

Supporting scientists in engaging with the public: The development and validation of an automated jargon identification program, the De-jargonizer

Tzipora Rakedzon¹, Elad Segev², Ayelet Baram-Tsabari¹, Noam Chapnik², Roy Yosef² ¹Technion-Israel Institute of Technology; ²Holon Institute of Technology

Introduction

Scientists often find communicating science to the public to be difficult, as it requires clear and nontechnical language, as well as a different text structure than academic writing. Previous research has attempted to identify jargon (Baram-Tsabari & Lewenstein, 2013; Sharon & Baram-Tsabari, 2013); however, no user-friendly tools exist. The De-jargonizer is an up to date, accurate and user-friendly tool to analyze scientific texts for a variety of audiences and levels.

Objectives and research questions

This project details the development and validation process of an automated jargon identification program.

Specifically, we ask:

1. Is the new jargon measurement comparable to other existing vocabulary systems, and what advantages does it provide?

2. Is jargon used at appropriate levels for popular science and academic writing among early career scientists?

Design and procedure

A crawler counted over 90 million words in all ~250,000 articles published in the BBC sites (including science related channels) during the years 2012-2015. Overall, ~600,000 word types were ordered by number of appearances. The system analyzes a text into three frequencies:

Word	Appearances	Examples
classification	in corpus	
High	Over 1000	Pressure, current
frequency		
Mid frequency	Over 50	Genetic, protein
Jargon	Less than 50	Phylogenetics, ions

The site then presents the reader with the number of words from each frequency and their percentage of the total words in the text. It also displays a color coded text (see figure below), allowing the reader to easily spot the jargon.



Validation process. RQ1: Three procedures were used to test the accuracy and validity of program.

Compleat Web VP Comparing scores on students' tests in an academic writing course (Rakedzon & Baram-Tsabari, 2017) using the De-jargonizer and the BNC-COCA VPcompleat program (Nation, 2012; Cobb, 2016) found a significant correlation at the 0.01 level (2-tailed).

Comparing academic abstracts to general audience summaries published in PLOS journal articles (5000 pairs): t-tests for change in vocabulary use show a significant effect of more frequent use of jargon in academic abstracts.

TED Comparison of non-science TED lectures, science TED lectures and academic scientific lectures: similar results were found by Sharon & Baram-Tsabari, 2013 and the De-jargonizer; i.e. significantly higher jargon use between non-science TED lectures and academic scientific lectures.

Technion Comparison of jargon use in pre and posttests by STEM graduate students at the Technion before and after an academic writing course with and without a science communication intervention.

hutzipi@technion.ac.il

The De-jargonizer site

zer	Multiple Text Grading	About	Instructions	Examp	oles
	Contact Us				
hea	rt is a proximal cause for heart failure. Not	wadays treatme	nt for rhythmic	Common	0.4
onic chan erate	pacemakers. Although an excellent solution ge, and the risk of contamination. Thus, a d from the person's own hair, and can diffe	Mid- Frequency:	13		
ty of ur ma	these heart-like cells to human physiology ajor goal is to characterize their electrical b	Rare:	3%		
ate adult human heart cells. Although much research is still due, this novel <mark>biological</mark> ic hopes.			Score:	90	

RQ2: Early career scientists' use of jargon





Results

Jargon use per condition for academic and popular science genres post control pre control (n=34) post comparison pre comparison (n= 174) post intervention pre intervention (n=17) post control pre control 4%, 99 (n=34) post comparsion 3%, 15 pre comparison (n= 174) post intervention %, 4 pre intervention 4 6 8 10 12 % of jargon





Jargon use in students' writing: Results were similar for all groups. Findings show jargon made up ~10% of academic texts, and ~6% of popular science texts - beyond the recommended 2% (Nation & Hu, 2000).

Discussion

The jargon identifier aims to support communicating science to a lay audience by identifying jargon. The jargon identifier:

- accurately measures jargon level in texts, finding good agreement with results from other studies on jargon (Sharon & Baram-Tsabari, 2013) and word frequency (Nation, 2012)
- can be used in assessment of science communication training programs and aid researchers in testing written, and possibly spoken, science communication messages