Measuring Mumbo-Jumbo: A Preliminary Quantification of the Use of Jargon in Science Communication

Aviv J. Sharon and Ayelet Baram-Tsabari

Technion – Israel Institute of Technology, Israel

aviv.sharon@gmail.com, ayelet@technion.ac.il

Abstract

Science communication training programs, increasingly found worldwide, aim to teach scientists skills for communicating about science with the public effectively. These include discussing science with less professional jargon. However, little attention has been paid to developing assessment methods for the outcomes of such programs, or for assessing the use of jargon.

Here we propose a novel approach to assessing scientific jargon in spoken English texts, using software and corpora (collections of texts) all freely available online. We compare the use of scientific jargon in existing science communication with its use among scientists. This study aims to develop a standardized "jargon index", which may be useful in formative or summative assessment.

Analyzed transcripts included academic speech and scientific TEDTalks. Within each transcript, words were annotated based on usage in (1) a scientific English corpus, and in (2) a general English corpus. Words appearing either only, or significantly more frequently, in the scientific corpus, were categorized as scientific jargon. These were also assigned "jargonness" scores, which reflect their obscurity.

Findings suggest that scientists use less jargon in communication with a general audience than with peers, but not always less obscure jargon. These findings may lay the groundwork for (self-)evaluating jargon through technology.

Keywords: Jargon, media training, scientists' understanding of the public

Introduction

To garner support and legitimacy for scientific endeavors, the scientific community has increasingly recognized the importance of communicating science to non-technical publics (hereafter "science communication") (Nisbet & Scheufele, 2009). Accordingly, a worldwide educational endeavor aims to teach scientists relevant communication skills. One of these skills is expressing ideas in one's domain of expertise while avoiding scientific jargon as much as possible (e.g., Baron, 2010; Dean, 2009; Meredith, 2010).

However, little attention has been paid to developing consistent methods to evaluate the outcomes of science communication training programs in general, or the use of jargon among trainees in particular (Baram-Tsabari & Lewenstein, 2012). This exploratory study strives to quantitatively assess the use of scientific jargon in science communication to develop a standardized, evidence-based "jargon index" based on some of the best practices in this field.

Literature Review

Science Communication Training and Evaluation

Although many science communication training programs exist worldwide, little attention has been paid to defining the goals learners should aim for in such programs, and how attainment of these goals should be evaluated. Such programs are typically evaluated using methods focusing on the learners' attitudes and reflections, rather than on their communication skills (Miller, Fahy, & The ESConet Team, 2009; Mulder, Longnecker, &

Davis, 2008). One conceptual framework outlines several measurable components of skills a scientist should have to communicate effectively (Baram-Tsabari & Lewenstein, 2012). Specifically, to effectively engage with the public, scientists are advised to convey meaningful scientific ideas without scientific jargon (e.g., Dean, 2009; Meredith, 2010). This change in speech patterns is important for effective communication of science both to ensure clarity and to promote positive views of science and scientists.

However, experts use jargon excessively for several reasons, such as lack of motivation and lack of skill. These may be reinforced by social norms, which are acquired when learning science and becoming enculturated into an academic community (Lemke, 1990) as well as by a cognitive bias called the "Curse of Knowledge". This bias causes individuals to overestimate what another person knows, because their judgment is impaired by their own knowledge. Evidence suggests this bias also causes individuals to overestimate public familiarity with technical terms (Hayes & Bajzek, 2008).

Thus overall, avoiding jargon for clarity's sake requires a conscious and deliberate effort to communicate clearly, which is an acquired skill demanding knowledge and experience (Stableford & Mettger, 2007). We argue here that this skill should be both explicitly taught in science communication training and rigorously assessed.

