidate expression (or "IDK" = "I Don't Know") and a degree of confidence in the decision (which is not currently used in aggregating the results of multiple classifiers). The currently implemented algorithms in the system are described below.

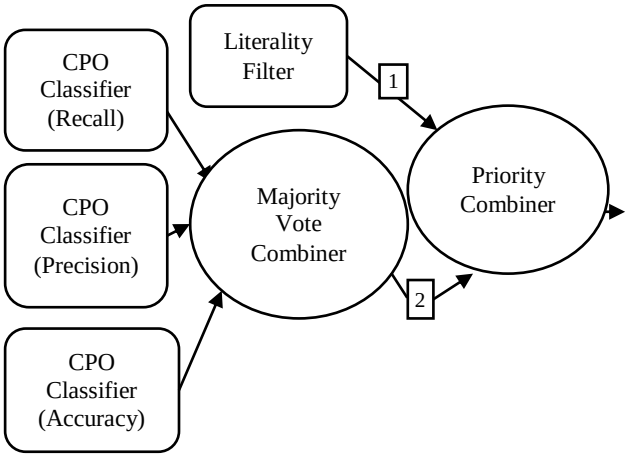


Figure 1: The LMID directed acyclic graph.

**Category Profile Overlap Classifier** The Category-Profile Overlap (CPO) Classifier identifies metaphors by looking at the amount of overlap between source and target words in terms of category information from the background corpus. For each source and target, the algorithm (i) checks the background corpus for concrete associated nouns, (ii)

determines the most common semantic categories for those nouns, and (iii) compares the overlap of these categories between the source and target, identifying a metaphor when the level of overlap is low. In a sense, this is a statistical version of the selection preference violation criterion (e.g., Wilks 1978, Mason 2004).

Pseudo-code for the algorithm is as follows:

1. Find the lemma of the source term (S), and the target term (T)
2. Let NOUNS be the N noun lemmas with highest pointwise mutual information (PMI) with S, with PMI greater than a threshold. If S is not found in the background corpus, return IDK
3. Let CONCNOUNS be the M most concrete words in NOUNS, based on precalculated concreteness scores
4. Let SCATS be the N semantic categories that appear for the most nouns in CONCNOUNS
5. Let TCATS be all semantic categories associated with T. If T is not found in the background corpus, return IDK
6. Calculate the overlap between SCATS and TCATS according to a predefined method (see below)
7. If the overlap is less than a given threshold provided, R, then return METAPHOR, else return LITERAL

Table 2: Parameter settings for CPO Classifier nodes.

| Parameter | Accuracy | Precision | Recall |
|---|---|---|---|
| PMI Threshold | 2 | 2 | 2 |
| Max Overlap | 0.4 | 0.08 | 0.1 |
| Num. Categories | All | 10 | 15 |
| Num. Nouns | 1,000 | 1,000 | 1,000 |
| Num. Concrete Ns | 100 | 100 | 100 |
| Overlap Type | Weighted | RatioB | RatioB |

The list of parameters of the algorithm is given in Table 2, along with the three sets of parameter values used in our experiments that were tuned for, respectively, high precision, high recall, and high accuracy on a development corpus in English. Each of the types of overlap are calculated as shown in Table 3.

Table 3: Calculating Overlap Types for the CPO.

| Overlap Type | Equation |
|---|---|
| Weighted | Dw / (Sw * Tw) |
| RatioA | Intw / Sw |
| RatioB | Dw / Sw |
| Union | \|Intersection(S,T)\| / \|Union(S, T)\| |
| Max | \|Intersection(S,T)\| / Max(\|S\|, \|T\|) |

**Literality Filter Classifier** There are many common words and phrases which are virtually certain to indicate that a phrase is non-metaphorical. Common examples include numbers ("14 books"), proper nouns, geographic or temporal adjectives ("Arizonan weather" or "later arrival"), and relational process verbs (e.g., "is" and "has").

This classifier contains lists of words in such categories for each language, and also checks if candidate terms contain numerals or are proper nouns. If listed words or numerals or proper nouns are found in the candidate, the classifier returns LITERAL; otherwise, it returns IDK.

