An Automated Jargon Identification Program: Evaluating STEM Students' Use of Jargon in Written Scientific Genres

Tzipora Rakedzon
Technion – Israel Institute of Technology
hutzipi@tx.technion.ac.il

Elad Segev
HIT – Holon Institute of Technology
elad1segev@gmail.com

Ayelet Baram-Tsabari
Technion – Israel Institute of Technology
ayelet@technion.ac.il

Abstract

Scientists are required to communicate science and research not only to other specialists in the field, but also to specialists from other fields, as well as to the public and policymakers. One fundamental suggestion when communicating with the public is to avoid professional jargon. However, avoiding jargon is difficult for scientists, and unfortunately, there is no standard to aid scientists in adjusting their written messages. In this project, we will present an up-to-date, user-friendly program for identifying jargon in written texts based on a corpus of ~90 million words published in the BBC website during the years 2012-2015. This program was compared and validated with other existing lists from the literature. Following, it was used to test pre and posttest writing samples from 222 students in a 14-week compulsory academic writing course for graduate STEM (science, technology, engineering and mathematics) students. Students' writing samples comprised two contrasting scientific genres: academic writing (abstract) and popular science (press release). The program identified and calculated the level of vocabulary and jargon used in pre and posttests. Results indicate that students use less jargon when writing to a popular audience. However, this was not always enough for the general audience, since the percentage of jargon words in the text still exceeded the recommended amount for adequate comprehension. We conclude that this tool may aid in teaching graduate students and future scientists to recognize and correct their texts to better communicate science to the public.

Keywords: science communication, jargon, academic writing, higher education.

Introduction

"Communication is the essence of science" (Garvey & Griffith, 1979), and communicating science is indeed an integral part of scientists' work. In academia, this is accomplished through writing scientific research articles to share findings with other experts, but recent trends have required scientists to communicate with the public. Scientists, therefore, must learn to write and speak science to the public, both to fulfill public interest (National Science Board, 2016) and aid the public in making educated decisions about the environment and their own health and well-being (Hackling, Goodrum, & Rennie, 2001).

Scientists often find this kind of communication with the public to be difficult: there are many courses for academic writing in universities around the world, but there is very little support for scientists who need to learn how to communicate with the public. Communicating science to the public requires clearer and nontechnical language, as well as a different text structure than academic writing (Baram-Tsabari & Lewenstein, 2013; Fahnestock, 1986; Muñoz, 2015; Rakedzon & Baram-Tsabari, 2016). In fact, vocabulary been implicated in reading comprehension in general as research estimates that readers must be familiar with 98% of words

in a text (Hu & Nation, 2000) to adequately comprehend a text (Schmitt, Jiang, & Grabe, 2011). As such, this means scientists should only use under 1 in 50 unfamiliar words when communicating with a specific audience. Research has shown that scientists do try and use less jargon when communicating with lay audiences but not always "less obscure jargon" (Baram-Tsabari & Lewenstein, 2013; Sharon & Baram-Tsabari, 2013). There are some existing courses which aim to teach scientists to avoid jargon, but this is taught subjectively, as no objective list or tool exists. Subjective perception is an insufficient strategy, as scientists are influenced by their own expertise and the "curse of knowledge".

Few attempts have already been made to contribute to the literature on supporting scientists learning in avoiding jargon. Baram-Tsabari & Lewenstein (2013) used *Google News* hits as a way to judge words on a 5-level scale from familiar to jargon; however, the *Google News* corpus changes its algorithm and updates its corpus regularly, producing constantly changing results for the same jargon at different periods. Sharon & Baram-Tsabari (2013) also used the British National Corpus (BNC) and Professional English Research Consortium (PERC) corpora to classify jargon used in transcripts of regular TED talks, science-related TED talks, and scientists' lectures for the scientific community. This study was limited by the sample size of the talks and lectures and the fact that it represents only spoken English; moreover, it did not produce a tool that is readily usable for training. Therefore, an up-to-date, user-friendly program is needed for aiding scientists and science communication educators to easily identify jargon, which would help in adjusting texts for general audiences. We introduce educational technology that helps scientists in learning and preparing written and spoken science for the public by identifying levels of vocabulary and categorizing jargon.