Scientific Jargon: Definition and Usefulness

A large body of research indicates that when people use language in different contexts, they make different choices of pronunciation, morphology, vocabulary, grammar and discourse features. In turn, this gives rise to different sets of human speech patterns, including *registers* of language, which are speech patterns "associated with situational contexts or purposes", such as legalese, the English of football commentaries, etc. (Biber, 1995, p. 1). Registers can be broadly or narrowly defined based on many variables, such as the roles and characteristics of the participants, the topic and purpose of the communication event and more (Biber, 1988). Thus, one can speak of a *scientific* register of English, used primarily by scientists when communicating about science with their colleagues and students. This work will specifically focus on the specialized vocabulary of the scientific register, or *scientific jargon*. We coin the neologism *"jargonness"* to refer to the degree to which the use of a word is *restricted* to the scientific register, i.e. rarely found outside it.

Evaluating Clarity in Science Communication

There exist several approaches to assess the understandability of any text, and in particular, to evaluate the clarity of a scientific text. One approach uses readability formulas, and another analyzes the vocabulary used based on short word lists (Ley & Florio, 1996). However, these approaches suffer from several drawbacks, mainly their limit of scope and flexibility. To capture a wide range of scientific words in a text, we opted to base our analysis on larger samples of the language. This method is called *corpus-based* linguistic analysis.

A *corpus* is a large collection of natural, authentic texts, written or spoken, which represent a language or language variety in machine-readable form, which may be annotated with various forms of linguistic information (McEnery, Xiao, & Tono, 2006).

Hypotheses

1. Jargon is *less pervasive* in popular science communication than within communication among scientists.
2. Effective science communication uses *less obscure jargon* than communication among scientists.

Methodology

Data Sources

Science Communication. As authentic examples of science communication, we used transcripts of (1) science-related "TEDTalks" (discussed here), and (2) a press conference about the discovery of a Higgs-like boson at CERN (see "External Validity").

TEDTalks are brief lectures featured in the TED conferences. TED (originally "Technology, Entertainment, Design") is a nonprofit organization that holds two annual conferences in California and Scotland on topics such as entertainment and design but also economics, science, and education. Online videos of TEDTalks have accumulated over half a billion views in total to date (Kessler, 2011). Given their popularity and their heavy emphasis on science, we drew on TEDTalks to analyze the best practices in using jargon in science communication.

Specifically, we retrieved all transcripts of TEDTalks tagged as "science" from 2010 and 2011 ("TED Science", 31 transcripts, 69,290 words in total, 2,235 words per transcript on average).

Communication among Scientists. To retrieve authentic examples of how scientists communicate with each other, we used (1) scientific transcripts from the Michigan Corpus of Academic Spoken English (MICASE; discussed here) and (2) scientific seminars about the discovery of a Higgs-like boson at CERN (see "External Validity"). MICASE is a corpus of transcripts totaling approximately 1.7 million words, collected and transcribed from recordings of lectures, classroom discussions and more at the University of Michigan (Simpson, Briggs, Ovens, & Swales, 2002). All transcripts categorized under "Physical Sciences and Engineering" and "Biological and Health Sciences" were included in the sample, except for those with titles containing the word "intro" or ending with the words "lab" or "study group". These were omitted in order to focus on scientific communication at an advanced undergraduate level and above ("MICASE", 43 transcripts, 487,671 words in total, 11,341 words per transcript on average).

Control. As a control group, we retrieved all transcripts of TEDTalks from the same years as the Science Communication group, as long as they were tagged as "design" but not also as "science"[1] ("TED Design", 28 transcripts, 53,780 words in total, 1,921 words per transcript on average).

Reference Corpora

General corpus. As a representative corpus of the English language, we used the British National Corpus (BNC), hereafter the "general corpus." This corpus contains ~100 million orthographic words, and was designed to represent a wide range of British English, as it was used between 1960 and 1993. It is publicly searchable via web interfaces such as BNCweb[2] (Hoffman & Evert, 2006).

Scientific corpus. To represent the scientific register of English, the Professional English Research Consortium (PERC) Corpus was used, hereafter the "scientific corpus". This corpus is a ~17-million-word corpus of English academic journal texts from high-impact science and technology journals. It is also publicly searchable via a web interface[3].

Isolating Uncommon Words

The more frequently a word occurs in the language as a whole, the higher the percentage of people who understand that word (Ley & Florio, 1996). Hence, we assumed that words of scientific jargon that impede clarity in science communication are relatively *rare* words. To isolate the uncommon words from our samples, we drew on existing lists of *common* words and excluded words on those lists from our sample.

