Institute of Informatics

Logics and Security Studies

Resource Validation Report

Kit Cho, Tomek Strzalkowski

May, 2015

ILS Tech Report: 17

Available for download from:

`http://www.ils.albany.edu`

Table of Contents

Summary

The REMND Project has developed an innovative and highly effective method for finding and understanding metaphors in four languages: English, Spanish, Russian, and Farsi. Our research demonstrated that metaphors can be reliably detected by tracking the use of highly imageable and concrete language. In addition, since most metaphors carry polarized affect, presence of affective language is another reliable indicator of metaphorical expressions.

In order to implement our method we constructed large lexicons with imageability and affective ratings in four languages. Since such resources were previously unavailable, we developed an automated procedure to expand relatively small research datasets created in psycholinguistic studies (largely in English), to obtain required resources, first in English and subsequently through automated translation in Spanish, Russian and Farsi. This procedure was described in detail in our final report submitted on March 31, 2015.

In this report, we describe the methodology used to validate the automatically obtained resources as well as the results of this validation. Overall, the results show that our expansion method is valid and the resulting resources are likewise valid and robust. We believe these new resources can be of significant interest to the research community, particularly in natural language processing and computational sociolinguistics.

ANEW Expansion

The well-known and highly cited corpus, which researchers currently consult for affective ratings of English words is the Affective Norms for English Words (ANEW) corpus (Bradley & Lang, 2009). The corpus consists of 2,477 words. The value for each word was obtained from ratings made by undergraduate students enrolled in a university. Ratings were made on a scale of 1 to 9, where a rating of 1 denoted highly negative, 9 denoted highly positive, with 5 being neutral. As of December 2014, the ANEW corpus has received over 1,400 citations on Google Scholar, making it evident that it has a large impact in the scientific community. However, one major limitation of the ANEW corpus is the relatively small size. More recently, Warriner, Kuperman, and Brysbaert (2013) created a more extensive affective norms corpus, collecting ratings for approximately 14,000 words. Their normative procedure was highly similar to that used by Bradley and Lang's (2009). One noteworthy difference, however, is that Warriner et al.'s (2013) ratings were gathered using participants recruited through Mechanical Turk. Although the Warriner et al.'s (2013) paper is an important extension of the ANEW corpus, 14,000 words may not be sufficient for researchers who are working with a large amount of text, as is often the case in fields such as natural language processing. Thus, a larger corpus was needed.

With respect to affect ratings for words other than English, at present, there is only one resource (Redondo, Fraga, Pardón, & Comesaña, 2007) that researchers rely on for affect ratings for Spanish words, and the number of words in their corpus is limited (1,034). More problematic is that there is no resource for Russian and Farsi affect word norms. Because affect ratings for words are used by researchers in a variety of disciplines, including psychology, linguistics, and computer science, creation of a comprehensive corpus that could be applied to different languages was much needed. The procedure we developed in the REMND project was described in our Final report. In this document, we describe the process used for validation of the automatically created resources. The first part of this document describes validation of our expansion procedure applied to English words. The second part then describes how we extended our validation to three other languages: Spanish, Russian, and Farsi. The third part discusses differences in validation results across the four languages.

English Corpus

As noted earlier, Warriner et al.'s (2013) procedure of collecting affect ratings was highly similar to those employed Bradley and Lang's (2009) with the exception that Warriner and colleagues recruited participants through Mechanical Turk (hereafter "Turkers"). Warriner et al. (2013) validated the reliability of their data by including 1,040 words in their study that were also in the ANEW corpus. A robust, positive correlation between the values of the words shared between the two corpora would provide strong evidence that the values obtained using Turkers are valid and as reliable as those collected using college undergraduates in a laboratory setting. Warriner et al. (2013) reported a correlation of .953, thus providing strong evidence of the validity of their validation method (and the reliability of collecting affect ratings from Turkers).

We briefly outline our method for automatically deriving affect ratings for single words; details are provided in our Final Project Report. Our method relies on imputing affect ratings for words that were derived from human raters to its first (most frequent) synsets, determined using Princeton's WordNet (Miller, 1995). WordNet is a large English lexical corpus with over 150,000 words, hierarchically organized in synsets that capture semantically equivalent words. For example, the first synsets to the word "building" are "edifice" and "construction". Thus, our method expansion will impute the affect value collected for "building" (hereafter source word) to both "edifice" and "construction" (hereafter expansion words). In some cases, multiple source words contribute to an expansion word, because the expansion word is the first synset of different source words (e.g., the expansion word "atrocious" is the first synset to the source words "horrible" and "awful"). In these cases, we took the average value of the source words to estimate the value of the expansion word.

Validation Method

We used two different approaches to provide converging evidence of the validity of our expansion method. The first approach is identical to that used by Warriner et al. (2013), which is to compare the affect values in our corpus to those obtained from a source that is established to be valid and reliable. Specially, we imputed values to expansion words using the words in ANEW as the source words. We then compared the values for the expansion words, which were derived automatically, to those obtained by Warriner et al. (2013), which relied on ratings made by human participants. A robust positive correlation between the expansion words' values obtained automatically through our expansion method and those obtained by Warriner et al. (2013) using human participants would thus support the notion that our method of expansion is valid.

The second approach compared the values imputed to the expansion words to ratings obtained using human subjects recruited through Amazon Mechanical Turk. In this second approach, we used expansion words that were derived using source words form Warriner et al.'s (2013) corpus. To ensure the reliability of our results, we also included words in our validation study that were collected by Warriner et al. (2013). Warriner et al. (2013) collected data for these words using Turkers, which is the method that we employ in our second validation approach. If our expansion method and results are reliable, we would expect a very strong positive correlation between the values obtained in the current study and those by Warriner et al. (2013).

Our validation approach for English words, which used two different procedures, will be described first. The procedure for our first validation approach is as follows. First, we took the 2,476 words in ANEW and imputed each word's affect value to the words first synsets (i.e., synonyms of the word's most common meaning). This method resulted in imputing values to 3,075 expansion words. Of these 3,075 words, 777 were excluded because they were compound words (e.g., "birthday suit") or short phrases (e.g., "bring out"). Another 1,285 (56%) words were excluded from analyses because the values for these words were not included in Warriner et al.'s (2013) norms. (The fact that 56% of our expansion words were not included in Warriner's normative study further reinforces the notion that a larger corpus was needed.) Our final validation set consisted of 1,013 words.

Results

In all our analyses reported in this document, we computed Pearson's correlation coefficient to compare the affect values from different sources. Our first analysis compared the values obtained by our automatic expansion method with those obtained by Warriner et al. (2013) from Turkers. The observed correlation was $r = .661$, a highly statistically significant result, $p < .001$ (see Figure 1). This robust correlation suggests that our expansion method is valid.