Objectives

This presentation details the use of an automated jargon identification program aimed at supporting the learning of scientists and the teaching and evaluation of science communication educators. It tests STEM graduate students' academic and popular writing by comparing their use of vocabulary before and after an academic writing course with and without a science communication intervention. Specifically, we ask:

1. Do students adapt their use of jargon to audience in written science genres following an academic writing course?
2. How does a popular science intervention affect the use of jargon in written science communication?

Methods

Design and Development

Over 90 million words were counted using a crawler in all ~250,000 articles published in the BBC website excluding science related channels (e.g. technology, science-environment, health) during the years 2012-2015. Overall, ~500,000 word types were ordered by their frequencies. Word type refers to unique words: for example, *value* and *values* are each unique word types, even though they belong to the same word family. This is similar to the programs used for comparison in the literature (e.g., BNC-COCA). Frequently used words may receive thousands of appearances and jargon may have only few appearances: e.g., *season* received 101,126 appearances, *pollution* 1,596 appearances, and *specifications*, 91 appearances.

The system classified each word into three levels: High frequency ('common,' over 1000 uses in the corpus), mid-frequency ('normal,' over 50 uses in the corpus, e.g. *protein*) and jargon ('rare,' words with less than 50 appearances, e.g. *dendritic*). It then presents the reader with a color-coded text so the reader can easily spot the jargon, and gives the number of words and percentage of the words from each type (see Figure 1).

The system is currently set to analyze the text for lay audiences and as such is set for the top 6500 word types as high frequency, based on literature suggesting that high frequency is

approximately the 2000 most frequent word families in English (Nation (2006) has estimated that 8,000 families are comprised of 34,660 words. Using this as a guideline, we created levels of frequency). The system also allows the option of adjusting levels for different level audiences: the interface allows the user to adjust the number of word types for high, low and jargon frequency (see right side, Figure 1 below) for other level readers. For example, when writing for more advanced readers with a higher-level vocabulary and the expectancy of a higher-level text, one could adjust the high frequency level to include 8000 instead of 6500 word types; for extremely low-level readers, the opposite could be done, creating a 4000-word type level for high frequency. Qualitative tests of the system, as well as suggestions here and in the literature, could produce cutoffs for individual use.

Overall, the system for jargon identification has been in development for 3 years. Five pilot versions and platforms have been tested both to ensure the validity of the lists, categories, and online interface.

Procedure

Sample. Following a three-stage validation (Rakedzon, Segev, & Baram-Tsabari, in preparation), the program was tested on pre and posttest academic and popular writing samples from 222 students from two consecutive semesters in an academic writing course for graduate STEM students in a leading Technological university. Pre/post tasks were administered at the beginning and repeated at the end of the 14-week semester to students participating in the course, mostly PhD students. Students were mostly native speakers of Hebrew (~70%), followed by Russian (13%), Arabic (5%) and other (12%) (speakers of other European or Asian languages or those who listed more than one native language). All students were required to take an internal English proficiency test for academic reading comprehension and grammar at the graduate level to be accepted to the course. The passing requirement was a score equivalent to B2 on the Common European Framework (Council of Europe, 2001).

Performance task. The pre and post task both asked: “Please describe your research, its context and implications for (A) a general audience (no science background) and (B) the academic community, in 150-250 words each”. The expectation of the tasks included writing in contrasting genres, i.e. an academic paper abstract to assess academic writing and a press release to assess popular science writing.

Experimental setup. We used a quasi-experimental design to examine whether a lesson about science communication in an academic writing course can improve graduate students’ popular science writing skills, and specifically, use of jargon in communicating research to the public. Three groups took part in the same academic writing lectures with the same teacher with one lesson dedicated to popular science writing (Table 1). In this lesson, the lecture included background on science communication, statistics about its usage, types (magazines, newspapers, books, press releases), and online example texts for students to emulate. Table 1 shows the intervention groups, and their corresponding conditions. T-tests were used to compare pre-post change in jargon use. To control for length, scores were divided by word number.

