A Text Corpora-Based Estimation of the Familiarity of Health Terminology

Qing Zeng[1], Eunjung Kim[1], Jon Crowell[1], Tony Tse[2]

Decision Systems Group, Harvard Medical School and Brigham & Women's Hospital, Boston, MA
{qzeng, ejkim, jcrowell}@dsg.harvard.edu
Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD
tse@ nlm.nih.gov

Abstract. In a pilot effort to improve health communication we created a method for measuring the familiarity of various medical terms. To obtain term familiarity data, we recruited 21 volunteers who agreed to take medical terminology quizzes containing 68 terms. We then created predictive models for familiarly based on term occurrence in text corpora and reader's demographics. Although the sample size was small, our preliminary results indicate that predicting the familiarity of medical terms based on an analysis of the frequency in text corpora is feasible. Further, individualized familiarity assessment is feasible when demographic features are included as predictors.

1 Introduction

Health literacy is "the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions" [1]. A large percentage of the US population has low health literacy, which has been linked to poor outcomes in health care [2,3,4,5]. Although health literacy includes reading, writing, speaking, listening, and numeric, cultural and conceptual knowledge components, reading comprehension of health terms and concepts has been the primary focus of health literacy tests.

Literacy is closely coupled to readability, a measure of the difficulty in reading text. Outside the health domain, grade level has been used to benchmark both a person's literacy level and the readability of text. Within the health domain, health literacy tests are modeled on functional literacy measures and grade-level scores [6,7]. For example, the two most frequently used tests – Test of Functional Health Literacy Assessment (TOFHLA) and Rapid Estimate of Adult Literacy in Medicine (REALM) – assign 3 and 4 levels of health literacy, respectively. However, because there is no comparable readability measure for health-related text, health care researchers rely on readability formulas developed since the 1940's for assessing general-domain text [8]. Key factors of these general readability formulas include vocabulary difficulty, sentence structure complexity, and text cohesion. Although existing formulas provide

In: Oliveira JL, Maojo V, Martin-Sanchez F, Pereira AS, editors. Biological and Medical Data Analysis: 6th International Symposium (ISBMDA 2005) Proceedings; 2005 Nov 10-11; Aveiro, Portugal. Berlin: Springer; 2005. p. 184-92.

"rough" estimates of readability for health-related text [9], a more accurate approach is needed, since most computer-based health communication is textual.

Assessing term difficulty is clearly a shortcoming of applying general readability measures to health-related content. The usual techniques such as counting the number of letters or syllables in words, or appearance on "easy word lists," often do not apply to health-related text, which typically contains technical terms. For instance, common words such as *operation* have more letters and syllables than technical terms like *femur*. Further, although *diabetes* and *menopause* are commonly recognized terms due to their prevalence and media coverage, general readability formulas consider them "unfamiliar" along with truly difficult terms such as *atelectasis* and *alveoli*. We focus exclusively on health terminology because sentence complexity and cohesion are domain-independent factors.

Ideally, data on vocabulary comprehension of health terms in the population would be obtained for such a study. Since it is not feasible to test *all* health terms on every subpopulation, we conducted a preliminary study using a convenience sample to estimate familiarity with a set of 34 terms that frequently occurred in health-related text corpora.

2 Methods

2.1 Familiarity of Sample Terms

We had previously created an instrument for evaluating health vocabulary familiarity as part of the ongoing Consumer Health Vocabulary Initiative project. The questionnaire, modeled on the TOFHLA, contained 34 multiple-choice questions, each assessing a commonly used health-related concept [10].

Two synonyms were selected for each concept: a consumer-friendly but precise term or Consumer-Friendly Display (CFD) name (*kneecap*) and an ambiguous term or jargon (*patellae*). We created two versions of the questionnaire, each consisting of 17 statements containing CFD names and 17 with jargon. For example, statement 1 below contains jargon in version A (*geriatric*) and a corresponding CFD name in version B (*elderly*).