Specifically, we omitted words belonging to the 9,000 most common word families in the English language (BNC Word Family Lists 1-9 from Heatley and Nation (1994); See Fig. 1, Step 1). Here, a word family is a set of morphologically related words, such as the root form "care" and its derived forms *cared, carer, carers, careful, carefully, careless, carelessness, cares, caring, carelessly, uncared* and *uncaring*. The number 9,000 was chosen because previous work has shown that 8,000 to 9,000 word families are needed to adequately comprehend written texts in English, such as newspapers, movie transcripts and novels without assistance (Nation, 2006).

The elimination of common words was done by using AntWordProfiler, a freeware software package that classifies words of groups of texts based on word lists, and can isolate words that belong to no list (Anthony, 2009). The program also generates statistics about the "tokens" (occurrences of words) and "types" (classes of words) in the texts, and these are presented in this study. Using the type-token distinction, the sentence "A rose is a rose is a rose" has eight tokens but only three types ("A", "rose" and "is").

Texts were analyzed based on BNC wordlists 1-9, included with the Range software package (Heatley & Nation, 1994). This left us with a set of relatively uncommon word types extracted from each type of transcripts.

Analysis of Uncommon Words

Each uncommon word type from our samples was evaluated in terms of its "jargonness" – the degree to which its use is *restricted* to the scientific variety, i.e. the degree of the word's obscurity to non-technical publics (Fig. 1, Step 2). To quantify this, two queries were conducted for each word type: (1) Its frequency per million words in the *general corpus* and (2) Its frequency per million words in the *scientific corpus*. To automate word frequency retrieval, a custom-made Python script was employed (Halwany, 2011).

Next, each word type was classified into one of five categories based on its relative frequencies in the two corpora (Fig 1., Step 2): (A) Words appearing exclusively in the scientific corpus, and not in the general corpus, e.g., "metalloproteases"; (B) Words appearing in both corpora, but with a higher normalized frequency in the scientific corpus, e.g., "thermodynamic"; (C) Words appearing in both corpora, but with a higher normalized frequency in the general corpus, e.g., "honeycombs"; (D) Words appearing exclusively in the general corpus, and not in the scientific corpus, e.g., "foolhardy"; (E) Words appearing in neither corpus, e.g., "kindergarteners" but also "neurofibroma".

Words from category B were further subdivided by the statistical significance of their specificity. Significance was determined by calculating the log-likelihood statistic for the frequencies of each word in the two corpora. Only uncommon words appearing *exclusively* in the scientific corpus (Fig 1., Category A), or appearing *significantly more frequently* in the scientific corpus than in the general corpus (Fig 1., Category B1, $p < 0.05$), were classified as scientific jargon.

Next, words appearing exclusively in the scientific corpus, or significantly more frequently in the scientific corpus than in the general corpus (Categories A or B1) were assigned jargonness scores.

Jargonness for each word was determined differently, depending on its presence in the general corpus: (1) If it appeared at least once in the general corpus (Category B1), jargonness was the common logarithm of the ratio of its normalized (i.e., per-million) frequencies in the scientific and general corpora.

The common logarithm was extracted from the frequency ratio because the same word may be found in different corpora, but with normalized frequencies that differ by several orders of magnitude. By extracting the common (base-10) logarithm of the quotient of frequencies, one easily notices the order of magnitude of this quotient.

(2) If a word existed *only* in the scientific corpus, and *not* in the general one (Category A), its jargonness was set at three, slightly below the maximal jargonness value found in this study (see "Results"). This means we made the conservative assumption that the word is three orders of magnitude (i.e., 1,000 times) more frequent in the scientific corpus than in the general one.