In some situations, common words or proper nouns may be used metaphorically; for example, "Ohio is the Arizona of the Midwest," "My brother is a real Marlon Brando," and "I think his car has only 3 cylinders." Such cases are not currently handled by the system and would be classified as LITERAL by the Literality Filter Classifier.

**Voting Combiner** This node classifies candidates by looking at the results of several other nodes. It takes in, as configuration, a list of nodes to include. When performing classification, it returns the value of the majority; for instance, if it has a list of five nodes and three of them return a classification of METAPHOR, this node will also return a classification of METAPHOR (3 beats 2). Future work will include a meta-classifier which attempts to learn which of the classifiers performs best on a given type of candidate expression.

**Priority Combiner** This node takes input from a set of nodes, each with a priority. It chooses for each candidate the highest-priority input answer which is not IDK.

## Experimental Results

The VU Amsterdam Metaphor Corpus (Steen, et al. 2010) is the closest thing available to a shared data set for linguistic and conceptual metaphor. The corpus consists of 200,000 words from the British National Corpus, divided into four genres: Academic, Fiction, News, and Conversation (spoken). This wide representation allows the comparison of performance across different registers and levels of formality. The corpus was annotated for metaphor by five trained linguists with a very high level of inter-rater agreement. Further, each instance of metaphor was sub-categorized as follows: (i) a metaphorically-used metaphor related word; (ii) a literally-used metaphor related word; (iii) personification; (iv) double metaphor (e.g., a word that is a part of two different metaphors); and (v) ambiguous text which raters were not sure was metaphoric or non-metaphoric (these account for only 7% of the test corpus).

A previous study (Dunn, 2013b) evaluated four linguistic metaphor identification systems on this corpus: an abstractness measurement system (Turney, et al. 2011), a semantic-similarity measurement system (Sporleder and Li 2010a,b), a source-target mapping system (Shutova, et al. 2013), and a domain interaction system (Dunn 2013a). The results discussed here compare the performance of the modular system described above with the performance of these systems. For the purposes of these results, the evaluation is conducted at the sentence-level, with the assumption that a sentence containing a metaphorically-used word is metaphoric. The evaluation only considers those sentences used in (Dunn 2013b),

for which all the systems had sufficiently robust representation, a total of 8,887 sentences.

Table 4 gives results computed using the entire corpus (i.e., not with separate models built for each genre, as done by Dunn 2013b; our results thus vary slightly from his). These results show both the performance of the modular system on the best available test corpus and in relation to four varied existing methods. The Modular system is clearly the highest performing, with an F1 that is 0.2 higher than the nearest competitor (0.703 vs. 0.506). Further, the Modular system has a good balance between precision (0.704) and recall (0.713), whereas other systems tend to raise one at the expense of the other (for example, the Source-Target system has a precision of 0.521, but a recall of only 0.183). One of the findings of the original study is that different algorithms perform well on some linguistic metaphors but not on others. This higher performance of the Modular system is what we would expect, then, because it allows a multi-faceted definition of linguistic metaphor.

Table 4: Results on VU Amsterdam Metaphor Corpus, By System (8,887 instances each).

| System | Precision | Recall | F1 |
|---|---|---|---|
| Modular | 0.704 | 0.713 | 0.703 |
| Domain Interaction | 0.501 | 0.565 | 0.502 |
| Source-Target | 0.521 | 0.183 | 0.411 |
| Similarity | 0.504 | 0.596 | 0.506 |
| Abstractness | 0.503 | 0.421 | 0.486 |

Table 5 shows a more detailed view of the Modular system's results. Genre here represents texts with different degrees of formality, different registers, and different conventions, allowing us to see where the system performs best. The highest F1 for our method (0.797) is on the News genre and the lowest (0.589) on the Academic genre. The Academic genre has a low precision (0.623) but a high recall (0.875); the Conversation genre (the only spoken genre) has the reverse: a low recall (0.372) but a high precision (0.713). One likely reason for this divide is that the Academic genre is the most formal and conventional, while the Conversation genre is the least formal and the least conventional. Thus, the Conversation genre likely contains many unusual and new linguistic metaphors which may be more difficult to detect. Also, the Conversation genre has shorter sentences with less syntactic structure, perhaps limiting the number of candidate expressions that are found (e.g., because there are fewer syntactic relations). The News genre tends to contain more stylistic metaphors (e.g., clever phrasings), which are easier to detect than the more pervasive and conventionalized metaphors which are common in the Academic genre.

