Introducing the LCC Metaphor Datasets

Michael Mohler, Mary Brunson, Bryan Rink, Marc Tomlinson
Language Computer Corporation
Richardson, Texas, USA
michael,mary,bryan,marc@languagecomputer.com

Abstract

In this work, we present the Language Computer Corporation (LCC) annotated metaphor datasets, which represent the largest and most comprehensive resource for metaphor research to date. These datasets were produced over the course of three years by a staff of nine annotators working in four languages (English, Spanish, Russian, and Farsi). As part of these datasets, we provide (1) metaphoricity ratings for within-sentence word pairs on a four-point scale, (2) scored links to our repository of 114 source concept domains and 32 target concept domains, and (3) ratings for the affective polarity and intensity of each pair. Altogether, we provide 188,741 annotations in English (for 80,100 pairs), 159,915 annotations in Spanish (for 63,188 pairs), 99,740 annotations in Russian (for 44,632 pairs), and 137,186 annotations in Farsi (for 57,239 pairs). In addition, we are providing a large set of likely metaphors which have been independently extracted by our two state-of-the-art metaphor detection systems but which have not been analyzed by our team of annotators.
Keywords: Metaphor, Dataset, Metaphoricity, Conceptual Metaphor, Affect

1. Introduction

Metaphor is everywhere in human language. It adorns our poetry, adds clarity to our science and philosophy, and serves as a fountain of language for new and complex ideas. For this reason, it is crucial that natural language processing (NLP) technologies be capable of correctly identifying and interpreting metaphor. Indeed, metaphor has been found to confound most language processing tools across a wide variety of applications – including textual entailment, text summarization, word sense disambiguation, semantic textual similarity, question answering, and event extraction. Unfortunately, research in metaphor to date has largely been hampered by the lack of size, quality, and diversity in the metaphor datasets available to researchers (Shutova, 2015). In this work, we present for the first time Language Computer Corporation's web-scale metaphor datasets, which have been produced by a team of nine annotators working in four languages (American English, Mexican Spanish, Russian, and Farsi) over a three-year period.

As a foundation for our metaphor annotation effort, we begin with George Lakoff's Contemporary Theory of Metaphor (Lakoff and Johnson, 1980; Lakoff, 1993), which popularized the idea of a conceptual metaphor mapping. Within the cognitive framework of a given conceptual mapping, terms pertaining to one concept or domain (the *source*) can be used figuratively to express some aspect of another concept or domain (the *target*). For example, the conceptual metaphor "Life is a Journey" indicates a medium within which the target concept "life" may be more easily discussed and understood. This particular mapping allows us to speak of one being stuck in a "dead-end" job, of a crucial decision as being a "fork in the road", or of someone's life "taking a wrong turn".

From this idea of the conceptual metaphor, we define our basic unit of annotation: a pair of syntactically-related candidate terms within a sentence which are evaluated for their ability to transfer ideas from a *source* to a *target* domain. For each pair of terms in our dataset, we provide annotations that include (1) a determination as to the pair's syntactic potential for metaphor; (2) an evaluation of the pair's metaphoricity on a four-point scale; (3) the identification of its underlying source and target domains; (4) an evaluation of its affective polarity – positive, negative, or neutral – and (5) an evaluation of its affective intensity on a four-point scale.

Unlike many existing metaphor datasets, the metaphor annotations we introduce in this work represent the full range of metaphoricity – novel metaphors, conventionalized metaphors, and non-metaphors (i.e., literal usages). Likewise, our annotations are not limited to pre-determined syntactic constructs or a reduced set of parts-of-speech.[1] At the domain level, our annotations cover 114 source domains and 32 target domains. Altogether, we provide 188,741 annotations in English (for 80,100 pairs), 159,915 annotations in Spanish (for 63,188 pairs), 99,740 annotations in Russian (for 44,632 pairs), and 137,186 annotations in Farsi (for 57,239 pairs). We also provide 182,305 likely metaphors which have been extracted by two independent metaphor detection systems (87,379 for English, 52,611 for Spanish, 19,387 for Russian, and 22,928 for Farsi).