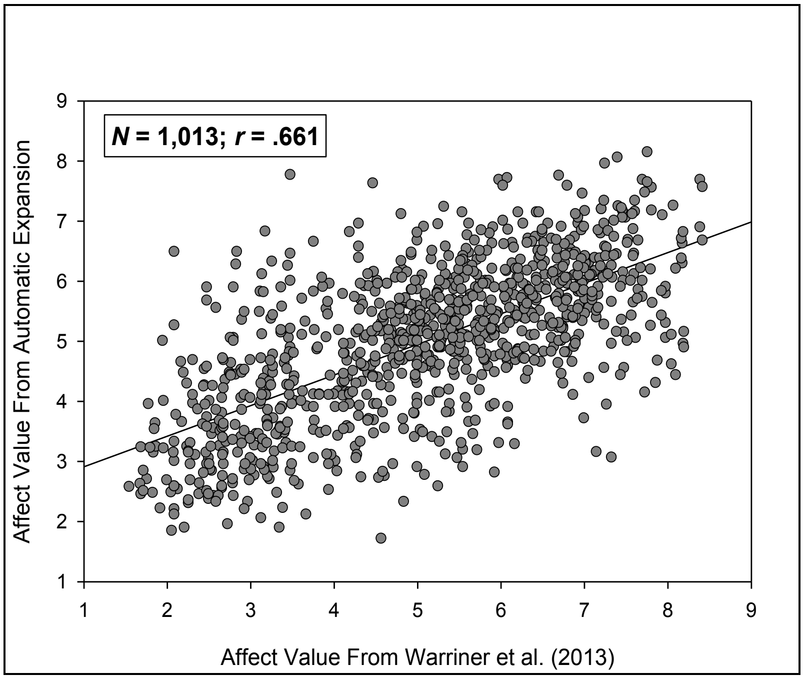


Figure 1: Scatterplot of the affect values of English words from the present expansion method of the ANEW corpus and those obtained by Warriner et al. (2013)

To provide convergent evidence of the validity of our expansion method, we conducted another validation. The procedure of our second validation approach mimicked that of the first approach with the exception that the source words were the 13,915 words in Warriner et al.'s (2013) corpus. Our expansion method yielded an additional 10,061 words. Of these words, 3,257 words were compound words or phrases. From the remaining 6,804 words, we randomly selected 235 words whose frequency of usage in the written text was variable (Log_HAL, taken from Balota et al., 2007) *Mean* = 5.17; *Standard Deviation* = 2.08; *Range* = .693-12.144). (This was to ensure that our sample stimuli were representative of words that appear in various written media, e.g., books, magazines.) We also included 40 words that were randomly selected from Warriner et al.'s (2013) corpus. These source words served as the check to ensure the reliability of the ratings of the results in our validation method. That is, if our sampling method and the responses given by Turkers are valid, we would expect that the affect values we obtained for these words in our study to be strongly correlated with the values obtained by Warriner et al. (2013). These 40 words and the 235 expansion words were randomly intermixed in a single list and presented in a new, randomized order for each Turker.

To maximize the likelihood that our Turkers would provide high-quality data and take the task seriously: (1) the description of our study stated that we are looking for native English speakers, and (2) we imposed a restriction such that only Turkers who have completed at least 1,000 studies, with an approval rating of 99% were allowed to participate. The instructions provided to Turkers were similar to those provided by Bradley and Lang (2010) and Warriner et al. (2013) to their participants. Turkers had an unlimited amount of time to answer each word but could not return to a word once they have indicated their response. We collected data from 17 Turkers.

To assess the reliability the ratings provided by Turkers in our study, we first assessed the correlation of the affect values for the 40 words for which data were collected by Warriner et al. (2013). The correlation for these words was nearly perfect, $r = .96$, giving us high confidence that our sampling method is reliable. The main analysis of interest was the correlation of the affect values for the 235 expansion words as derived automatically from our expansion method and those given by Turkers. The correlation for this analysis was .759, a highly statistically significant result, $p < .001$ (see Figure 2).

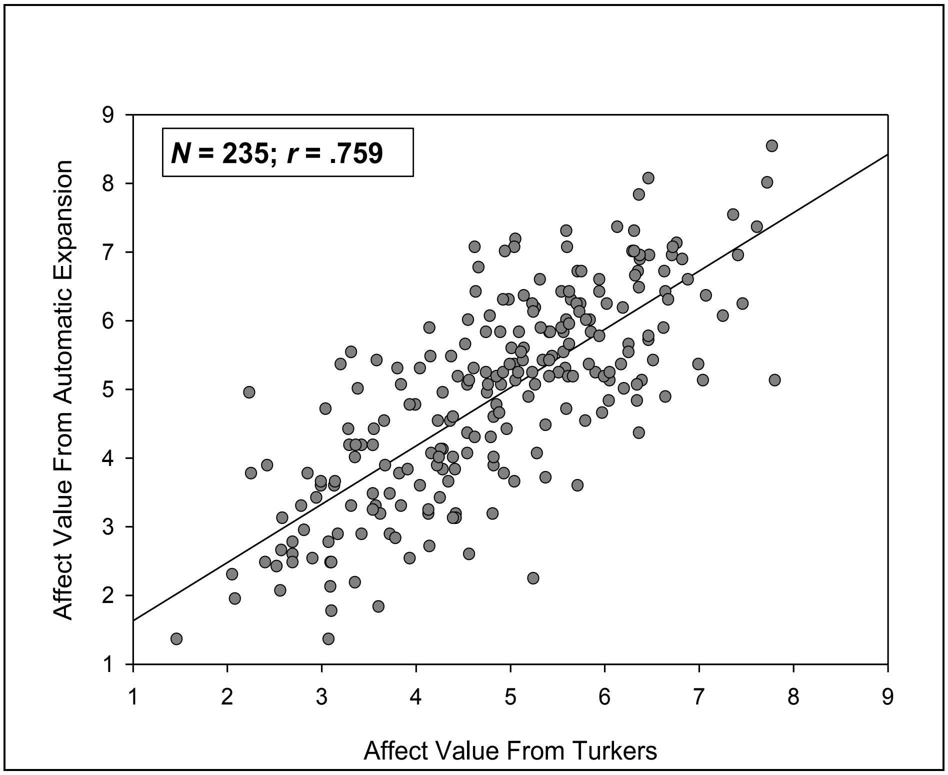


Figure 2: Scatterplot of the affect values of English words from the expansion words of the Warriner et al.'s (2013) corpus and those obtained from MTurkers.

We also analyzed the precision of our expansion by assessing the correlations based on the word's part of speech (i.e., adjective, noun, or verb). Although there are some words that have multiple parts of speech, for example, "DOG" as in the animal used as a noun, or "DOG" used as a verb to denote following someone, we categorized words in its most frequent usage; thus, in the aforementioned example, "DOG" would be categorized as a noun. The correlation coefficient for adjectives, nouns and verbs were all robust and comparable, .747, .748, and .802, thus further supporting the notion that our expansion method is valid.