Version A:

1. A geriatric person is one who is _____.
    A. Very old
    B. lanky and good looking
    C. well groomed
    D. aggressive and loud

2. You are in trouble when alcohol is detected _____.
    A. in your skin while you're sun bathing
    B. in your eye while you're reading
    C. in your heart while you're exercising
    D. in your blood while you're driving

3. If you have a cerebrovascular accident it means that _____.
    A. you broke a bone
    B. you were unable to make it to the bathroom on time
    C. you had a heart attack
    D. a blood vessel in your brain ruptured or clogged

Version B:

1. An elderly person is one who is _____.
    A. Very old
    B. lanky and good looking
    C. well groomed
    D. aggressive and loud

2. You are in trouble when ethanol is detected _____.
    A. in your skin while you're sun bathing
    B. in your eye while you're reading
    C. in your heart while you're exercising
    D. in your blood while you're driving

3. If you have a stroke it means that _____.
    A. you broke a bone
    B. you were unable to make it to the bathroom on time
    C. you had a heart attack
    D. a blood vessel in your brain ruptured or clogged

We did not use TOFHLA for this preliminary study because it contains only a few difficult terms. This questionnaire, on the other hand, assesses a variety of technical terms.

The questionnaire was administered to a convenience sample of 21 people recruited from the Brigham and Women's Hospital in Boston and local churches. The inclusion criteria were: non-clinician, 18 years of age or older, and the ability to read and write in English. Each participant also provided demographic information: age, gender, race, ethnicity, first language, profession, and education level.

The authors used the completed questionnaires to calculate term familiarity scores. Each statement was given a score of 1 if completed with the correct term; otherwise, it was given a 0. We then estimated the familiarity score for each term across the population by averaging all scores for that term across the sample.

To obtain baseline data for term familiarity, we employed two methods commonly used by general readability formulas: (1) counting syllables per word and (2) consulting the Dale-Chall List [11]. Words that contain 3 syllables or more have been deemed "difficult" by some readability measurements. Words that do not appear on the Dale-Chall list, which contains about 3,000 words claimed to be understandable by 80% of fourth graders, have also been deemed "difficult." The familiarity score of difficult and easy words was calculated by syllable count as well as using the Dale-Chall list.

2.2 Health-Related, Text Corpora-Based Features

General readability formulas rely on term frequency counts from newspaper articles. (Term frequency is the number of occurrences of a term within a corpus). Because of the scarcity of health-related terms in such general text sources, we obtained three health-related corpora:

1. MEDLINE® abstracts: the National Library of Medicine (NLM) MEDLINE indexes publications from all disciplines in the health domain. While coverage of health-related terms is very broad, MEDLINE's content focuses more on biomedical research than clinical practice. Jargon usage is prevalent.[1]
2. MedlinePlus®: MedlinePlus is a high quality consumer health information Web site developed by the NLM. Because it is tailored for a lay audience, MedlinePlus terminology consists of a mixture of lay terms and jargon.[2]
3. MedlinePlus logs: Log data (i.e., user-submitted queries) are one of the best sources of consumer health language. A limitation is that the authors of consumer-generated text (e.g., newsgroup postings, email messages, or queries) tend to be more motivated and better educated than the general population.

Table 1. Text corpora used by the study

Corpus	Size (no. of words)	Date	Author	Audience
MEDLINE	45,924,958	Jan 1987 - Dec 1991	Professional	Professional
MedlinePlus	3,717,365	Sep. 2003	Professional	Lay
MedlinePlus log	28,797,199	Oct. 2002 - Sep. 2003	Lay	N/A

2.3 Non-Health Related Features

The three health-related corpora do not provide sufficient representation of the health term usage or exposure of lay people with lower literacy levels. Thus, we used the word list from the popular Dale-Chall readability formula as a supplement. The percentage of words in a term that belongs to the Dale list of easy words is treated as a feature.

Another non-health related characteristic is word length: difficult words tend to be longer than easy words. Even though there are many exceptions to the rule, word length is nonetheless a useful feature.

[1] For information on MEDLINE, see http://www.nlm.nih.gov/pubs/factsheets/medline.html. A text corpus of MEDLINE abstracts is available from http://trec.nist.gov/data/t9_filtering.html

[2] For information on MedlinePlus, see http://www.nlm.nih.gov/medlineplus/faq/faq.html

2.4 Familiarity Predication for the Sample Population

A support vector machine (SMV) for familiarity prediction was developed using the following feature variables: term frequency in the three health text corpora, percentage of easy words from the Dale-Chall list, and average word length. The mean familiarity score served as the outcome measure. For evaluation, 10-fold cross validation was performed.

2.5 Individualized Familiarity Predication

While predicating term feasibility for a population is useful, the variation in even a relatively small population can be large. For example, different ethnic groups in US may share characteristics (e.g., language), but have individual differences in other ways (i.e., are heterogeneous) – such as age, education level, gender, profession, and other demographic factors. One approach is to acquire a large and diverse sample and treat demographic variables as features or predictor variables.