The remainder of this work is presented as follows. In Section 2, we discuss existing resources for metaphor at a large and a small scale and compare them to the LCC Metaphor Datasets that we describe here. In Section 3, we describe the provenance of our underlying data sources and the methodology by which we have selected the pairs to be annotated. In Section 4, we detail the types of annotations provided in these datasets, along with our repository of source and target concepts. In Section 5, we indicate the level of agreement between the annotators on a variety of dimensions. Finally, we conclude in Section 6.

2. Related Work

The only resource which approaches our LCC Metaphor Datasets in scale is the CCM conventionalized metaphor databank (Levin et al., 2014), released by researchers

[1] Our annotations contain nouns, verbs, multi-word expressions, adjectives, and (rarely) adverbs.

at Carnegie Mellon University. Like us, they provide metaphors covering a variety of syntactic relationships for English, Spanish, Russian, and Farsi. However, their corpus only consists of between 4,000 and 8,000 sentences in each language and only considers conventionalized (that is, very common) metaphors that they have extracted from a document corpus in a targeted manner. By contrast, our dataset consists of both conventionalized and novel metaphors which have been extracted from text in a variety of ways.

Three additional datasets deserve mention, as they have served as the *de facto* standard training and evaluation datasets for the majority of all metaphor-based research over the past decade. The first is the TropeFinder (TroFi) dataset (Birke and Sarkar, 2006), which consists of 2,146 metaphoric and 1,593 literal usages of 51 English verbs. These sentences were taken from a single domain – The Wall Street Journal – and represent the most frequently used dataset for metaphor-based evaluation.

A second frequently used metaphor dataset is the VU Amsterdam corpus (Steen et al., 2010) which consists of 200,000 words taken from the academic, fiction, and news subdomains of the British National Corpus (BNC) (Burnard and Berglund, 2007). This data was then labeled according to the MIP annotation process (Pragglejaz Group, 2007). Five annotators labeled each word of the corpus as figurative (specifying metaphor/personification/other) or literal. Once metaphoric prepositions and ambiguous metaphors are removed from consideration, the total size of this corpus is 6,893 sentences, with 4,553 sentences containing metaphor and 2,340 sentences with no metaphor (Dunn, 2014).

A third commonly used metaphor dataset is that of Shutova (2010), which was likewise taken from the BNC. This dataset was produced by sampling various genres, including literature; newspaper/journal articles; essays on politics; international relations and history; and radio broadcasts. The resulting corpus consists of 761 English sentences annotated independently by three native English speakers. These annotators were asked to (1) classify the verbs as metaphorical or literal and (2) identify the conceptual mappings for the verbs. Metaphoric nouns and adjectives were not considered.

From an historical perspective, the Berkeley Master Metaphor List (Lakoff, 1994) deserves mention. This resource consists of 208 conceptual metaphors (e.g., A_Problem_Is_A_Body_Of_Water), along with a handful of examples of each. While this represents a solid resource for linguistic inquiry into the phenomenon of conceptual metaphor, it is unsuitable for use in developing tools for metaphor-focused NLP insofar as the size is limited, it consists of no negative instances of metaphor, and it is not represented in a way that is conducive to training and evaluating a real-world metaphor processing system.

Other metaphor-related datasets include one containing 115 Italian literary metaphors (Bambini et al., 2014); a collection of around 8,600 mind-related metaphors from 18th century literature (Pasanek and Sculley, 2008); and a similar collection of 1,100 mental metaphors;[2] a list of around 200 metaphors (out of context) provided by Li et al. (2013); the Hamburg Metaphor Database (Reining and Lönneker-Rodman, 2007), which contains 1,656 annotated French and German metaphors; the dataset of Tsvetkov et al. (2014), which contains 884 literal and 884 metaphoric adjective-noun pairs (without context); and the conceptual schemas dataset (Gordon et al., 2015), which provides a rich interpretive framework for conceptual metaphor, but does not itself address metaphor detection.