The following formula summarizes this calculation:

$$\text{Jargonness} = \begin{cases} \log_{10}\left(\dfrac{frequency_{scientific}}{frequency_{general}}\right) & (frequency_{general} > 0) \quad \text{Category B1} \\ 3 & (frequency_{general} = 0) \quad \text{Category A} \end{cases}$$

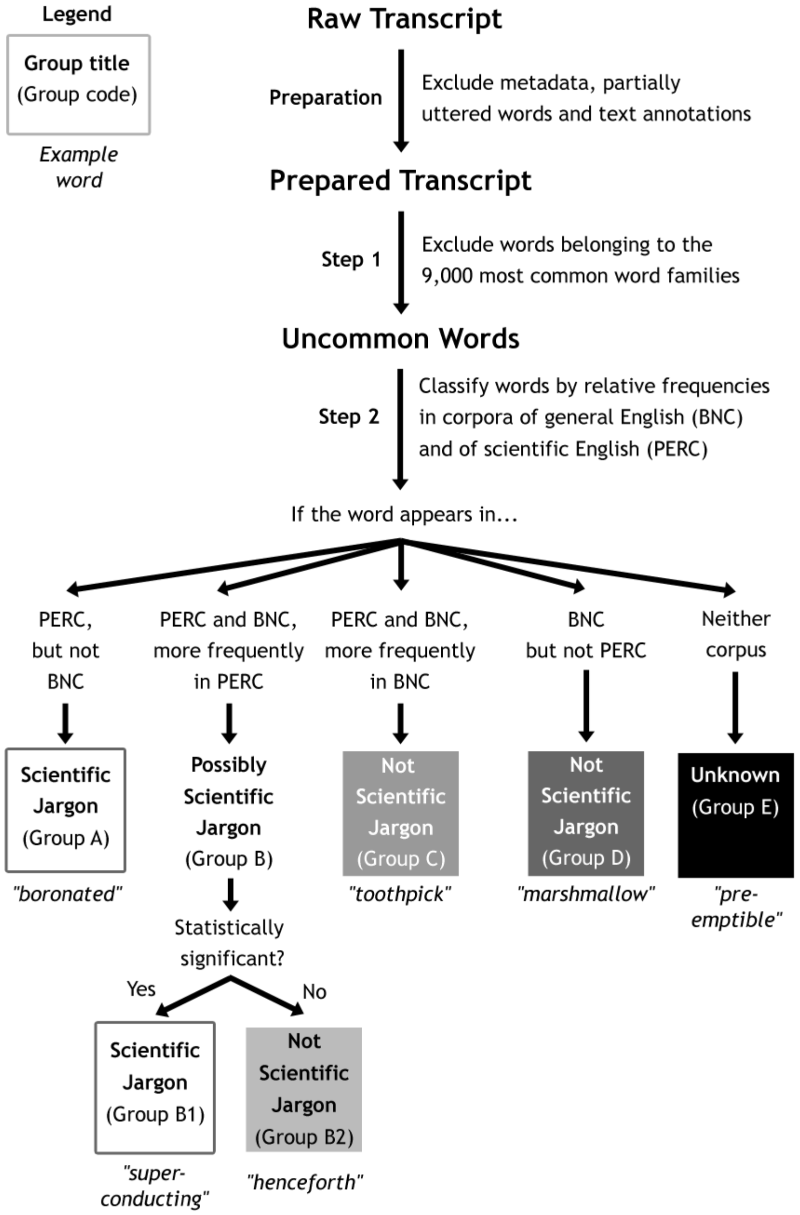
Legend
Group title
(Group code)
Example word
Raw Transcript
Preparation
Exclude metadata, partially uttered words and text annotations
Prepared Transcript
Step 1
Exclude words belonging to the 9,000 most common word families
Uncommon Words
Step 2
Classify words by relative frequencies in corpora of general English (BNC) and of scientific English (PERC)
If the word appears in...
PERC, but not BNC
PERC and BNC, more frequently in PERC
PERC and BNC, more frequently in BNC
BNC but not PERC
Neither corpus
Scientific Jargon (Group A)
"boronated"
Possibly Scientific Jargon (Group B)
Not Scientific Jargon (Group C)
"toothpick"
Not Scientific Jargon (Group D)
"marshmallow"
Unknown (Group E)
"pre-emptible"
Statistically significant?
Yes
No
Scientific Jargon (Group B1)
"super-conducting"
Not Scientific Jargon (Group B2)
"henceforth"

Figure 1. Methodology for classification of words in a spoken text to jargon and non-jargon.