Having established that our expansion procedure is an acceptable approach to deriving affect values for single words in English, we now tested how well affect values for words in English correlate with their foreign-language translations equivalent, specifically, Spanish, Russian, and Farsi. Because our validation approach for Spanish differed from that used for Russian and Farsi (for reasons explained below), we describe our validation protocol for Spanish first and separate from that used for Russian and Farsi.

Validation of Spanish Corpus

At present, there is only one corpus that researchers consult for affective ratings of Spanish words (Redondo, Fraga, Padrón, & Comesaña, 2007). This corpus was constructed by translating English words in the ANEW corpus to their Spanish equivalent and then the affect values of the words were rated by native Spanish speakers across three universities in Spain. Across 1,034 words, Redondo et al. (2007) reported that the Pearson's *r* correlation coefficient between the ratings of the words in English and their Spanish translation equivalent, was .916. The robust correlation suggests that the affect rating for words in English are comparable to its Spanish translation equivalent. For the present validation protocol, we report two methods in which we further test whether the affect ratings for English words can be used for their Spanish equivalents.

Method and Results

Each English word was translated into its Spanish equivalent using Google Translate. In our first validation method, we also compared the precision of using Google Translate in translating English words to its Spanish equivalent. Redondo et al. (2007) used trained linguists to translate English words; however, such an approach is not feasible when one is dealing with a large corpus. Using Google Translate, we were able to match 88% (906/1,034) of the words in our corpus to those used by Redondo et al. (2007). The other 12% mismatch were due to errors in differences in part of speech (e.g., Google Translate provided the verb form of the word, whereas Redondo et al. used the noun form).

A comparison of the affect ratings derived automatically using our expansion method to those collected by human participants in Redondo et al. (2007), yielded a correlation of *r* = .905 (based on 906 words). When we removed 9 cases for which the standardized residuals was greater than 3.5 (i.e., outliers), the correlation increased slightly to *r* = .918 (see Figure 3). These results provide support for the claims that the affect values derived automatically using our expansion method are reliable and that using the English affect ratings for its Spanish translation equivalent is valid.

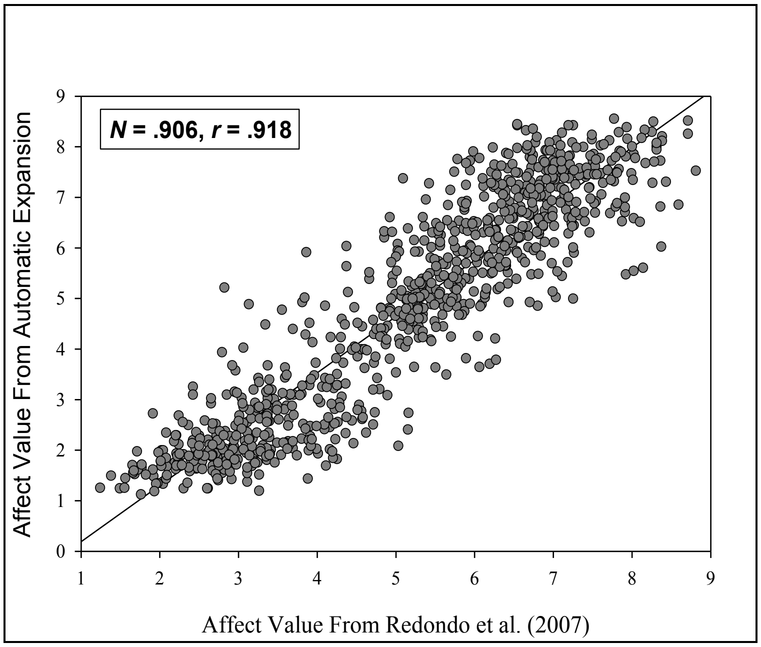


Figure 3: Scatterplot of affect ratings of Spanish words from the current expansion method compared to the values reported by Redondo et al. (2008).

To provide convergent evidence of the reliability of our expansion procedure, we conducted a second validation study in which we compared the affect ratings derived using our expansion method to those given by Spanish bilinguals recruited through Amazon Mechanical Turk. (The whole task was presented to participants in Spanish.) Only Turkers who had completed at least 500 tasks with a 99% approval rate were invited to participate. Each Turker rated 240 words, one at a time, with the words being presented in a different, randomized order for each Turker. Forty words were selected from the Redondo et al. (2007) norms. The 40 words that were selected ranged from being highly negative (e.g., tóxico [toxic], terrible [terrible], entierro [burial]) to highly positive (e.g., gatito [kitten], comer [eat], miel [honey]) and served as words for which we used to determine whether the participant was fluent in Spanish and/or was taking the task seriously. That is, if a participant were to indicate ratings for these words that were highly discrepant to those reported by Redondo et al. (2007), we would exclude this participant's data from the analysis. All the words used in this validation program (as well as those used for Russian and Farsi) were also in the second validation protocol (described above) for the English corpus. We used the same words (i.e., translation equivalents) for all languages so as to ensure that any differences in the results across the languages were not due to a different set of words used in one language but not the others.

Data from 18 participants were collected. Of these 18 participants, the data from 6 were discarded. One participant completed the task in eight minutes (the average time to complete the task was 29 minutes). The data for the other 5 participants were discarded because an examination of their ratings for the 40 words that were taken from Redondo et al.'s corpus (i.e., quality control words) raised suspicion that they were not fluent in Spanish and/or were not taking the task seriously. For example, these participants gave the same highly positive rating (i.e., 9) to words such as cárcel [jail], tóxico [toxic], terrible [terrible].

First, we considered the correlation for the ratings for the 40 words that were selected from Redondo et al.'s (2007) corpus. The correlation between the ratings given by participants in our study compared to those given by participants in Redondo et al.'s (2007) study was $r = .916$, $p < .001$. This robust correlation suggests that the ratings in our sample are reliable. We then considered the correlation between the ratings given by participants in our study compared to those derived using our automatic expansion method for the other 200 words. This analysis yielded a robust correlation of $r = .851$, $p < .001$ (see Figure 4), providing further evidence that our method of automatically computing affect ratings for Spanish words is valid.

Similar to the part of speech analyses for English, the correlation coefficient for adjectives, nouns, and verbs, were all high and comparable, .884, .845, and .849, respectively.