Results

Overall, graduate students used more jargon, which is appropriate for academic writing. In contrast, jargon was a confusing issue for students when writing for the public. On one hand, a majority of the students on both the pre and posttest did clearly use less jargon in the popular as opposed academic part, with approximately 10% of text using jargon for academic and 6% for popular science. This unfortunately does not reach the ideal 2% of unfamiliar words in a text recommended by the literature. The results in Table 2 are presented per intervention.

Table 1. Intervention groups and experimental setup

Intervention Group	Academic writing course	Pre task	Feedback on pre task	Popular science lecture	Press release task with feedback	Formal letter task with feedback	Post task
PSLpress (n = 94)	V	V	V	V	V	-	V
PSLletter (n = 81)	V	V	V	V*	-	V	V
Comparison (n = 30)	V	V	V	-**	-	V	V
Control (n = 17)	-	V	-	-	-	-	V

Notes:

V = completed stages; - = stages not included for this group; PSLpress = popular science lecture-press release task group; PSLletter= the popular science lecture and formal letter task group; comparison is the regular academic writing course without intervention; control includes students without an academic writing course or intervention;

* these groups received a lecture on popular science and press releases; they also received a shorter version of the lecture on formal letters, CVs and email writing to keep the number of lectures for all conditions equal.

** This group received a full lecture on formal letters, CVs and email writing.

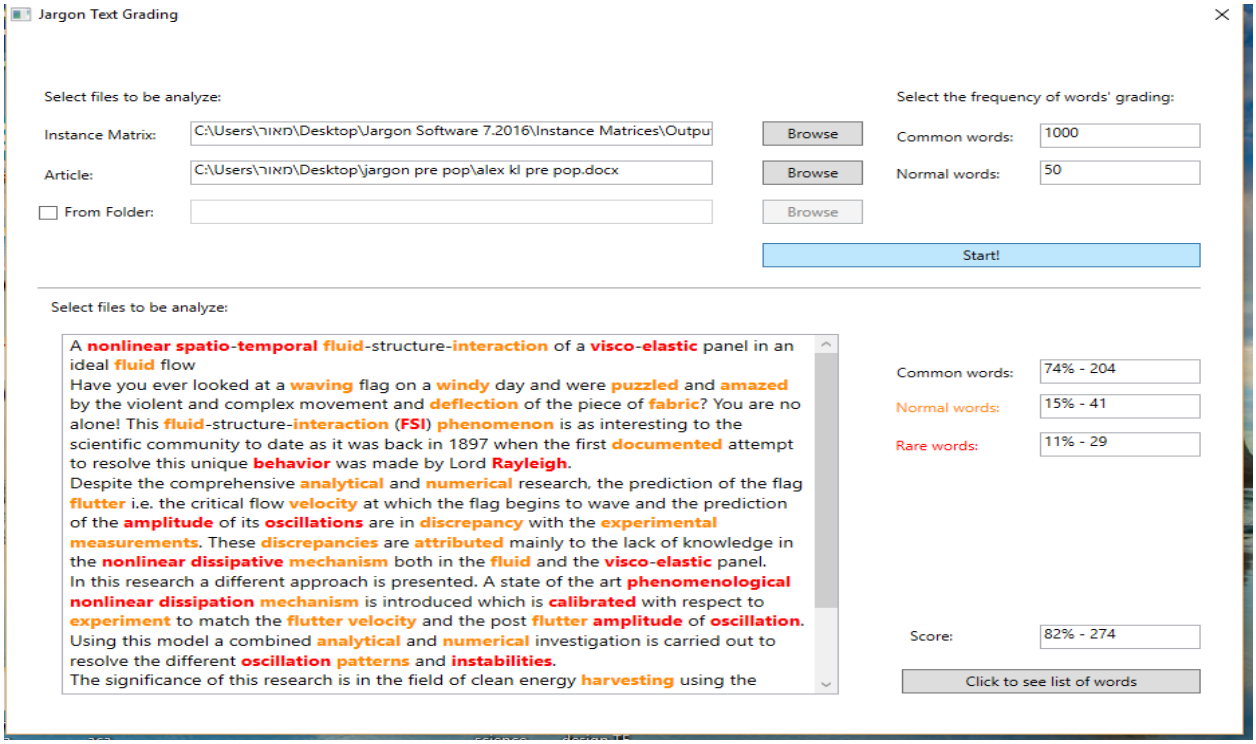

Figure 1. Example color-coded text from jargon identifier

Table 2. Pre-post change in vocabulary use

Group	Text types	Vocabulary types	Mean pre	Mean post	Mean dif.	Std. error mean	Sig.
Comparison	Pre-post Academic	High freq	.7500	.737667	–.0123333	.0098224	.219
		Mid-freq	.131000	.143000	.0120000	.0080715	.148
		Jargon aca	.119667	.120333	.0006667	.0060635	.913
	Pre-post Popular science	High freq	.8090	.7963	–.01267	.01081	.251
		Mid-freq	.116667	.126333	.0096667	.0063875	.141
		Jargon	.073333	.077667	.0043333	.0072822	.556
	Posttest: academic v. popular science	High freq	.737667	.7963	.0586667	.0095861	.000*
		Mid-freq	.143000	.126333	–.0166667	.0057602	.007*