Results

*Identification of Uncommon Words***.** Counting both in tokens and in types, scientific academic speech (MICASE) had a larger proportion of uncommon *types* and uncommon *tokens* than science communication (TED Science) and control transcripts (TED Design) (Table 1) (Two independent 3-sample proportion tests, p < 0.001 in each). In other words, there was a difference in the prevalence of rare words (not necessarily jargon) between the academic scientific speech, science communication and control transcripts.

	MICASE (Communication among scientists)	**TED Science (Science communication)**	**TED Design (Control)**
Initial word count (tokens)	487,671	69,290	53,780
Uncommon words (tokens) **(Absolute number and % of initial word tokens)**	12,909 (2.65%)	1,439 (2.08%)	995 (1.85%)
Initial word count (types)	14,088	6,578	5,936
Uncommon words (types) **(Absolute number and % of initial word types)**	3,636[(a)] (25.81%)	841 (12.79%)	663 (11.17%)

Table 1. Proportions of uncommon words in three collections of transcripts (Step 1). "Tokens" are occurrences of words and "types" are classes of words. [(a)] 65 two-letter words were omitted from this group to reduce the number of partially transcribed words in the sample.

*Proportion of Jargon within Uncommon Words***.** The proportions of scientific jargon within uncommon types varied significantly between the groups of texts. Scientific jargon made up 43.3% of the uncommon types in MICASE, compared to 37.5% of uncommon TED Science types and 19.2% of uncommon types in TED design (Table 2; 3-sample proportion test, p < 0.001). Thus scientific jargon was more prevalent in academic scientific speech than in science communication by a factor of 1.15 (2-sample proportion test, p < 0.01).

Category	Sub-Category	Uncommon types found in…	Scientific Jargon?	Examples	MICASE (Comm. among scientists, n = 3571)	TED Science (Science comm., n = 841)	TED Design (Control, n = 663)
A	–	**The scientific(a) corpus but not in the general one(b)**	Yes	"allergenicity" "postsynaptically"	184 (5.15%)	30 (3.57%)	15 (2.26%)
B	–	**Both scientific and general corpora, and more frequently in the scientific corpus (total)**			1640 (45.93%)	347 (41.26%)	155 (23.38%)
	B1	Of which *significantly*(c) more frequent in the scientific corpus	Yes	"ethanol" "photoreceptor"	1362 (38.14%)	285 (33.89%)	112 (16.89%)
	B2	Of which *not significantly* more frequent in the scientific corpus	No	"hallucinogen," "prerecorded"	278 (7.78%)	62 (7.37%)	43 (6.49%)
C	–	**Both scientific and general corpora, and more frequently in the general corpus**	No	"hyperactive" "decaffeinated"	419 (11.73%)	139 (16.53%)	145 (21.87%)
D	–	**The general corpus, but not in the scientific one**	No	"brunch" "choreography"	667 (18.68%)	208 (24.73%)	245 (36.95%)
E	–	**Neither the general nor the scientific corpus**	–	"essentialistic" "velociraptor"	661 (18.51%)	117 (13.91%)	103 (15.54%)
		Total scientific jargon(d)	**Yes**		**1,546 (43.3%)**	**315 (37.5%)**	**127 (19.2%)**
		Total not scientific jargon(e)	**No**		**1,364 (38.20%)**	**409 (48.63%)**	**433 (65.31%)**

Table 2. Distribution of uncommon types in two corpora of scientific and general English (Step 2). Percentages are calculated of total uncommon types extracted from each data source.

(a) PERC (Professional English Research Consortium) Corpus; (b) BNC (British National Corpus); (c) Log-likelihood, $p < 0.05$; (d) Sum of types in categories A & B1; (e) Sum of types in categories B2, C & D.