3. Data Sources

Unlike many existing datasets, the LCC Metaphor Datasets have been constructed for the express purpose of enabling and improving the development and evaluation of a system for enterprise-quality, open-text metaphor identification. To this end, we have built the majority of our annotations on top of a collection of publicly available web documents. For American English, we have made use of ClueWeb09 corpus[3] consisting of around 1 billion English language web pages and a supplemental dataset consisting of text scraped from the popular Debate Politics online forum.[4] For Mexican Spanish, we have employed the Spanish Gigaword corpus (Linguistic Data Consortium, 2011), which has been post-processed to restrict the IP address of the web page to Mexico in order to avoid both Peninsular and South American variants of the language. For Russian, we have made use of the 2.2 billion token RuWac corpus,[5] which represents a snapshot of the Russian web. For Farsi, we made use of the Hamshahri corpus[6] of Iranian newswire text.

From these corpora, we have produced annotations in three ways. The first set of annotations (SYS) consists of a validated subset of the output from two metaphor detection systems: (1) our machine-learning based, tiered metaphor detection system (Bracewell et al., 2014); and (2) our standalone semantic generalization system (Mohler et al., 2015). On a monthly basis, these systems were run over a shifting portion of the datasets described. The subset to validate (on a weekly basis) was selected within an *ad hoc* active learning environment which extracted validation pairs with a variety of characteristics, including (dis)similarity to existing annotations, target/source diversity, syntactic relation diversity, system confidence, and system disagreement. While the result of this annotation method was a large and diverse set of metaphoric and literal pairs, it suffered from an unavoidable sample bias in that it was known, *a priori*, that one or more of our algorithms was able to extract each of the metaphors. In short, this dataset enabled us to estimate system precision, but not system recall.

Our second dataset (REC) was developed to address this problem by providing a more natural source of training data with a significant number of useful non-metaphor annotations. In this case, individual documents were selected automatically to be annotated thoroughly for a full set of lex-

[2] http://www.cs.bham.ac.uk/ jab/ATT-Meta/Databank

[3] http://www.lemurproject.org/clueweb09.php/

[4] http://www.debatepolitics.com/

[5] http://corpus.leeds.ac.uk/tools/ru/ruwac-parsed.out.xz

[6] http://ece.ut.ac.ir/dbrg/hamshahri/

Language	Dataset	Non-Syntactic	Metaphors	Non-Metaphors	Unclear	Language	Dataset	Non-Syntactic	Metaphors	Non-Metaphors	Unclear
	ANN	0	8,597	0	140		ANN	0	4,227	0	358
EN	REC	21,124	371	1,731	319	ES	REC	18,807	158	603	763
	SYS	6,857	11,101	22,434	7,425		SYS	2,215	6,799	20,554	8,703
	UNV	0	87,397	0	0		UNV	0	52,611	0	0
	TOTAL	27,981	107,466	24,165	7,884		TOTAL	21,022	63,795	21,157	9,824
	ANN	0	3,334	0	310		ANN	0	2,122	0	208
RU	REC	19,609	1,048	3,181	654	FA	REC	23,195	313	968	146
	SYS	2,045	4,360	7,501	2,590		SYS	7,359	6,575	10,588	5,765
	UNV	0	19,387	0	0		UNV	0	52,611	0	0
	TOTAL	21,654	28,129	10,682	3,554		TOTAL	30,554	61,621	11,556	6,119

Table 1: The datasets associated with each language. Non-syntactic pairs are indicated with a -1.0 metaphoricity score in the dataset. Other scores are categorized according to their average metaphoricity score across all annotators for that pair. Non-metaphors are those with $0 \leq \text{score} < 0.5$. Metaphors are those with $1.5 < \text{score} \leq 3$. Unclear are those with $0.5 \leq \text{score} < 1.5$.