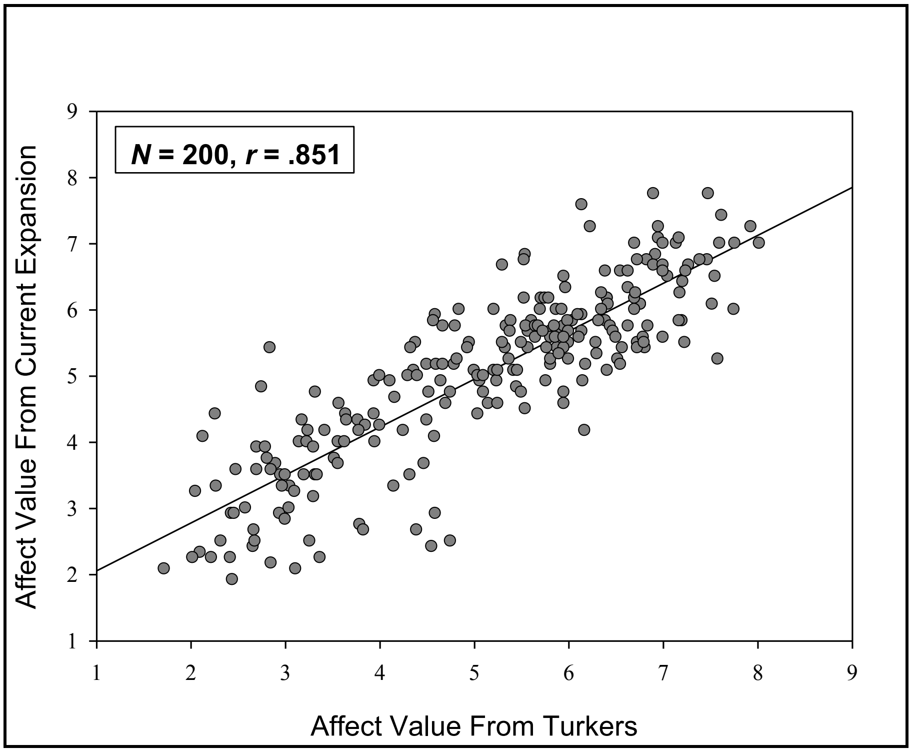


Figure 4: Scatterplot of affect ratings of Spanish words from the current expansion method compared to the values from Turkers.

Validation of Russian and Farsi Corpus

Unlike English and Spanish for which there are (some) existing resources that researchers can consult for affect norms (though, as noted above, these resources are highly limited), there is no resource for Russian and Farsi affect norms. As a result, a somewhat different validation procedure was needed. Because we used a nearly identical procedure to validate these two resources, they will be discussed together.

Method and Results

The 240 words used for validation were the same words (translation equivalents) used in the English and Spanish validations. As with the Spanish validation protocol, to assess the precision of Google Translate, we translated each word to its foreign language equivalent and had a trained native speaker of each language verify the translation provided by Google Translate. The percentage of English words that was incorrectly translated to Russian and Farsi were 5.4% and 10.3%, respectively. Thus, although there are differences in the level of precision of Google Translate for the two languages, the overall error rate for both languages is low. For the purposes of validating our corpora, we used the corrected translation of each word.

Fourteen Turkers who are fluent in Russian, and 5 Turkers who are fluent in Farsi participated in the validation experiment. Because it is more difficult to recruit Turkers who speak these two languages through Amazon Mechanical Turk, we allowed Turkers who had completed at least 100 hits with a 96% approval rate to participate in the study. To ensure that our participants were fluent in these two languages, we added a 10-item grammar test toward the end of the survey. The grammar test assessed participants' ability to detect common grammatical errors such as subject-verb agreement and word tense. For each sentence, participants had to indicate whether there was a grammatical error. (Five sentences contained an error.) Chance performance was 50%, and the data from Turkers whose score was below 60% were excluded from all analyses. Despite the small sample size (5) in our Farsi validation, when we computed the degree of agreement on the affect rating of the words among our sample, the intraclass correlation (inter-rater agreement, see McGraw & Wong, 1996; Shrout & Fleiss, 1979) yielded a coefficient of .84. (The coefficient value ranges from 0-1, with a higher value indicating greater agreement. A value of .70 is typically accepted as good agreement; thus, our obtained value of .84 indicates that the participants showed high level of agreement in their ratings of the words.)

The overall correlation of the affect values given by Turkers and those derived from our expansion method was .878 for Russian (after removing one outlier; see Figure 5), and .839 for Farsi (see Figure 6). Thus these results support the conclusion that affect values for English words are translated to their Russian or Farsi equivalent, the affect values for the words are largely retained.

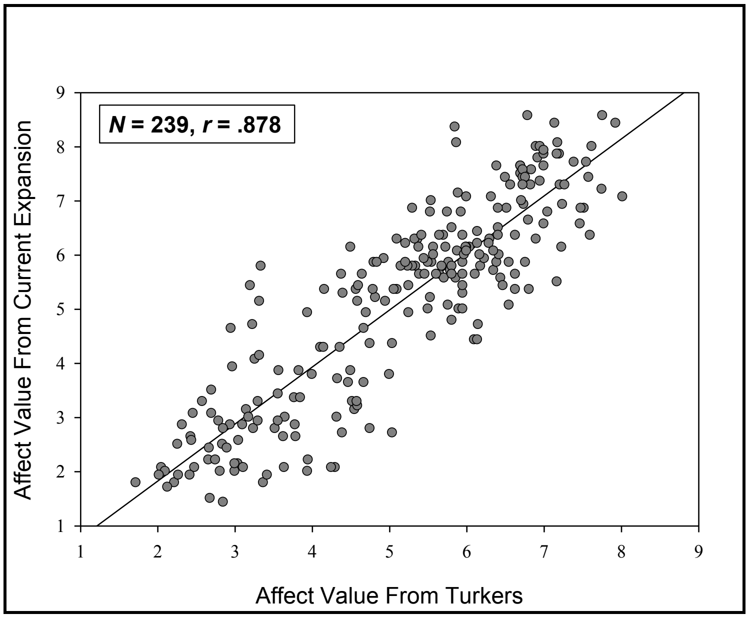


Figure 5: Scatterplot of affect ratings of Russian words from the current expansion method compared to the values from MTurkers.