Table 6 shows more detailed results of the Modular system by the sub-type of metaphor. The two highest categories are Personification and Double Metaphor, with 86.4% and 85.4% of linguistic metaphors found, respectively. Performance is highest on these metaphors because the notion of category overlap applies especially strongly to personifications, in which an inanimate object takes on human properties. Double metaphors are relatively easy to detect because there are two metaphors present, doubling the chance of identification (although, on the other hand, making it more difficult to identify the conceptual metaphor). The lowest performance is on the catch-all category of Metaphorically-used Metaphor-Related words. This does not tell us much because this is simply the category for everything which is not a personification or a double metaphor.

Table 5: Results of Modular System on VU Amsterdam Metaphor Corpus, By Genre.

| Subset | Precision | Recall | F1 |
|---|---|---|---|
| Total | 0.704 | 0.713 | 0.703 |
| Genre: Academic | 0.623 | 0.875 | 0.589 |
| Genre: Fiction | 0.694 | 0.642 | 0.680 |
| Genre: News | 0.832 | 0.806 | 0.797 |
| Genre: Conversation | 0.713 | 0.372 | 0.715 |

Table 6: Results of Modular System on VU Amsterdam Metaphor Corpus, By Sub-Type.

| Subtype | Found | Missed | % |
|---|---|---|---|
| Metaphor-Related Word | 2,387 | 1,154 | 67.4% |
| Personification | 448 | 70 | 86.4% |
| Double Metaphor | 106 | 18 | 85.4% |
| Ambiguous Metaphor | 533 | 164 | 76.4% |

## Error Analysis

To help understand the weaknesses in the system, we examine in detail several examples of false positives. First, some errors are caused by the linguistic resources. For example, the sentence "His problem will be that Mrs. Thatcher might decide to do it first and then she will garner the votes" was identified as a metaphor, with the source being *decide* and the target being *votes*. Such a candidate expression may have been metaphoric, as in "Money decides the votes, not reason." However, in this false positive, *votes* is not in a syntactic relationship with *decide*, and so it should not have been a candidate in the first place. A similar example is "Name badges are worn by staff," in which the extracted source is *worn* and the extracted target is *staff*; however, these are again not in a direct dependency relationship. "Worn staff" is metaphoric, but that is not an actual candidate expression in this sentence.

Another cause of false positives is the many ways of naming or referring to a real-world entity. For example, the sentence "The Bhopal accident killed 2000 and injured 200,000 more" is falsely identified as metaphoric with the source *killed* and the target *accident*. On the other hand, the phrase "Toxic gases killed..." would not have been identified as metaphoric, and the phrase "The toxic management killed..." would have actually been metaphoric. The difficulty is in knowing what the phrase is referring to. For example, the very same candidate source and target would have been

metaphoric in the sentence "My spaghetti sauce accident killed the ambiance of the evening." Thus, there are many cases in which more information is required than simply the extracted candidate expression with its source and target elements. In other words, more context is necessary to determine the referents of the candidate expression.

One cause of false negatives is the failure to extract candidate expressions, especially in cases where the metaphoric material is spread across a sentence and not explicitly present in a single syntactic structure. For example, in the sentence "The admission which the Prime Minister wrung from him could hardly be said to have been grudging," the metaphoric relation is between *wrung* and *admission*. However, in the structure of the sentence, *admission* is dependent on *wrung* only as the antecedent of *which*; this is made even more difficult by the OSV order of the relative clause.