*Jargonness***.** Next, the level of jargonness of the scientific jargon was examined across the three groups of texts. Jargon types in academic speech (MICASE) were more obscure than jargon in science communication. In fact, the median MICASE jargon word had a jargonness value of 1.21, and thus was significantly greater than the median in TED Science, which was 1.078 (Wilcoxon-Mann-Whitney (WMW) Test. $p < 0.001$). TED Science jargon did not have significantly different jargonness than the jargon from the control group, TED Design, whose median jargonness value was 1.022 (WMW Test, not significant).

External Validity. The method was re-applied to compare the prevalence and obscurity of jargon in scientific seminars about the discovery of a Higgs-like particle ($n_{\text{seminars, tokens}}$ = 1,572; $n_{\text{seminars, types}}$ = 473), versus statements in the press conference on the same topic by the same two spokespeople at CERN ($n_{\text{press conf., tokens}}$ = 1,645; $n_{\text{press conf., types}}$ = 501). The scientific seminars contained a higher proportion of uncommon types than the press conference (5.92% vs. 2.59%; 2-sample proportion test, $p < 0.01$). In both cases, most of these uncommon types were scientific jargon: 23 of the 28 uncommon types in the seminars (82%), and 10 of the 13 uncommon types in the press conference (77%). Overall, the scientific seminars contained a higher proportion of jargon types than the press conference by a factor of 2.4.

The median jargonness of jargon types, however, was greater in the press conference (1.65) than in the seminars (1.33; WMW Test: $p < 0.05$). Thus when discussing the discovery of a Higgs-like boson, the spokespeople used over twice as much scientific jargon when addressing the scientific community as when addressing the public, but the jargon used when addressing the public was more obscure.

Comparison of Cases. Among the uncommon words in each group, academic scientific speech contained significantly more jargon than science communication in both cases examined. This confirms the first hypothesis. As for the jargonness (obscurity) of the scientific jargon, the data present a more nuanced picture. In one case (MICASE vs. TED Science) the jargon used in science communication had lower jargonness than the jargon in speech among scientists, but in another (Higgs boson seminar vs. Higgs boson press conference) the reverse was true. In other words, in both comparisons, *less* jargon was used in science communication than in academic speech, but only in one comparison was the jargon used when addressing the public *less obscure* than the jargon in academic speech, as hypothesized.

The observed shift in lexical choice might partly be explained as a result of speakers tailoring their utterances to suit a general audience. The method suggested here appears to be sensitive to such differences in the use of jargon in speech tailored to different audiences and rehearsed to different degrees.

Study Limitations

Limitations of the method used. This method treats different orthographic word types separately and assigns them different jargonness scores, including word pairs such as "algorithm" vs. "algorithms" and "vapor" vs. "vapour", although both words in each pair are probably equally "jargony" to non-technical publics.

Next, this analysis ignores the context in which scientific jargon appears, such as if it is clarified. Also, the method ignores different meanings of homographs (e.g., "kitchen sink" vs. "carbon sink"). Finally, it evaluates each word separately, ignoring the difficulty of understanding phrases that may have unique meanings in science (e.g., "the big bang").

Data set sizes. The data sets analyzed (TEDTalks, MICASE, etc.) are rather limited in size for a corpus-based study, making statistical inference difficult.

Different settings. The recordings from the TED conferences were carefully planned and rehearsed monologues for a mostly passive audience, while the MICASE transcripts also include spontaneous conversations. Hence, some differences in the use of jargon may be explained by variations in settings, familiarity and advanced planning of speech, rather than by the intended audience.

Discussion

To the best of our knowledge, this is the first quantitative measure of the proportion and "jargonness" of scientific jargon in science communication. Although preliminary in nature, the method is sensitive to both the pervasiveness and obscurity of the jargon used in a text.

The method can be used for several purposes: (1) Self-evaluation of the jargonness of single words and prevalence of jargon in entire texts; (2) Comparison of student performance before and after training in science communication; (3) Comparison of the effectiveness of different science communication classes.