		Jargon	.120333	.077667	–.0426667	.0079211	.000*
PSLpress	Pre-post Academic	High freq	.7654	.741170	–.0242553	.0052516	.000*
		Mid-freq	.136809	.149362	.0125532	.0042913	.004*
		Jargon	.097340	.109362	.0120213	.0040320	.004*
	Pre-post Popular science	High freq	.8256	.8000	–.02564	.00610	.000*
		Mid-freq	.120426	.136702	.0162766	.0045927	.001*
		Jargon	.054362	.062340	.0079787	.0042930	.066
	Posttest: academic v. popular science	High freq	.741170	.8000	.0588298	.0061512	.000*
		Mid-freq	.149362	.136702	–.0126596	.0041039	.003*
		Jargon	.109362	.062340	–.0470213	.0047539	.000*
PSLletter	Pre-post Academic	High freq	.7765	.749506	–.0270370	.0059743	.000*
		Mid-freq	.138025	.154074	.0160494	.0050267	.002*
		Jargon	.085185	.096173	.0109877	.0037417	.004*
	Pre-post Popular science	High freq	.8217	.8080	–.01370	.00676	.046*
		Mid-freq	.122840	.134815	.0119753	.0046195	.011*
		Jargon	.055679	.056420	.0007407	.0039471	.852
	Posttest: academic v. popular science	High freq	.749506	.8080	.0585185	.0058613	.000*
		Mid-freq	.154074	.134815	–.0192593	.0041825	.000*
		Jargon	.096173	.056420	–.0397531	.0039479	.000*
Control	Pre-post Academic	High freq	.7629	.768235	.0052941	.0099306	.601
		Mid-freq	.127647	.120588	–.0070588	.0077062	.373
		Jargon	.110000	.109412	–.0005882	.0080198	.942
	Pre-post Popular science	High freq	.8194	.8200	.00059	.00797	.942
		Mid-freq	.120000	.109412	–.0105882	.0068315	.141
		Jargon	.060588	.069412	.0088235	.0080386	.289
	Posttest: academic v. popular science	High freq	.768235	.8200	.0517647	.0100086	.000*
		Mid-freq	.120588	.109412	–.0111765	.0072194	.141
		Jargon	.109412	.069412	–.0400000	.0121268	.005*

*$p < 0.05$.

Analysis of both intervention groups, PSLpress and PSLletter (Table 2), showed similar results. For academic writing, we can see that high frequency words showed a significant trend of decreasing in PSLpress and PSLletter, $t(93) = –4.619$ $t(80) = –4.526$, respectively. This was

evident in the significant expected gain in mid-frequency words and jargon use, $t(93) = 2.981$; $t(80) = 2.937$, over the course of a semester.

For popular science writing, we also see a significant lower percentage of high frequency words [PSLpress $t(93) = –4.203$; PSLletter $t(80) = –2.029$], replaced by mid frequency words. In terms of jargon, we do find a significant trend.