### Cross-Linguistic Results

To test the system cross-linguistically, we evaluated it on comparable corpora in English, Spanish, Russian, and Farsi. These corpora were constructed by selecting sentences from web-scraped blogs, automatically extracting candidate expressions, and then having each candidate annotated as metaphoric or non-metaphoric by two native speakers, who discussed the annotations until reaching agreement.

Table 7: Results across languages.

| Language | Sentences | Precision | Recall | F1 |
|---|---|---|---|---|
| English | 5,000 | 0.310 | 0.750 | 0.440 |
| Russian | 999 | 0.584 | 0.327 | 0.538 |
| Spanish | 236 | 0.603 | 0.251 | 0.614 |
| Farsi | 1,296 | 0.837 | 0.175 | 0.833 |

Results are given in Table 7 for metaphor identification using CPO with Accuracy optimized settings. While the non-English results are not as good overall as in English, they are among the best for this task for each of these languages. They therefore show how a language-independent system can attain good accuracy for multiple languages with no language-dependent tuning beyond gathering a corpus and developing relatively small lexical resources. It should be noted that the results for this evaluation of English are lower than for the previous evaluation; the previous evaluation is more reflective of actual performance while this one inherits certain errors from the annotation scheme employed, errors which the VU Amsterdam Metaphor Corpus, because it was manually produced and corrected, does not face. More importantly, this evaluation also uses only one instance of the CPO instead of the larger system. This difference in performance shows the importance of the multi-faceted approach to metaphor identification.

### Degrees of Language Independence

These cross-lingual results bring up the issue of degrees of language independence: how well does the underlying algorithm work across languages, and how easily is it expanded to new languages? No system is entirely language independent, insofar as NLP resources like syntactic parsing and POS tagging are required. Thus, the baseline expectation is that some basic processing is available for a new language. This leaves two language dependencies: first, the lexical semantic properties; second, the candidate extraction patterns. The lexical semantic properties in the current implementation require some language-specific resources, specifically a machine-translation system for expanding the Concreteness, Familiarity, and Imageability values for nouns and a WordNet equivalent for determining the inventory and membership of semantic categories in the language.

Currently, these are manually designed resources for each new language (or rather, require such manually designed resources). This requirement could be reduced, however, by developing an independent system for automatically building a dictionary of Concreteness, Familiarity, and Imageability ratings for nouns in a new language using only the background corpus. While that is not a part of this project, the point is that complete language independence is only possible if basic processing resources can also be automatically created for a new language. Tsvetkov, et al. (2014) present a language-independent approach to metaphor identification which also is language independent only to the degree that basic resources are available for each language (in this case, a machine-readable bilingual dictionary between the new language and English). Complete language independence requires either basic resources for each language or a system for automatically creating such resources.

## Conclusions

In this paper we have shown a language independent metaphor identification system that relies on only a small amount of semantic information and performs as well as or better than other extant systems in both English and other languages. The modular architecture will allow us to integrate multiple metaphor identification methods in a single system, which is important because different metaphors are best identified using different sorts of linguistic cues.

The most immediate area for future work is to implement and integrate more metaphor identification modules and investigate how they can be best combined. There are two particularly promising techniques for negotiating the results of many classifier modules: (i) combination methods that use estimates of a module's confidence in its identifications to better determine an aggregate response, and (ii) meta-classifiers that learn better aggregation functions, using the level of confidence of each classifier module and properties of the candidate expression as features.

### Acknowledgements

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0021. 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.

## References

Assaf, D., Neuman, Y.; Cohen, Y.; Argamon, S.; Howard, N.; Last, M.; Koppel, M. (2013). Why "dark thoughts" aren't really dark: A novel algorithm for metaphor identification. In *2013 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)*, 60-65. IEEE.

Black, M. (1954). Metaphor. *Proceedings of the Aristotelian Society, New Series*, 55, 273-294.

Bracewell, D.; Tomlinson, M.; and Mohler, M.. (2013). Determining the Conceptual Space of Metaphoric Expressions. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Vol. 1*, 487-500. Berlin, Heidelberg: Springer-Verlag.