Future research in the evaluation of science communication skills could develop in many directions. Some possibilities include: (1) A systematic analysis of human ratings of the words' perceived jargonness and of human performance on vocabulary tests may help validate the method proposed, or point to necessary improvements; (2) More statistical measures for "jargonness" should be tested, in conjunction with validation; (3) Interactions between vocabulary choice and situational and personal variables merit further investigation.

Some open questions are: Does rehearsing a message reduce its jargonness? How is the use of jargon affected by training or experience in science communication?

Answering these questions may shed light on the intricate language choices made in science communication. More importantly, it may help scientists learn to talk about science with the public fluently and clearly, and perhaps with less mumbo jumbo.

References

Anthony, L. (2009). *AntWordProfiler*. Tokyo, Japan: Waseda University. Retrieved from http://www.antlab.sci.waseda.ac.jp/

Baram-Tsabari, A., & Lewenstein, B. V. (2012). An instrument for assessing scientists' written skills in public communication of science. *Science Communication*. doi:10.1177/1075547012440634

Baron, N. (2010). *Escape from the ivory tower: a guide to making your science matter*. Washington: Island Press.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1995). *Dimensions of register variation*. Cambridge: Cambridge University Press.

Dean, C. (2009). *Am I making myself clear? A scientist's guide to talking to the public*. Cambridge, Mass.: Harvard University Press.

Halwany, N. (2011). *FreqGrabber*. Retrieved from https://github.com/nadavh/freq_grabber

Hayes, J. R., & Bajzek, D. (2008). Understanding and Reducing the Knowledge Effect: Implications for Writers. *Written Communication*, *25*(1), 104–118. doi:10.1177/0741088307311209

Heatley, A., & Nation, I. S. P. (1994). *Range*. New Zealand: Victoria University of Wellington. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Hoffman, S., & Evert, S. (2006). BNCweb (CQP-edition): The marriage of two corpus tools. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods* (pp. 177–195). Frankfurt am Main: Peter Lang.

Kessler, S. (2011, June 27). With 500 Million Views, TED Talks Provide Hope for Intelligent Internet Video. *Mashable*. Retrieved May 20, 2012, from http://mashable.com/2011/06/27/ted-anniversary/

Lemke, J. L. (1990). *Talking science: language, learning, and values*. Norwood, N.J.: Ablex Pub. Corp.

Ley, P., & Florio, T. (1996). The use of readability formulas in health care. *Psychology, Health & Medicine*, *1*(1), 7–28. doi:10.1080/13548509608400003

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. London [etc.]: Routledge.

Meredith, D. (2010). *Explaining research: How to reach key audiences to advance your work*. New York, N.Y.: Oxford University Press.

Miller, S., Fahy, D., & The ESConet Team. (2009). Can Science Communication Workshops Train Scientists for Reflexive Public Engagement?: The ESConet Experience. *Science Communication*, *31*(1), 116–126. doi:10.1177/1075547009339048

Mulder, H. A. J., Longnecker, N., & Davis, L. S. (2008). The State of Science Communication Programs at Universities Around the World. *Science Communication, 30*(2), 277–287. doi:10.1177/1075547008324878

Nation, I. S. P. (2006). How Large a Vocabulary is Needed For Reading and Listening? *Canadian Modern Language Review/ La Revue canadienne des langues vivantes*, *63*(1), 59–82. doi:10.3138/cmlr.63.1.59

Nisbet, M. C., & Scheufele, D. A. (2009). What's next for science communication? Promising directions and lingering distractions. *American Journal of Botany*, *96*(10), 1767–1778. doi:10.3732/ajb.0900041

Simpson, R. C., Briggs, S. L., Ovens, J., & Swales, J. M. (2002). *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan. Retrieved from http://micase.elicorpora.info/

Stableford, S., & Mettger, W. (2007). Plain Language: A Strategic Response to the Health Literacy Challenge. *Journal of Public Health Policy*, *28*(1), 71–93. doi:10.1057/palgrave.jphp.3200102

[1] Three transcripts of talks that we considered overly technical were omitted from this group.

[2] http://bncweb.lancs.ac.uk

[3] http://scn.jkn21.com/~perc04/