Language	Dataset	Non-Syntactic	Metaphors	Non-Metaphors	Unclear	Language	Dataset	Non-Syntactic	Metaphors	Non-Metaphors	Unclear
	ANN	0	1,104	0	17		ANN	0	492	0	44
EN	REC	6,335	102	567	107	ES	REC	7,057	59	203	251
	SYS	1,206	1,830	3,740	1,257		SYS	354	1,125	3,466	1,473
	TOTAL	7,541	3,036	4,307	1,381		TOTAL	7,411	1,676	3,669	1,768

Table 2: The reduced datasets in English and Spanish, which are being released as a free resource for research purposes. These datasets were created by randomly sampling the larger datasets for a size of approximately 1/8 of the annotator example dataset, 1/3 of the recall validation dataset (by document), and 1/6 of the system validation dataset. The unvalidated dataset was not included.

emes associated with one of our target concepts. For each sentence in these documents that contained a target lexeme, annotators were asked to evaluate as candidate source lexemes all collocations, nouns, verbs, and adjectives in the same sentence. This is the most "natural" of our datasets.

Our third dataset (UNV) represents a set of pseudo-annotations. That is, it consists of the joint output of our two state-of-the-art metaphor detection systems. The first of these systems (Bracewell et al., 2014) is a machine-learning based approach to metaphor detection which takes into account syntactic, lexical, psycholinguistic, and selectional interactions between the metaphor pair itself and with the wider context. The second system (Mohler et al., 2015) is example-based and employs a generalization component to map unseen examples to the example base using language-independent ontological information and semantic parsing. These datasets, therefore, represent the set of metaphor pairs which both systems have independently judged to be metaphoric, but which our annotators did not have the opportunity to inspect manually. While this dataset is unvalidated, it represents a huge resource of highly likely metaphors across a variety of source/target domains.

Finally, we provide an additional dataset (ANN) which consists of examples selected directly from the web. For this dataset, annotators were instructed to find representative (non-conventionalized) metaphors for a variety of lexical realizations for each of our source and target concepts (e.g., they were asked to find Government_As_Monster metaphors). While this dataset is a good source of novel metaphors for a variety of source and target domains, it only contains annotations for the targeted pair selected by the annotators and does not include any non-metaphor annotations.

The size of each of these datasets is shown in Table 1, along with their metaphoricity breakdown as being either non-syntactic, metaphoric, non-metaphoric, or unclear. In addition to the full dataset, which is being released under licensing constraints, we are providing an abridged set of datasets for English and Spanish, which can be used at no cost for research purposes. These datasets consist of a randomly sampled subset of the larger datasets without respect to metaphoricity, affect, source concept, or target concept. The size of each of these reduced datasets is indicated in Table 2. Both the full datasets and the free datasets can be found at our company website.[7]

4. Annotation Types

As part of our annotation process, our annotators were asked to provide information on four aspects of a candidate source/target pair. The first step of the annotation process was to indicate whether or not the syntactic relation between the terms is capable of serving as an avenue for metaphor. This annotation is useful primarily in training a system to limit the number of term pairs to consider before applying more robust (and costly) components to the metaphor detection problem. For all syntactically relevant pairs, the annotators were then asked to rate the metaphoricity of the pair on a four-point scale. Then, for all pairs that were assigned a positive metaphoricity score, annotators were asked to rate one or more system-generated mappings to a CM source domain, to provide a better mapping if necessary, and to rate the affective polarity and intensity of the pair.

[7] http://www.languagecomputer.com/metaphor-data.html

Rating	Description	Example
0	No Metaphoricity	Unlawful means to **gain** personal wealth should be prohibited.
1	Possible/Weak Metaphor	Our goal is **increasing** the income of the low-paid.
2	Likely/Conventional Metaphor	Defining a poverty line is necessary to analyze the **depth** of poverty.
3	Clear Metaphor	Indochina would have been a safe **haven** for democracy.

