The Validation of MRCPD Cross-language Expansions on Imageability Ratings

Ting Liu[1], Kit Cho[2], Samira Shaikh[1], Mehrdad Mirzaei[1], Tomek Strzalkowski[1]

[1] University at Albany, State University of New York

[2] University of Houston-Downtown

E-mail: tliu@albany.edu, kitwcho@gmail.com

Abstract

In this article, we present a method to validate a multi-lingual (English, Spanish, Russian, and Farsi) corpus on imageability ratings automatically expanded from MRCPD (Liu et al., 2014). We employed the corpus (Brysbaert et al., 2014) on concreteness ratings for our English MRCPD+ validation because of lacking human assessed imageability ratings and high correlation between concreteness ratings and imageability ratings (e.g. r = .83). For the same reason, we built a small corpus with human imageability assessment for the other language corpus validation. The results show that the automatically expanded imageability ratings are highly correlated with human assessment in all four languages, which demonstrate our automatic expansion method is 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.

Keywords: MRC Psycholinguistic Database (MRCPD), wordnet, imageability, validation

1. Introduction

Scientists engaged in research using lexical stimuli must take into account a multitude of variables associated with words (e.g. familiarity, frequency of occurrence) that could potentially obscure the interpretation of their results. The Psycholinguistic Database (MRCPD) (Wilson, 1988) provides 26 linguistic and psycholinguistic variables for over 150 thousand English words and has served as a valuable resource for researchers in a variety of disciplines such as linguistics, psychology (Coltheart, 1981), and computational linguistics (Broadwell et al., 2013, Strzalkowski et.al. 2013). However, not all words have ratings for all 26 variables. For example, only 9,240 (6%) of the total words in MRCPD have an *imageability* value (i.e., how easily and quickly the word evokes a mental image).

The traditional way of analyzing the variables requires human analysis and judgments, which is accurate but time consuming and expensive (Higgins, Xi, Zechner, & Williamson, 2011, Brysbaert et al. 2014). Therefore, in Liu et al. (2014), we presented an approach to automatically expand the MRCPD database by adding imageability ratings for an additional 106,911 words and subsequently translating the expanded MRC into Spanish, Russian, and Farsi. Using an automated approach has the advantage of gathering data quickly and with fewer resources. However, as with any other method, validation of the method is required. In the present study, we report two independent validation techniques and 1. assessed the reliability and validity of the expanded results, and 2. tested whether the imageability ratings of English words are retained when they are translated into their foreign-language equivalents.

The present study explored the validation by comparing the imageability ratings for English words to their Spanish-, Russian-, and Farsi-translation equivalents. To our knowledge, there are only two corpora of imageability ratings for non-English words (Italian – Della Rosa et al., 2010; Spanish - Sebastián Gallés, 2000) and those corpora are quite limited (417 words for Italian corpus and 6K for Spanish corpus). 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 article is structured as follows: Section 2 describes how imageability is used in previous research and the resources available to obtain imageability ratings for English words. Section 3 gives the algorithm of our expansion method to expand the English corpora within and across languages. Section 4 demonstrates result validation.

2. Imageability in language research

Imageability of words is of high interest to language researchers because imageability affects human 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 and concreteness ratings for 9,240 English words. The data for this corpus come from 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 100 (low imageability) to 700 (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. Although Brysbaert et al. (2014) collected ratings for the dimension of concreteness only, 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. Therefore, this corpus could be a good candidate for us to validate our expanded MRCPD on imageability ratings.

3. MRCPD expansion and validation

3.1 MRCPD expansion in English and across Languages

Our expansion method relies on imputing imageability values of words (e.g., dog) found in the MRCPD to their synonyms and hyponyms (e.g., puppy, pooch, mutt) from their most frequently used senses (Liu et al. 2014). 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).

Creating an extended imageability lexicon for other languages would require the same procedure if there is a human assessed imageability corpus as seed and a fully functioned electronic lexicon for expansion. However, no other language has all required resources. Therefore, an alternative way is to translate the expanded English MRCPD+ database into another language through a mechanical process, e.g., Google Translate, with scores averaged in case of many-to-one translations. Table 1 shows the size of translated MRCPD+ in four languages

Language	Size of translated MRCPD+
English	126,693
Spanish	127,591
Russian	125,691
Farsi	101,167

Table 1. MRCPD+ on Imageability Ratings in 4 languages

3.2 Validation method

To validate our automatic expansion method, we compared the values obtained using the 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, $r = .919$ (1st column in Table 2).

To test the validity of our expansion, we selected 5500 words that were both in our expansion and in Brysbaert's and were listed in both with the same part of speech and only one part of speech. We add the condition of only one part of speech because all words in Brybaert's corpus have only one single part of speech. Of these words, the majority were nouns (2880), followed by verbs (2101), with the fewest being adjectives (519).

	MRC ori	MRC exp
Overall	.919	.673
Nouns	.930	.766
Verbs	.837	.563
Adjectives	.835	.686

Table 2 The Correlations between MRCPD original/expanded and Brysbaert et al.'s (2014) Corpus

3.3 Results and discussion

We calculated the correlations between the values derived from our expansion method (obtained automatically) with those obtained by Brysbaert et al.'s (2014) using human raters (Amazon MTurkers). The overall correlation was moderate ($r = 0.673$). However, analyzing the data in more detail revealed large variability in the level of correlation by part of speech. As shown in the second column of Table 2, the correlation was much higher for nouns ($r = .766$) and adjectives ($r = .692$) than for verbs ($r = .563$).