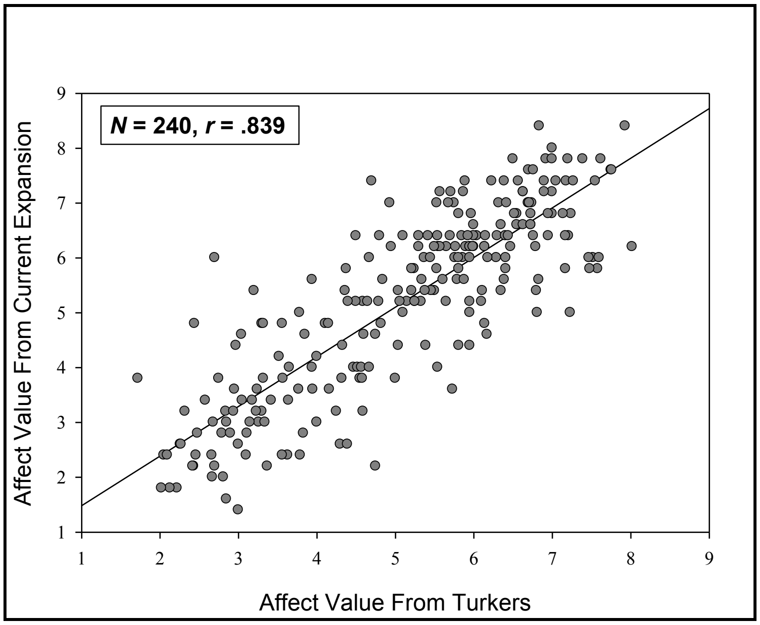


Figure 6: Scatterplot of affect ratings of Farsi words from the current expansion method compared to the values from MTurkers.

For the part of speech analyses, similar to the English and Spanish results, the correlation for adjectives, nouns, and verbs words were comparable and robust, .911, .867, .872. However, for Farsi, the correlation, though still robust, was significantly lower for adjectives (.772) relative to the comparable nouns (.842) and verbs (.899). (The difference between adjectives and verbs was significant at the .05 level, two-tailed, and the difference between adjectives and nouns being significant at the one-tailed level.)

Cross-cultural Comparisons

In this section, we list the 20 words that had the largest differences in affect ratings across languages. These data are presented in Table 1. Because the main purpose of the study was to validate our expansion method rather than test aprori hypotheses of differences in the affect of a specific set of words across cultures, we hesitate to make firm conclusions on the data presented in Table 1. However, we hope that these results will encourage future researchers to further explore these intriguing differences. For example, in our study, speakers of English, Spanish and Farsi rated the word "native" as being slightly positive (average: 6/9). However, Russian speakers rated "native" as being highly positive (8.36/9). One might speculate whether this difference reflects a difference in the degree of nationalism by speakers of Russian vs. English, Spanish, or Farsi.

Table 1: The top 20 words for which affect ratings differed across languages. Participants who took the foreign language version of the task received only the translated word.

	Language			
Word	**English**	**Spanish**	**Russian**	**Farsi**
argumentative	3.20	3.50	5.43	5.40
courtroom	2.84	5.42	2.50	3.20
daytime	7.58	5.25	7.43	5.80
derogate	2.70	3.92	3.50	6.00
evict	3.04	4.60	3.00	2.14
fond	6.79	6.20	5.58	8.57
flagrant	4.00	4.25	8.21	4.20
gloat	4.30	5.00	2.07	2.60
hospital	5.04	4.42	2.71	5.40
inject	3.34	3.50	5.79	3.00
livid	3.42	3.40	4.17	1.93
merciful	7.00	5.58	7.86	8.00
native	5.85	5.75	8.36	6.40
nursery	5.73	5.67	6.14	3.60
recycle	6.14	6.20	7.58	4.43
skirmish	4.25	3.20	4.17	2.07
sneaky	3.94	4.00	4.92	2.00
sociopath	2.44	1.92	2.57	4.80
stealer	2.13	4.08	1.71	1.80
uprise	6.17	4.17	6.00	4.60

Conclusions

Overall, we obtained robust correlations in the affect ratings of words that were automatically derived compared to those obtained using human participants. In principle, the current expansion method is appropriate to use on all studies that have gathered affective norms data using human participants and methods that are both valid and reliable. Although researchers who are interested in obtaining ratings for additional words can conduct their own normative study using procedures similar to those employed by Bradley and Lang (2009) and Warriner et al. (2013), which had a group of participants rate each word on its affect, such a procedure is not ideal because it may require a lot of resources. For example, Warriner et al.'s (2013) normative study of 14,000 words collected data from as many as 1,827 participants. Thus, valid and reliable methods to automatically compute affect ratings, an approach that we used to create the present corpus, is clearly a more desirable option because it requires fewer resources.

Our results also showed that the results from our method of expansion are generalizable to words in Spanish, Russian, and Farsi. At present, there is a very small corpus (about 1,000 words; see Redondo et al. (2007) for affect values for Spanish words, and there are no resources for Russian and Farsi words. Thus, the results of the present study should be of high interest to the scientific community. However, it should be noted that because we used Google Translate (rather than expert linguists) to translate English words to their foreign-language equivalent, we do not anticipate that all words will be translated accurately and thus the affect values for these words may be inaccurate. Based on the results of our study, we estimate that no more than 10% of the words will be incorrectly translated.

Our expansion technique also raises interesting questions for future researchers to investigate. One question to consider is whether our expansion method is also valid for other psychological constructs that have been collected using human participants. For example, the dimension of arousal (i.e., the intensity of emotion evoked by a word) is one variable that is of interest to many researchers. Another question is whether compound words (e.g., “holy scripture”) or short phrases (e.g., “word of god”) derived from our expansion method correlate with their source words (i.e., “bible”).

MRC Imageability Expansion

Words differ on many properties. One of the more widely studied properties is its imageability, which refers to the extent that the word can be experienced through one or more of our senses. For example, a word such as “mountain” is highly concrete because we can easily imagine and evoke a mental image of a mountain. A word such as “sugar” is also imageable because the *sensation* associated with sugar can be easily experienced by having someone taste sugar. Words that are not imageable cannot be easily experienced through our senses and demonstrated. These words include “justice” and “romantic”, and are words (or concepts) that often have to be explained using other words in the lexicon. For example, to explain the word “justice” to someone, you may tell that person to think of a courtroom with lawyers debating in front of a judge and jury.

Imageability of words is of high interest to researchers because they affect many cognitive processes. For example, words high in imageability are more memorable (Pavio, 1971), are acquired/learned at an earlier age (Morris, 1981), and are more likely to be used in metaphorical language (Broadwell et al., 2013). Coltheart (1981) created a corpus (MRC Psycholinguistic Database) of imageability ratings for 4,808 (unique entries) English words. The data for this corpus come from a various previously-published papers by other researchers, all of whom collected imageability ratings by asking participants to rate each word on its degree of imageability, typically on a scale from 1 (low imageability) to 5 (high imageability)[1], and to date, the MRC corpus is the most widely known and used corpus among researchers who are interested in obtaining imageability ratings. More recently, Brysbaert, Warriner, and Kuperman (2014) created a highly comprehensive corpus by expanding Coltheart’s (1981) corpus to 37,058 words.[2] Concreteness ratings, which are highly correlated with imageability ratings (see Footnote 2), have also been obtained for languages other than English. For example, Della Rosa et al. (2010) created their own corpus of 417 Italian words, collecting their data using a procedure similar to the other aforementioned corpora.