Table 3: Example sentences for each metaphoricity rating. In the examples, the source lexeme is **bolded** while the target lexeme is underlined.

Target Concept	EN	ES	RU	FA	Target Concept	EN	ES	RU	FA
ABORTION	120	50	12	0	BUREAUCRACY	3,240	3,649	2,595	1,188
CLIMATE_CHANGE	394	115	26	0	CONTROL_OF_GUNS	3,342	0	0	0
DEBT	101	386	211	556	DEMOCRACY	3,505	5,723	2,964	6,284
DEMOGRAPHICS	161	106	26	0	DISEASE	1,604	171	57	0
DRUG_TRAFFICKING	352	91	0	0	ELECTIONS	3,085	3,978	1,635	2,638
GOVERNMENT	4,592	9,613	5,729	7,477	GUN_DEBATE_GROUPS	451	0	0	0
GUN_OWNERSHIP	1,026	0	0	0	GUN_RIGHTS	2,244	0	0	0
GUNS	7,304	0	0	0	GUN_VIOLENCE	568	0	0	0
INTELLECTUAL_PROPERTY	738	57	15	0	ISLAMIC	281	24	78	0
MARRIAGE	278	79	118	0	MENTAL_CONCEPTS	3,331	614	186	0
MIGRATION	765	183	63	0	MONEY	1,387	2,310	1,459	995
POLITICIANS	419	191	85	0	POVERTY	3,303	5,132	1,845	3,092
RELIGION	1,979	664	502	0	TAXATION	3,223	5,407	2,612	2,104
TAXES	1,114	68	243	81	TAXPAYERS	316	151	65	5
TERRORISM	389	111	49	0	WEALTH	2,146	3,095	2,589	2,398
WEAPONS_OF_MASS_DESTRUCTION	85	15	14	0	WELFARE	246	88	32	0

Table 4: Counts for each of our target domains excluding those from the UNV dataset.

4.1. Syntactic Potential

A significant amount of previous work on metaphor has followed Krishnakumaran and Zhu (2007) in concentrating on three classes of metaphoric syntax – (1) nouns linked by IS-A relationships, (2) verbs with their subjects or objects, and (3) nouns with their adjectives. While this represents a significant percentage of metaphors in English, we instructed our annotators to consider a far wider range of potential syntactic realizations – including, but not limited to, those employing prepositions, relative clauses, and complements. In short, we flag (as syntactically non-relevant) only those cases in which the source and target lexemes chosen (1) are grammatically unrelated or only distantly related in the sentence; (2) do not relate to each other meaningfully; and (3) are incapable of forming a metaphor.

4.2. Metaphoricity

For all syntactically relevant pairs, annotators were asked to judge metaphoricity according to criteria comparable to the MIP annotation guidelines (Pragglejaz Group, 2007). In order to enable a finer-grained understanding of metaphoricity, we instructed the annotators to employ a four-point metaphoricity scale to allow for more nuance than a simple literal/metaphoric flag.[8] In particular, annotators were asked to rate the degree of metaphoricity between the candidate source and target terms on a scale from 0 to 3, taking into consideration the context of the entire sentence (and, if available, additional paragraph-level context). Example sentences, along with a shorthand description of each metaphoricity score, can be found in Table 3.

[8]For a beneficial discussion of the advantages of fine-grained metaphoricity, see Dunn (2014)

In making this determination, annotators were asked to consider the following criteria:

Sensory Perceptibility How easily the source can be perceived by the senses (e.g., seen, heard, etc.).

Expressiveness How vivid the language being used is.

Uncommonness How frequently the metaphor is encountered.

Non-literalness How jarring the metaphor would be if it were to be read and interpreted literally.

Metaphors with higher metaphoricity are expected to exhibit a higher degree of the features above, whereas metaphors with lower metaphoricity are expected to exhibit lower sensory perceptibility and to be less expressive, more common, and less jarring when read and interpreted literally.