To figure out what could cause the relatively low correlation for verbs, we extracted all the verbs from MRCPD used to expand the 2101 verbs for the above set. Then we calculated the correlations of the verbs' imageability rates between MRCPD and Brysbaert. It turned out the correlation is only .687, which is not as

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

strong as the correlations from nouns (.86) and adjectives (.817)[2]. Therefore, it is reasonable that verbs correlation is lower than nouns' and adjectives'.

The second issue we found is from the Wordnet synonym design. Synonyms express the same meanings and therefore should be expected to convey similar imageability and concreteness. Accordingly, our expansion method would impute the same score for all words in a synset. However, we noticed that in some cases, the synonyms placed in at least some synsets have greatly varied levels of imageability and concreteness (As an example, "person" synset includes synonyms with quite different imageability values in the original MRC, person: .803, individual: .629, mortal: .574), which led to assigning imageability scores that were significantly out of sync with human assessment.

4. 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 208 words randomly selected from expanded corpus noted above (nouns: 74, verbs: 69, adjectives: 75), which were translated into their foreign-language (Spanish, Russian, and Farsi) translation equivalent using Google Translate and only 6% of translations were corrected by our linguist experts[3].

The translated words were rated by human participants (Amazon Mechanical Turk workers – Spanish: 15, Russian: 12, Farsi: 3) who are fluent in that language. Because the number of raters for Farsi was low, 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, which is much higher than .70, typically accepted as good agreement.

Table 3 Correlations between imageability ratings of words in English and their foreign-language translation equivalents

	Spanish	Russian	Farsi
Overall	.848	.781	.680
Adjectives	.762	.667	.484
Nouns	.871	.847	.731
Verbs	.742	.570	.630

Table 3 displays the correlations between imageability ratings of words in English and their foreign-language translation equivalents. We used the values collected by Brysbaert et al. (2014) as English imageability rates and for the foreign-language translation imageability values, we collected the ratings from Turkers, as noted above.

The overall correlations were high across all three languages, thereby showing that the imageability values for English words are largely retained when they are translated to a foreign language. 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 better than either adjectives or verbs. One reason for this difference could be that referents (or meanings) of nouns are much easier to hone-in and these meanings are largely independent of cultural context. For example, the meanings of words such as "peasant" and "mountain" are likely to activate very similar representations for all languages, whereas adjectives such as "flying" and "aural" 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. As an example, in English the imageability score for "aural" is .50 (moderate), while in, Farsi it is .93 (very high).

5. Conclusions

The present study 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 the overall imageability values obtained by our automatic expansion method correlated modestly with values obtained using human participants (r = .673), the correlation was stronger for nouns (r = .766) and adjectives (r = .686) than for verbs (r = .563).

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.

[2] It is same as we calculate correlation on verbs between MRCPD and Brysbaert corpus

[3] We build all our own validation data because there's no resources (in Spanish, Russian, and Farsi) available for validation.

6. Acknowledgement

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.

7. References

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*, 290-313.

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchinson, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D. L., Simpson, G.B., & Treiman, R. (2007). The English lexicon project. *Behavioral Research Methods*, 39(3): 445-459.

Bird, H., Howard, D., & Sue, F. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function *words. Behavioral Research Methods, Instruments, and Computers*, 33(1): 73-79.

Broadwell, G.A., Boz, U., Cases, I., Strzalkowski, T., Feldman, L., Taylor, S., Shaikh, S., Liu, T., & Cho, K. (2013).
Using Imageability and Topic Chaining to Locate Met aphors in Linguistic Corpora. In *Proceedings of International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction.* 2013. Washington D.C.

Brysbaert, M., Warriner, A.B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46, 904-911.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3^{rd} ed.). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, 33A: 497-505.

Google Scholar. (2012). Retrieved from http://scholar.google.com.

Green, M. C., & Brock, T.C. (2002). In the mind's eye: Transportation-imagery model of narrative persuasion. In Melanie C. Green, Jeffrey J. Strange, and Timothy C. Brock (Eds.), *Narrative impact: Social and cognitive foundations* (315-341). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Hargreaves, I.S., & Pexman, P.M. (2012). Does richness lose its luster? Effects of extensive practice on semantic richness in visual word recognition. *Frontiers in Human Neuroscience*, 6: 1-11.

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the auto- mated scoring of spontaneous spoken responses. *Computer Speech and Language*, 25(2), 282–306. doi: 10.1016/j.csl.2010.06.001

Liu, Ting, Kit Cho, George Aaron Broadwell, Samira Shaikh, Tomek Strzalkowski, John Lien, Sarah Taylor, Laurie Feldman, Boris Yamrom, Nick Webb, Umit Boz, Ignacio Cases, Ching-sheng Lin. (2014) Automatic Expansion of the MRC Psycholinguistic Database Imageability Ratings. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).

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

Pavio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review,* 76(3): 241-263.

Ratcliff., R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, 33: 510-532.

Rossiter, J. R. (1982). Visual imagery: Applications to Advertising. In Andrew Mitchell (Ed.), *Advances in Consumer Research Vol. 9: Association for Consumer Research*: 101-106.

Sebastián Gallés, Núria (coord.), M. Antònia Martí Antonín, Manuel Francisco Carreiras Valiña, and Fernando Cuetos Vega. LEXESP, Léxico Informatizado del Español Programa CORCO. Barcelona: Edicions de la Universitat de Barcelona, 2000.

Strzalkowski, T., Broadwell, G.A., Taylor, S., Feldman, L., Shaikh, S., Liu, T., Yamrom, B., Cho, K., Boz, U., Cases, I., and Elliot, K. (2013.) Robust Extraction of Metaphor from Novel Data. *In Proceedings of Workshop on Metaphor in NLP, NAACL.* 2013 Atlanta.

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