The REMND Project introduced a new procedure to gather imageability ratings for single words. In all the aforementioned studies, the researchers had human participants rate each English word on its imageability value. Although this is a very straightforward method for researchers who wish to gather normative data for their own set of words, it is not ideal because it is time consuming, and in some cases, costly. As an example, Brysbaert et al. (2014)’s corpus was created by creating 210 lists (surveys) of 300 words and the words were rated by a total of 4,237 participants who were recruited online using Amazon Mechanical Turk and were monetarily compensated in exchange for their participation. In REMND we used a new, automated method to gather imageability ratings for words. Using an automated approach has the advantage of gathering data quickly and with fewer resources; however, this method requires an independent validation step

[1] We have normalized all scores to fall within the (0, 1) range

[2] Although Brysbaert et al. (2014) collected ratings for the dimension of concreteness, previous research has shown that concreteness ratings are highly correlated with imageability ratings (e.g., *r* = .83, Pavio, Yuille, and Madigan, 1968); thus, many researchers, including Brysbaert et al. (2014) have used these two terms interchangeably. In addition, the instructions participants receive for imageability and concreteness normative studies are highly comparable in that participants are told that words that are high in imageability (or concreteness) should arouse a sensory experience.

to guarantee reliability and validity of the results. The validation method we used is described below.

The second purpose of our analysis was to test whether the imageability ratings of English words are retained when they are translated into their foreign-language equivalents. The present study explored this question by comparing the imageability ratings for English words to their Spanish-, Russian-, and Farsi-translation equivalents. To our knowledge, there is only one corpus of imageability ratings for non-English words (Italian; Della Rosa et al., 2010) and that corpus is very limited (417 words). One reason why imageability ratings are not readily available for other languages may be due to resources required in order to gather these data. Thus, a validation that shows that imageability ratings for English words can generalize to their foreign-language translation equivalent will allow researchers to study imageability of words in other languages.

This chapter is divided into two main sections. We first briefly describe our method to automatically compute imageability ratings for English words (full description can be found in the REMND Project Final Report). We then describe the procedure that we used to validate our expansion method. The second section describes the procedure we employed to test whether imageability ratings for English words are generalizable to their foreign-language (Spanish, Russian, and Farsi) equivalents.

Automatic Computation of Imageability Ratings for English Words

Our expansion method relies on imputing imageability values of words (e.g., dog) found in the MRC Psycholinguistics Database to their synonyms and hyponyms (e.g., puppy, pooch, mutt). Synonyms and hyponyms were identified using Princeton's WordNet (Miller, 1995), which is a large English lexical database with over 150,000 words, hierarchically organized into synsets that capture semantically equivalent words (synonyms). To provide an example of how our expansion method works, suppose that we need to find the imageability score for the word "somebody". WordNet places "somebody" in a synset along with the following words: "person", "individual", "mortal", and "soul". If the imageability values of any of these words are known, they will be averaged to obtain the value for "somebody", as well as any other unscored word in this synset. The underlying assumption here is that words assigned to a synset, as synonyms, would have the same or very close imageability scores, while their hyponyms would have scores that can only be higher as we move down the hierarchy. Thus our expansion method is fairly conservative.

To validate our automatic expansion method, we compared the values obtained using our expansion method to those collected by human participants, as was the case in the construction of the Brysbaert et al.'s (2014) corpus. The simplest and most direct way to assess the validity and precision of our expansion method is to compute a correlation coefficient of the imageability values for words present in both corpora. A high agreement between the values in the two corpora would be supported by the presence of a positive correlation. It should be noted that Brysbaert et al.'s (2014) study included words that are found in the MRC Psycholinguistics Database. They reported a robust correlation between the values of the words from both corpora, r = .919. This finding suggests that the results from Brysbaert et al. (2014) are valid and that collecting data online using Amazon Mechanical Turk (Turkers) is a valid alternative to collecting data using college undergraduates (MRC Psycholinguistics Database).

Validation Method

The test the validity of our expansion, we randomly selected 176 words that were in our and in Brysbaert et al.'s (2014) corpora. Of these 176 words, the majority were nouns (n = 74), followed by verbs (n = 69), with the fewest being adjectives (n = 33). The frequency of occurrence (measured using Log HAL, see Balota et al., 2009) of these three different type of words were statistically equivalent, F = .844, p = .732 (Mean$_{\text{ADJECTIVES}}$: 7.42, Mean$_{\text{NOUNS}}$: 7.08, Mean$_{\text{VERBS}}$: 6.87).

Results and Discussion

In all analyses reported in this paper, outliers, defined as a data point with a standardized residual greater than 3.5 were removed from all analyses. After removing one outlier, the overall correlation between the values derived from our expansion method (obtained automatically) with those obtained by Brysbaert et al.'s (2014) using human raters (Turkers) was moderate in strength, r = .652. However, analyzing the data by the word's part of speech revealed large variability in the level of precision of our automatic expansion method. As can be seen in Figure 7, the level of precision of our expansion method was much higher for Nouns (r = .762) and Verbs (r = .617) than for Adjectives (r = .228). The difference between Nouns and Adjectives and Verbs and Adjectives were both statistically significant, p < .001, two-tailed. The difference between Nouns and Verbs was just shy of statistical significance, p = .051.