4.3. Domain Mappings

For each pair of source/target terms which were determined to have a positive metaphoricity, the annotators were asked to determine the appropriate source and target domains (i.e., to define the underlying conceptual metaphor). For this purpose, they were provided with two lists – one containing 32 target domains (see Table 4) and one containing 114 source domains (see Table 5). In particular, annotators were provided with a single CM target domain and two CM source domains produced by our conceptual metaphor mapping system (Mohler et al., 2014). The annotators were then asked to rate each domain separately for the degree to which the source or target lexeme (disambiguated by context) fits into that domain. Annotators produced ratings on

ABYSS	ACCIDENT	ADDICTION
A_GOD	ANIMAL	A_RIGHT
AVERSION	BACKWARD_MOVEMENT	BARRIER
BATTLE	BLOOD_STREAM	BLOOD_SYSTEM
BODY_OF_WATER	BUILDING	BUSINESS
CLOTHING	COMPETITION	CONFINEMENT
CONTAINER	CONTAMINATION	CONTROL
CRIME	CROP	DARKNESS
DELICACY	DESIRE	DESTROYER
DISEASE	DOWNWARD_MOVEMENT	EMOTION_EXPERIENCER
EMPLOYEE	ENABLER	ENERGY
ENSLAVEMENT	FABRIC	FACTORY
FAMILY	FIRE	FOOD
FORCEFUL_EXTRACTION	FORWARD_MOVEMENT	FURNISHINGS
GAME	GAP	GEOGRAPHIC_FEATURE
GIFT	GOAL_DIRECTED	GOURMET_CUISINE
GREED	HAZARDOUS_GEOGRAPHIC_FEATURE	HEAT
HIGH_LOCATION	HIGH_POINT	HUMAN_BODY
IMPURITY	INDUSTRY	INSANITY
JOURNEY	LEADER	LIFE_STAGE
LIGHT	LIVESTOCK	LOW_LOCATION
LOW_POINT	LUXURY	MACHINE
MAGIC	MATERIAL	MAZE
MEDICINE	MONSTER	MORAL_DUTY
MOVEMENT	MOVEMENT_ON_A_VERTICAL_SCALE	NATURAL_PHYSICAL_FORCE
OBESITY	OBJECT_HANDLING	PARASITE
PATHWAY	PHYSICAL_BURDEN	PHYSICAL_HARM
PHYSICAL_LOCATION	PHYSICAL_OBJECT	PLANT
PLIABILITY	PORTAL	POWER
POSITION_AND_CHANGE_OF_POSITION_ON_A_SCALE		PRESSURE
PROTECTION	RACE	RESOURCE
RULE_ENFORCER	SCHISM	SCIENCE
SECURITY	SERVANT	SHAPE
SIZE	STAGE	STORY
STRENGTH	STRUGGLE	TEMPERATURE
THEFT	TOOL	TRIBUTE
UPWARD_MOVEMENT	VEHICLE	VERTICAL_SCALE
VISION	WAR	WEAKNESS
WEATHER		

Table 5: The full set of source concepts available as part of the LCC Metaphor Datasets.

a scale from 0 to 3, where 0 represents no relationship between the lexeme and the CM domain, and 3 represents a very strong, apt relationship between the two. For example, the CM source domain ANIMAL should be rated as a 0 for a source lexeme of "building", but as a 3 for a source lexeme of "salamander".

If a better (or equally good) source or target domain existed that was not provided by the system, the annotators were then instructed to select from a drop-down menu the best CM source or target domain available for that lexeme. This list included an option of "OTHER", which was to be selected if none of the domains in the list represented the most appropriate mapping.

4.4. Affect Annotations

Along with the CM source and target domain annotations, the annotators were asked to rate the affect, or emotional impact, of the source and target lexemes in the cases where these formed a valid linguistic metaphor (i.e., where the metaphoricity score was greater than 0). Ratings were made on a scale from -3 to 3, as shown in Table 6.