Carbonell, J. G. (1980). Metaphor: a key to extensible semantic analysis. In *Proceedings of the 18th annual meeting on Association for Computational Linguistics*, 17–21. Association for Computational Linguistics.

Dunn, J. (2013a). What metaphor identification systems can tell us about metaphor-in-language. In *Proceedings of the First Workshop on Metaphor in NLP*: 1-10. Stroudsburg, PA: Association for Computational Linguistics.

Dunn, J. (2013b). Evaluating the premises and results of four metaphor identification systems. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Vol. I*, 471-486. Berlin: Springer.

Gandy, L.; Allan, N.; Atallah, M.; Frieder, O.; Howard, N.; Kanareykin, S.Argamon; S. (2013). Automatic Identification of Conceptual Metaphors With Limited Knowledge. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 328–334. AAAI Press.

Hovy, D., Srivastava, S., Jauhar, S. K., Sachan, M., Goyal, K., Li, H., Hovy, E. (2013). Identifying Metaphorical Word Use with Tree Kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, 52-57. Stroudsburg, PA: Association for Computational Linguistics.

Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and Thought, 2nd edition*, 202–251. Cambridge, UK: Cambridge Univ Press.

Lakoff, G., and Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Lakoff, G., and Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. Chicago: University of Chicago Press.

Li, L., and Sporleder, C. (2010a). Using Gaussian Mixture Models to Detect Figurative Language in Context. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 297-300. Stroudsburg, PA, USA: Association for Computational Linguistics.

Li, L., and Sporleder, C. (2010b). Linguistic Cues for Distinguishing Literal and Non-literal Usages. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 683–691. Stroudsburg, PA, USA: Association for Computational Linguistics.

Mason, Z. (2004). CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System. *Computational Linguistics*, 30(1), 23–44.

Mohler, M.; Bracewell, D.; Tomlinson, M.; and Hinote, D. (2013). Semantic Signatures for Example-Based Linguistic Metaphor Detection. In *Proceedings of the First Workshop on Metaphor in NLP*, 27–35. Stroudsburg, PA: Association for Computational Linguistics.

Neuman, Y.; Assaf, D.; Cohen, Y.; Last, M.; Argamon, S.; Howard, N.; and Frieder; O. (2013). Metaphor identification in large texts corpora. *PloS one*, 8(4).

Shutova, E.; Teufel, S.; and Korhonen, A. (2013). Statistical Metaphor Processing. *Computational Linguistics*, 39(2), 301–353.

Steen, G. (2007). *Finding metaphor in grammar and usage: A methodological analysis of theory and research*. Amsterdam: John Benjamins.

Steen, G.;. Dorst, A.; Herrmann, B.; Kaal, A.; and Krennmayr, T.. (2010). Metaphor in usage. *Cognitive Linguistics*, 21(4), 765–796.

Strzalkowski, T.; Broadwell, G.; Taylor, S.; Feldman, L.; Shaikh, S.; Liu, T; Elliot, K. (2013). Robust Extraction of Metaphor from Novel Data. In *Proceedings of the First Workshop on Metaphor in NLP*, 67–76. Association for Computational Linguistics.

Tsvetkov, Y.; Mukomel, E.; and Gershman, A.. (2013). Cross-Lingual Metaphor Detection Using Common Semantic Features. In *Proceedings of the First Workshop on Metaphor in NLP*, 45–51. Stroudsburg, PA: Association for Computational Linguistics.

Tsvetkov, Y; Boytsov, L.; Gershman, A.; Nyberg, E.; Dyer, C. (2014). Metaphor Detection with Cross-Lingual Model Transfer. *Proceedings of 2014 Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics.

Turney, P.; Neuman, Y.; Assaf, D.; and Cohen, Y. (2011). Literal and Metaphorical Sense Identification Through Concrete and Abstract Context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 680–690. Stroudsburg, PA, USA: Association for Computational Linguistics.

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

Wilks, Y. (1978). Making preferences more active. *Artificial Intelligence*, 11(3), 197223.