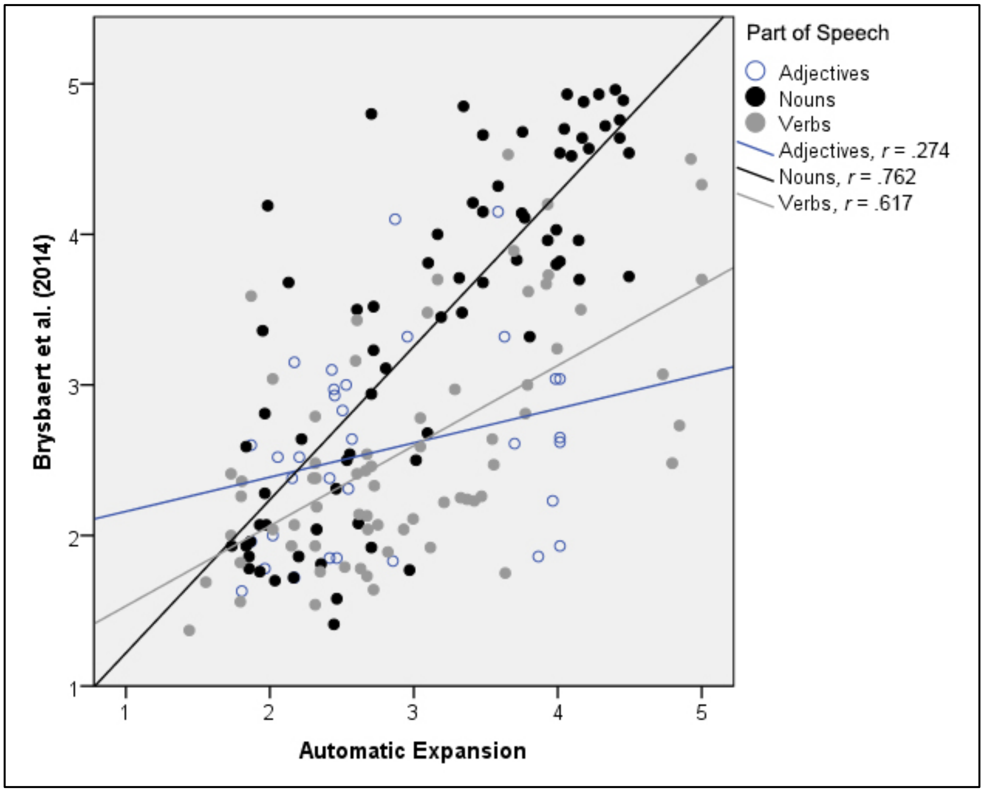
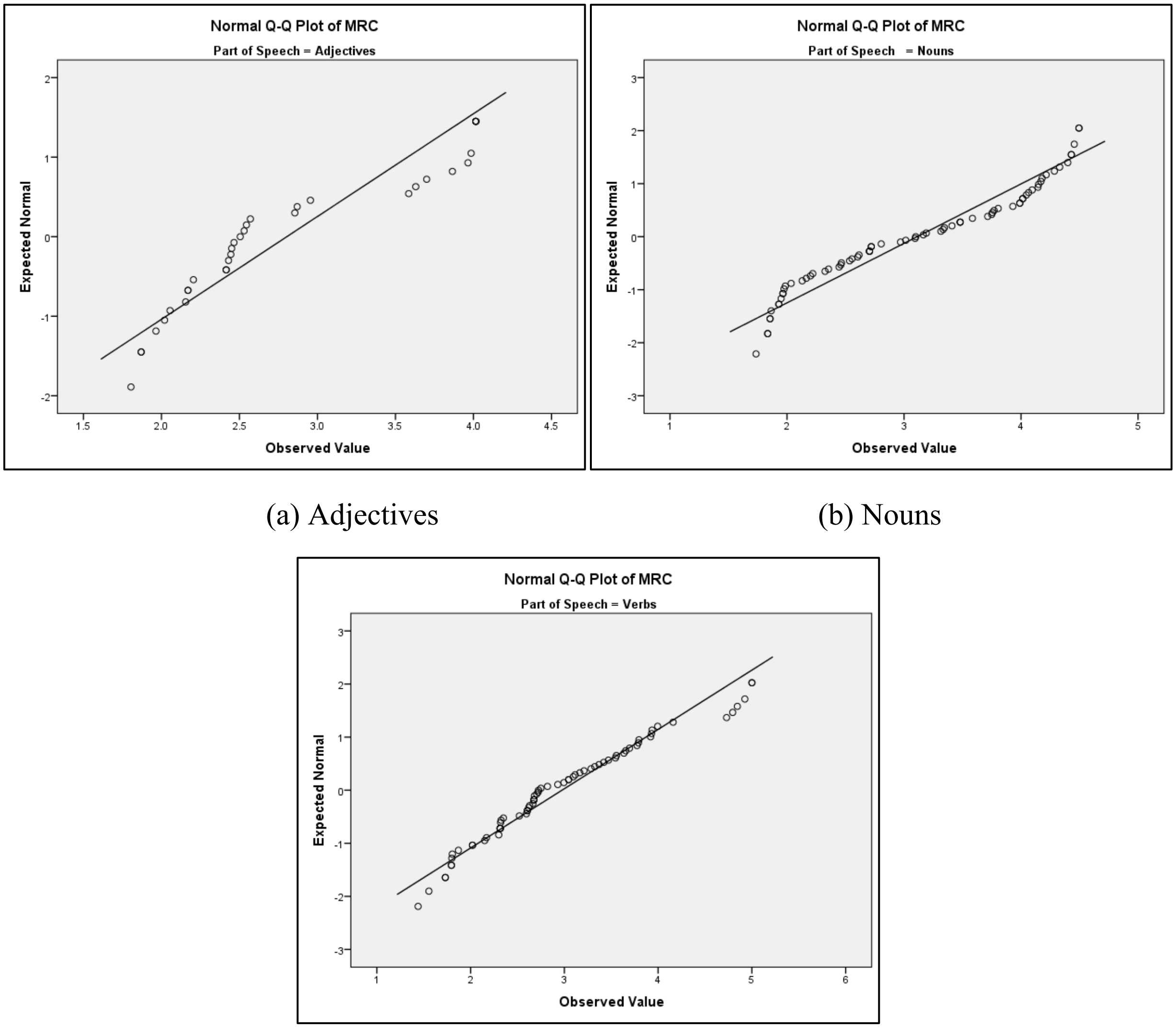


Figure 7: Imageability values for words obtained using an automatic expansion method vs. values reported by Brysbaert et al. (2014) as a function of the word's part of speech (adjectives, nouns, or verbs).

A closer examination of the distribution of the imageability values in our expansion shows that the values for adjectives deviated from normality more so than the distribution of values for either nouns and verbs (see Figure 8).



(a) Adjectives

(b) Nouns

(c) Verbs

Figure 8: Normality distribution plots of the MRC Expansion Values for (a) Adjectives, (b) Nouns, and (c) Verbs.

Thus, although the overall correlation of the affect values obtained in our expansion method with those obtained manually using human raters (via Brysbaert et al., 2014) is moderately strong, because the correlation for adjectives was weak, at present, we recommend only using data for nouns and verbs.

One reason for relatively low correlation may be the procedure that was followed in assigning scores through expansion. In this method, all imageability scores were propagated on all words within a Wordnet synset, starting from the known core (the original MRC) and propagating through the Wordnet hierarchy. When more than one word had a prior score in a synset, they were divided into these that were above 0.7 threshold (determined to be a strong indicator for metaphors) and those that fell below it. The score assigned to the remaining words in the synset (and also propagated to hyponyms) was then based on the average score from the larger of the two groups, or the overall average if the groups were of equal size. We noticed that, as a result, some words in the synset got their imageability scores significantly out of sync with human assessment (mostly higher or less often lower). This appears to mean that the synonyms placed in at least some synsets have greatly varied levels of imageability and concreteness. This is unexpected, and may potentially point to flaws in the Wordnet design. Possibly this suggests that some highly concrete synsets were "padded" with less concrete or highly ambiguous "synonyms", which should have been placed in their own synsets (the example used above of *person* synset seems to support this observation, e.g., *person: .94, individual: .70, mortal: .39, soul: .37*). We should note that this phenomenon does not occur with ANEW, where synonyms appear to carry comparable polarity scores.