Our annotators were instructed to consider both the source and the target concepts when determining the overall affective qualities of the metaphor pair. This allowed the annotators to make a determination in cases where the source term was ambiguously good or bad, depending on the target term. For example, in "We need to **fight** poverty", the "fight poverty" metaphor should be assigned a positive affect, since poverty is typically a bad thing (negative), and thus, fighting against poverty is a good thing (positive). On the other hand, the affect of "These politicians are **fighting** the poor" should be negative since causing harm to the poor of society is bad. These considerations bear a strong similarity to the notion of *Affect Calculus*, first described by Strzalkowski et al. (2014),

In order to simplify the process, the annotators were instructed not to take into consideration any external modifiers that altered the modality or polarity of the metaphor. Thus, "poverty is a **bottomless pit**" would receive the same affect rating as "poverty is not a **bottomless pit**", since the metaphor being used in both cases is a negative one. Similarly, "poverty might be a **bottomless pit**" was to be rated

Rating	Description	Example
-3	Strongly Negative	Communism must be the natural form of government for people who **worship** the government.
-2	Moderately Negative	When you have **piles** of debt, it is hard not to stay up at night worrying about it.
-1	Weakly Negative	Georgia will face a further **test** of its democracy.
0	Neutral	Rebuilding a strong middle class strengthens our tax **base**.
1	Weakly Positive	They would not lower the cost of goods when the taxes **came down**.
2	Moderately Positive	If you aren't able to **pull yourself out of** poverty, it's clearly because you are lazy.
3	Strongly Positive	Do you think that Liberal Democracy is the **apex** of the world's governments?

Table 6: Example sentences for each affect rating. In the examples, the source lexeme is **bolded** while the target lexeme is underlined.

as if it did not contain the qualifying modal word "might". Additionally, the annotators were instructed to rate affect based on the perspective of the author of the text, not their own. As such, "Taxes are an effective **medicine**" would be rated as having positive affect, even if the annotator him/herself did not agree with the statement.

5. Agreement

As a part of our annotation effort, we assigned candidate pairs to particular annotators with a mixture of singly annotated pairs and multiply annotated pairs. This was done both to ensure the quality of the resource we were creating and to enable us to address any misunderstandings that the annotators experienced. As such, a significant number of the metaphor pairs included in this resource were annotated by more than one person as shown in Table 7.

	EN	ES	RU	FA
1 Annotator	46,311	36,490	19,812	22,443
2 Annotators	5,756	4,506	3,341	3,786
3 Annotators	21	1,067	57	589
4 Annotators	1	8	0	0

Table 7: The number of metaphor pairs singly and multiply annotated across all languages.

By the end of the project, we were periodically assessing the inter-annotator agreement for each type of annotation in each language. The results of the most recent such analysis (December 2014) are shown in Table 8.[9] It should be pointed out that the relatively low agreement on affect polarity in English was due to annotator misunderstanding regarding the annotation of a controversial (in the U.S.) target domain – CONTROL_OF_GUNS. One of our annotators considered increased gun control a good thing, while the other considered it a bad thing, which resulted in a disagreement about the overall affect calculus. This effect was not found for other target domains and led to our clarifying the annotation guidelines for affect as described in Section 4.4.

6. Conclusion

The datasets we provide here represent a new and invaluable resource for research in metaphor detection and conceptual metaphor mapping, based upon their size, diversity, and reliability. It is our sincere hope that researchers in metaphor across a wide range of disciplines can make substantial use of this resource in both studies of metaphor in language and in the further development of tools for detecting, processing, and interpreting metaphor as a key, but often overlooked, component of open-domain natural language understanding.

[9]Note that this is not the same as the inter-annotator agreement of the released datasets themselves. Rather, it represents a snapshot of the annotation quality as of December 2014, after many months of annotator training and feedback.