Despite the small sample size in this pilot study, the participants came from varying backgrounds, which allowed us to experiment with some demographic variables. Logistic regression model was used for familiarity prediction. The dependent variable familiar is a dichotomous variable coded "1" if the participant answer was right and "0" if wrong or missing. Five term variables (average length, query log frequency, Medline frequency, 4th grader level test, MedlinePlus frequency) and seven demographic variables (gender, native language, race, job, age, ethnicity, and education level) were used.

Table 2. Categorical Variables Coding

Demographic Variable	Categories	Parameter Coding (1)	Parameter Coding (2)
Education	High School	1.000	0.000
	College	0.000	1.000
	Graduate	0.000	0.000
Age	Middle	1.000	0.000
	Young	0.000	1.000
	Older	0.000	0.000
Language	English	1.000	
	Non-English	0.000	
Race	White	1.000	
	Non-White	0.000	
Job	Non-Professional	1.000	
	Professional	0.000	
Ethnicity	Non-Hispanic	1.000	
	Hispanic	0.000	
Gender	Female	1.000	
	Male	0.000	

All term variables are continuous variables. Among the demographic variables, age was converted to categorical data with three levels. We coded age as "Young" when between 20 and 35 years; "Middle," between 36 and 50; and "Older" when over 51. Education levels also had three levels: "High school," "College," and "Graduate." Other variables were all dummy variables. Table 2 describes the categorical variables coding.

3. Results

3.1 Sample data

We recruited 21 participants of varying socioeconomic background (see Table 3). All completed the questionnaire, although some questions were left blank, as we instructed participants not to guess the answers.

Among the 68 terms, 19 consisted of words on the Dale-Chall list and 31 consisted of 1- or 2-syllable words. The remainder were regarded as difficult words. The average familiarity score of the terms is 0.77, with more than half of the terms scoring 0.9 or higher. In other words, at least 50% of the terms were recognized by over 90% of the participants. The familiarity scores of the difficult and easy terms, as judged by the number of syllables and the Dale-Chall list alone, do not appear to be reliable indicators of term familiarity in the test population (see Figure 1).

Table 3. Demographics (n=21)

Variables	Subgroup	Frequency
Gender	Female	5 (23.8%)
	Male	16 (76.2%)
Age	20 ~ 35	9 (42.9%)
	36 ~ 50	5 (23.8%)
	51 over	7 (33.3%)
Education	High School	8 (38.1%)
	College	5 (23.8%)
	Graduate	8 (38.1%)
Native Language	English	16 (76.2%)
	Non-English	5 (23.8%)
Race	White	13 (61.9%)
	Black	4 (19.0%)
	Asian	1 (4.8%)
	American Indian	1 (4.8%)
	Other	2 (9.5%)
Ethnicity	Hispanic	1 (4.8%)
	Non-Hispanic	17 (81%)
	N/A	3 (14.2%)
Occupation	Non-Professional	12 (57%)
	Professional	9 (43%)