One interesting difference arises when we evaluate the academic posttests and compare it to the popular science posttest. Here we see that the popular science posttest indeed has a significantly lower percentage of jargon use [PSLpress $t(93) = –9.891$; PSLletter $t(80) = 10.070$], indicating some overall awareness of adaptation for lay audiences.

For the comparison group (the group who had the course without the popular science intervention), results do not show any significant change in jargon use for popular science, but they too used more jargon in popular science tests than academic tests. The control group essentially showed deterioration since they did not participate in any course or intervention: there was a nonsignificant trend of less jargon for the academic post and more jargon for the popular science post. However, they did know, similar to all groups, to use less jargon on popular science posttest than the academic post.

Discussion and conclusions

Overall, several trends indicate that STEM graduate students are aware that they need to use less jargon for a general audience in comparison with an academic audience. However, the numbers still show that a significant difference between genres does not necessarily produce a text for a general audience. A text comprising ~6% jargon is likely not enough for the general audience, considering readers need to be familiar with 98% of words in a text for good comprehension (Hu & Nation, 2000).

These results must be considered in light of several limitations. First, language is dynamic, and new words are constantly entering the language. Some words become more common with time, such as names of diseases, demanding frequent updates of the corpus, which we intend to do. Secondly, assessing academic as opposed to popular science writing reaches beyond word choice and jargon, including differences in sentence structure, style, and syntax. This can be measured using other tools such as manual rubrics (Rakedzon & Baram-Tsabari, 2016).

Beyond this specific intervention, this tool provides an up to date, accurate and user-friendly tool to analyze scientific texts for a variety of audiences and levels. This tool is aimed to aid scientists when writing for the public to identify the jargon which would complicate their text for the lay audience. This would enable them, for example, to find which terms in the text need to be replaced or explained. This also may aid those training for science communication. This may include scientists, or even graduate students and future scientists, by teaching them to recognize and correct their texts for communicating with the public as part of their enculturation during their graduate studies. Moreover, it can be used in assessment of science communication training programs. Finally, this will also aid researchers in testing many types of written, and possibly spoken, science communication at several levels and in different disciplines for various audiences. We hope the tool will also be expanded with time to include a corpus that updates with time, and expanded to include use for other languages as well.

References

Baram-Tsabari, A., & Lewenstein, B. V. (2013). An Instrument for Assessing Scientists' Written Skills in Public Communication of Science. *Science Communication*, *35*(1), 56-85.

Council of Europe. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Fahnestock, J. (1986). Accommodating Science: The Rhetorical Life of Scientific Facts. *Written Communication*, *3*(3), 275-296.

Garvey, W. D., & Griffith, B. C. (1979). Scientific communication as a social system. *Communication: The Essence of Science*, 148.

Hackling, M., Goodrum, D., & Rennie, L. (2001). The state of science in Australian secondary schools. *ECU Publications Pre. 2011*. Retrieved from http://ro.ecu.edu.au/ecuworks/4682

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language*, *23*, 403-430.

Muñoz, V. L. (2015). The vocabulary of agriculture semi-popularization articles in English: A corpus-based study. *English for Specific Purposes*, *39*, 26-44.

Nation, I. (2006). How Large a Vocabulary is Needed For Reading and Listening? *Canadian Modern Language Review*, *63*(1), 59-82.

National Science Board. (2016). *Science And Engineering Indicators*.

Rakedzon, T., & Baram-Tsabari, A. (2016). Assessing and improving L2 graduate students' popular science and academic writing in an academic writing course. *Educational Psychology*. Retrieved from http://dx.doi.org/10.1080/01443410.2016.1192108

Rakedzon, T., Segev, E., & Baram-Tsabari, A. (2016). An automatic jargon identifier for scientists engaging with the public and for science communication educators. *Manuscript in Preparation*.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. *The Modern Language Journal*, *95*(1), 26-43.

Sharon, A. J., & Baram-Tsabari, A. (2013). Measuring mumbo jumbo: A preliminary quantification of the use of jargon in science communication. *Public Understanding of Science (Bristol, England)*, *23*(5), 528-546.