	EN	ES	RU	FA
Syntactic Relatedness	90.5%	92.6%	95.1%	86.6%
Metaphoricity	92.8%	92.1%	87.4%	87.1%
Source Relatedness	95.1%	88.6%	97.4%	87.0%
Target Relatedness	94.3%	93.7%	94.3%	91.8%
Affect Polarity	82.1%	91.0%	96.1%	100.0%

Table 8: Snapshot analysis of our inter-annotator agreement at the end of the project. Note that syntactic relatedness and affect polarity are binary decisions, while metaphoricity and source/target relatedness are graded on a scale. For the latter, differences of 1 point (or less) were considered to be representative of agreement.

7. Acknowledgments

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-0025. 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."
Furthermore, we would like to thank each of our annotators without whom this resource would not exist.

8. Bibliographical References

Bambini, V., Resta, D., and Grimaldi, M. (2014). A dataset of metaphors from the italian literature: Exploring psycholinguistic variables and the role of context.

Birke, J. and Sarkar, A. (2006). A clustering approach for nearly unsupervised recognition of nonliteral language. In *EACL*.

Bracewell, D., Tomlinson, M., Mohler, M., and Rink, B. (2014). A tiered approach to the recognition of metaphor. In *Computational Linguistics and Intelligent Text Processing*.

Burnard, L. and Berglund, Y. (2007). Exploring bnc xml edition with xaira. In *28th Annual Conference of the International Computer Archive for Modern and Mediaeval English (ICAME)*.

Dunn, J. (2014). Measuring metaphoricity. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics*, volume 2, pages 745–751. Association for Computational Linguistics Stroudsburg, PA.

Gordon, J., Hobbs, J., May, J., Mohler, M., Morbini, F., Rink, B., Tomlinson, M., and Wertheim, S. (2015). A corpus of rich metaphor annotation. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 56–66. Association for Computational Linguistics.

Krishnakumaran, S. and Zhu, X. (2007). Hunting elusive metaphors using lexical resources. In *Proceedings of the Workshop on Computational approaches to Figurative Language*, pages 13–20. Association for Computational Linguistics.

Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*, volume 111. Chicago London.

Lakoff, G. (1993). The contemporary theory of metaphor. *Metaphor and thought*, 2:202–251.

Lakoff, G. (1994). *Master metaphor list*. University of California.

Levin, L., Mitamura, T., Fromm, D., MacWhinney, B., Carbonell, J., Feely, W., Frederking, R., Gershman, A., and Ramirez, C. (2014). Resources for the detection of conventionalized metaphors in four languages. In *Proceedings of the Ninth Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland*.

Li, H., Zhu, K. Q., and Wang, H. (2013). Data-driven metaphor recognition and explanation. *TACL*, 1:379–390.

Mohler, M., Rink, B., Bracewell, D., and Tomlinson, M. (2014). A novel distributional approach to multilingual conceptual metaphor recognition.

Mohler, M., Tomlinson, M., and Rink, B. (2015). Cross-lingual semantic generalization for the detection of metaphor. *Computational Linguistics and Intelligent Text Processing. Springer*.

Pasanek, B. and Sculley, D. (2008). Mining millions of metaphors. *Literary and linguistic computing*, 23(3):345–360.

Pragglejaz Group. (2007). MIP: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.

Reining, A. and Lönneker-Rodman, B. (2007). Corpus-driven metaphor harvesting. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 5–12. Association for Computational Linguistics.

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

Shutova, E. (2015). Design and evaluation of metaphor processing systems. *Computational Linguistics*.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A., Krennmayr, T., and Pasma, T. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

Strzalkowski, T., Shaikh, S., Cho, K., Broadwell, G. A., Feldman, L., Taylor, S., Yamrom, B., Liu, T., Cases, I., Peshkova, Y., et al. (2014). Computing affect in metaphors. *ACL 2014*, page 42.

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. In *Proceedings of ACL*.

9. Language Resource References

Linguistic Data Consortium. (2011). *Spanish Gigaword Third Edition*. Linguistic Data Consortium, Spanish Gigaword, 1.0, ISLRN 595-627-966-073-3.