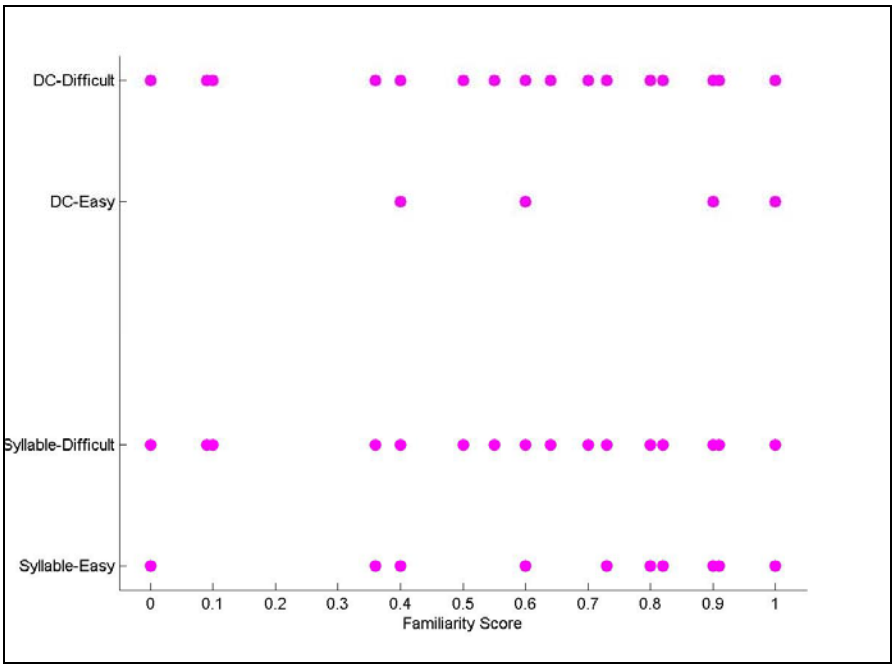

Fig. 1. Distribution of familiarity scores of difficult and easy terms according to the number of syllables and the Dale-Chall easy word list. Some easy terms scored low on the familiarity scale and some difficult terms scored high on the scale.

3.2 Prediction of Average Familiarity

Using a support vector machine (SVM) to predict the average familiarity score of terms based on text corpora resulted in limited success. In 10-fold cross-validation (n=68), the average performance of the SVM is: Mean absolute error=0.196; Root mean squared error=0.293. When reviewing the evaluation results, we found that prediction of terms with the lowest familiarity scores was the least accurate. We suspect that this may be a result of lack of samples with low familiarity scores – only 9 terms scored less than 0.5.

3.3 Individualized Familiarity Predication

A logistic regression model was created for predicting whether a term is familiar to a person with particular background characteristics. As our analysis showed that MEDLINE frequency, MedlinePlus frequency, Gender, Age, and Ethnicity were not significant for predicting familiar terms for this sample, we removed these variables from the model.

Table 4. Logistic regression analysis result

Variable	Coefficient	Standard Error	Significance (p)	Odds Ratio
AvgLength	0.098	0.044	0.028	1.103
QueryLog	22.490	3.162	0.000	5.85E+09
Dale-Chall	1.370	0.353	0.000	3.937
Language(1)	0.790	0.309	0.010	2.204
Race(1)	0.624	0.241	0.010	1.866
Job(1)	-0.647	0.245	0.008	0.523
Education			0.004	
Education(1)	-0.963	0.296	0.001	0.382
Education(2)	-0.735	0.294	0.012	0.479
Constant	-0.413	0.465	0.374	0.662

The final logistic regression model is as follows:

$$\begin{aligned} A &= -0.413 + \text{Log(odds of finding the right answer)} \\ &= -0.413 + (0.098 \times \text{average length}) + (22.490 \times \text{query log}) \\ &\quad + (1.37 \times \text{Dale-Chall}) + (0.790 \times \text{language(1)}) \\ &\quad + (0.624 \times \text{race}) - (0.647 \times \text{job(1)}) \\ &\quad - (0.963 \times \text{education(1)}) - (0.735 \times \text{education(2)}) \end{aligned} \quad (1)$$

$$\text{Probability that the term is familiar to the reader} = \frac{\exp(A)}{1+\exp(A)} \quad (2)$$

Please note that the model and the variables reflect this study's participants' familiarity with a sample of terms. For a different population, the significant variables and the model will be likely to change.

Using 10-fold cross validation (n=714), the regression model performed moderately well:

- Correctly classified instances = 574 (80.4 %)
- Mean absolute error = 0.273
- Root mean squared error = 0.371
- Area under the ROC = 0.796

4. Discussion

This preliminary study showed that predicting text corpora-based term familiarity for health vocabulary is feasible. We measured 21 participants' familiarity with 68 terms through a TOFHLA-style questionnaire. Our attempt to predict average term familiarity from text corpra-based frequencies and other term characteristics resulted in moderate success (10-fold cross validation: mean absolute error = 0.196, root mean squared error = 0.293). Predicting term familiarity from the reader's demographics and the term characteristics generated reasonable results (10-fold cross validation: mean absolute error = 0.273, root mean squared error = 0.371, area under the ROC = 0.796).

Because of the considerable amount of health-related materials available and the large numbers and diverse consumers of these materials, there is a need to measure the readability of health content and make appropriate matches between consumers and content. Use of word length and the Dale-Chall easy word list to identify health terms that are difficult for consumers to comprehend can only provide a very rough estimate of term familiarity, as illustrated by Figure 1. For example, some short words might be incomprehensible and a single list does not reflect the health literacy level of all consumers. In contrast, our approach could provide a more refined and group-specific estimate of familiarity with health-related terms.

Because the number of terms and participants in this preliminary study is small, the predictive models are not likely to be applicable to the general population. In fact, it is not meaningful to estimate familiarity for the general population. Rather, models should be developed for the targeted audience populations. For instance, some public health campaigns may focus on low literacy group or minor groups, while other information might be intended for health-literate readers.