Imageability Ratings Across Languages

To test whether imageability ratings for English words are retained when translated into their foreign-language equivalents, we started with the same 176 words noted above, but increased the size of our sample such that we had 75 words from each part of speech, thus yielding a total sample of 225 words. These 225 words were translated into their foreign-language (Spanish, Russian, and Farsi) translation equivalent using Google Translate. For each language, we had a trained linguist validate the translation provided by Google Translate. Overall, the translation provided by Google Translate were largely acceptable, and only 8.4% of the words in Spanish, 4.9% of the words in Russian, and 5.3% of the words in Farsi, were corrected by our trained linguists.

The translated words were rated by human participants (Turkers), fluent in that language. Although Brysbaert et al. (2014) showed that researchers could collect high-quality imageability ratings from Turkers, to further ensure that we would attract Turkers who would take the task seriously, only Turkers who have completed at least 500 surveys with a 99% approval rate were invited to participate. To ensure that participants were indeed fluent in the foreign-language of interest, we administered a grammar test at the end of the survey. The grammar test assessed MTurkers' ability to identify common grammatical errors (e.g., inconsistent use of word tense, incorrect syntax, subject-verb agreement errors) in single sentences, and only the data from MTurkers who scored 60% or higher on this test were analyzed.

We assessed whether imageability values for words in English are retained when translated into their foreign-language equivalent by computing a correlation between the imageability values of the words in English and their foreign-language equivalents. For English words imageability values, we used the values collected by Brysbaert et al. (2014), and for the foreign-language translation imageability values, we collected the ratings from Turkers, as noted above. Furthermore, to increase the similarity in the data collection protocol between our surveys and Brysbaert et al.'s (2014), we translated their English instructions to their foreign-language equivalent.

Results and Discussion

The data for the Spanish, Russian, and Farsi are based on ratings obtained from 15, 12, and 3, Turkers, respectively. Because the number of raters for the Farsi was low (partly due to a small number of available testers and then rigorous screening criteria we implemented), one might be concerned by the reliability of our results. To address this issue, we computed the degree of agreement among the raters in Farsi (inter-rater agreement, see McGraw & Wong, 1996; Shrout & Fleiss, 1979). Our analysis yielded an inter-rater agreement of .856. A value of .70 is typically accepted as good agreement; thus, our obtained value of .856 indicates that the participants showed high level of agreement in their ratings of the words. The high level of agreement is not too surprising given the rigorous screening criteria that we used to ensure that we recruited only fluent speakers of Farsi in our study. The data for all languages are displayed in Table 2.

Table 2: Correlations between imageability ratings of words in English and their foreign-language translation equivalents. Within each column in the part of speech categories, values that share the same superscript are statistically significant at the .05, two-tailed level.

	Spanish	Russian	Farsi
Overall	.848	.781	.680
Part of Speech			
Adjectives	$.762^{a}$	$.667^{a}$	$.382^{ab}$
Nouns	$.871^{ab}$	$.847^{ab}$	$.731^{a}$
Verbs	$.742^{b}$	$.570^{b}$	$.630^{b}$

First, we consider the overall correlations (see the top portion of Table 2). Overall, the correlations were high, thereby showing that the imageability values for English words are largely retained when they are translated to a foreign language, specifically, Spanish, Russian, or Farsi. The overall correlation was greater for Spanish (.848) than for Russian (.781, $p = .034$) and Farsi (.680, $p < .001$). Within each language, there are differences in the correlation as a function of the words' part of speech. Specifically, the imageability values of nouns are retained much more than both adjectives and verbs. One reason for this difference could be that referents (or meanings) of nouns are much easier to hone-in and these meanings are typically unambiguous. For example, the meanings of words such as "peasant" and "mountain" are likely to activate very similar representations for all languages, whereas an adjectives such as "flying" and "glazed" or verbs such as "pierce" and "learning" are more likely to evoke different representations by individuals and therefore judgments to these words are more variable across individuals and across cultures.

Conclusions

The study reported in this chapter sought to validate an automatic method to derive imageability ratings for English words and to test whether imageability ratings for English words are correlated with their foreign-language translation equivalents. Although imageability values obtained by our automatic expansion method correlated modestly with values obtained using human participants (*r* = .652), the correlation was much stronger for nouns (*r* = .762) and verbs (*r* = .617) than for adjectives (*r* = .274) Thus, at present, we recommend using our automatic expansion values for nouns and verbs. We should note, however, that the relatively lower correlation is partially explained by larger than expected variability of imageability scores for words classified as synonyms in the Wordnet lexical database. This unexpected discovery may potentially point to some design flaws in Wordnet, and would merit further investigation.

Overall, and more importantly, our method shows that the automatic imputation of values to words via a Wordnet lexical hierarchy is indeed a valid approach and one that we hope future researchers would exploit to automatically derive other lexical characteristics for words (e.g., affect, arousal). We also showed that imageability values for words in English are largely preserved when they are translated into their foreign-language equivalents. This finding presents opportunities for researchers to study how imageability of words affects cognitive processes in languages other than English. Although the present study only examined three languages Spanish, Russian, and Farsi, we note that they represent languages from different language families (i.e., Romance, Slavic, and Iranian). Thus, we have reasons to believe that imageability ratings for English words would generalize to foreign languages besides the one examined in this report.

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-0024. The U.S. Government is authorized to reproduce and distribute reprints or 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

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459.

Barber, H. A., Otten, L. J., Kousta, S. T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language, 125*, 47–53.

Bradley, M., & Lang, P.J. (1999). Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical report C-1, Gainesville, FL. The Center for Research in Psychophysiology, University of Florida.

Broadwell, G. A., Boz, U., Cases, I., Strzalkowski, T., Feldman, L., Taylor, S., & Webb, N. (2013). Using imageability and topic chaining to locate metaphors in linguistic corpora. In *Social Computing, Behavioral-Cultural Modeling and Prediction* (pp. 102-110). Springer Berlin Heidelberg.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology, 33*, 497–505.

Della Rosa, P. A., Catricalà, E., Vigliocco, G., & Cappa, S. F. (2010). Beyond the abstract—concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. *Behavior Research Methods, 42*, 1042–1048.

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

Miller, G. A. (1995). WordNet: A Lexical database for English. Communications of the ACM, 38(11): 39-41.

Morris, P. E. (1981). Age of acquisition, imagery, recall, and the limitations of multiple-regression analysis. *Memory & Cognition, 9*, 277–282.

Paivio, A. (1971). Imagery and verbal processes. New York: Holt, Rinchart, and Winston.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76*, 1-25.

Redondo, J., Fraga, I., Padrón, I., & Comesaña, M. (2007). The Spanish adaptation of ANEW (affective norms for English words). *Behavior Research Methods, 39*, 600-605.

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

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*, 1191-1207.