Another limitation is that we extracted term frequencies from the corpora without considering the morphological and lexical variations (e.g. number, tense) of the terms, which influences the calculation of term frequencies. For representation of term usage in media coverage and common English language, text corpora like the Reuters® collections will be better than the Dale-Chall list of easy words.

It may be argued that the underlying relationship between readability and the feature variables cannot be captured by support vector machines or logistic regression. On the other hand, applying some other methods including neural networks to this data set yielded almost identical or worse results – this may not be the case if more sample data are available or different features are used.

Another limitation of the reported approach is that only surface-level familiarity is measured, and not deeper knowledge of the concepts. For instance, although many consumers may recognize the term *heart attack*, few will know its precise definition or risk factors. That a participant answers a multiple choice question containing a term correctly does not indicate full comprehension of the underlying concept. In a related project (www.consumerhealthvocab.org), we manually reviewed concepts and assigned consumer-friendly display names to them. In the manual review process, reviewers not only consider whether a term is recognizable, but also its relationship with existing medical concept(s) found in the NLM Unified Medical Language System® (UMLS®). On the other hand, comprehensive, systematic manual review is lim-

ited by labor costs, while automated methods can be easily applied to large number of terms.

For future work, we would like to extend the questionnaire to include more terms and test their familiarity on a larger, more diverse sample population. It would also be interesting to evaluate the health literacy of participants and explore the relationship between health literacy level and term familiarity.

Research on readability and learning has indicated that providing material of an appropriate level is important to readers of all levels: providing materials that are either too difficult or too easy impairs a reader's ability to absorb new information. Suggesting materials at an appropriate readability level requires differentiating between health terms that consumers are likely to find familiar and unfamiliar (or "difficult") and knowing what a consumer or group is likely to comprehend. Thus, we believe our preliminary study on accurate, user-specific estimations of term familiarity is a necessary step towards improving health communication.

5. Acknowledgement

We thank our collaborators at the Consumer Health Vocabulary Initiative: Guy Divita, Allen Browne, and Laura Roth. This research is funded in part by the NIH grant R01 LM07222.

6. References

1. Ratzan, S.C., and R.M. Parker. (2000). Introduction. In: National Library of Medicine Current Bibliographies in Medicine: Health Literacy. Selden, C.R., Zorn, M., Ratzan, S.C. and R.M. Parker (Eds). NLM Pub No. CBM 2000-1. Bethesda, MD: National Institutes of Health, U.S. Department of Health and Human Services.
2. Rudd, R., B. Moeykens, et al. (2000). Health and Literacy: A Review of Medical and Public Health Literature. Annual Review of Adult Learning and Literacy. J. Comings, B. Garner and C. Smith. San Francisco, CA, Jossey-Bass. 1: 158-199.
3. Osborne, H. (2004). Health Literacy From A To Z : Practical Ways To Communicate Your Health, Jones & Bartlett Pub.
4. (2004). "AHRQ, IOM weigh in on developing a health-literate America." Qual Lett Healthc Lead 16(5): 6-8.
5. McCray, A. T. (2005). "Promoting health literacy." J Am Med Inform Assoc 12(2): 152-63.
6. Davis, T. C., S. W. Long, et al. (1993). "Rapid estimate of adult literacy in medicine: a shortened screening instrument." Fam Med 25(6): 391-5.
7. Parker, R. M., D. W. Baker, et al. (1995). "The test of functional health literacy in adults: a new instrument for measuring patients' literacy skills." J Gen Intern Med 10(10): 537-41.
8. Zakaluk, B. L. and S. J. Samuels (1988). Readability: Its Past, Present, and Future, Intl Reading Assn.
9. Gemoets, D., Rosemblat, G., Tse, T., and R. Logan. (2004). Assessing readability of consumer health information: an exploratory study. Medinfo. 2004: 869-73.
10. Zeng, Q.T., Tse, T., Crowell, J., Divita, G., Roth, R., and Browne, A.C. (2005). Identifying consumer-friendly display (CFD) names for health concepts. Technical Report, DSG-TR-

2005-003. Boston: Decision Systems Group (DSG), Brigham and Women's Hospital, Harvard Medical School.
11. Chall, J. S. and E. Dale (May 1, 1995). Readability Revisited: The New Dale-Chall Readability Formula, Brookline Books.