Public Understanding
of Science

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

Journal:	*PUBLIC UNDERSTANDING OF SCIENCE*
Manuscript ID:	PUS-12-0108.R2
Manuscript Type:	Theoretical/ research paper
Keywords:	media and science, interaction experts/ publics, popularization of science, science communication

Abstract:	Leaders of the scientific community encourage scientists to learn effective science communication, including honing the skill to discuss science with little professional jargon. However, avoiding jargon is not trivial for scientists for several psychological and sociological reasons, and this demands special attention in teaching and evaluation. Despite this, no standard measurement for the use of scientific jargon in speech has been developed to date. Here a standard yardstick for the use of scientific jargon in spoken texts, using a computational linguistics approach, is proposed. Analyzed transcripts included academic speech, scientific TEDTalks, and communication about the discovery of a Higgs-like boson at CERN. Findings suggest that scientists use less jargon in communication with a general audience than in communication with peers, but not always less obscure jargon. These findings may lay the groundwork for evaluating the use of jargon.

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

Abstract

Leaders of the scientific community encourage scientists to learn effective science communication, including honing the skill to discuss science with little professional jargon. However, avoiding jargon is not trivial for scientists for several reasons, and this demands special attention in teaching and evaluation. Despite this, no standard measurement for the use of scientific jargon in speech has been developed to date. Here a standard yardstick for the use of scientific jargon in spoken texts, using a computational linguistics approach, is proposed. Analyzed transcripts included academic speech, scientific TEDTalks, and communication about the discovery of a Higgs-like boson at CERN. Findings suggest that scientists use less jargon in communication with a general audience than in communication with peers, but not always less obscure jargon. These findings may lay the groundwork for evaluating the use of jargon.

> "We don't understand our audience well enough – we have not taken the time to put ourselves in the shoes of a neighbor, the brother-in-law, the person who handles our investments – to understand why it's difficult for them to hear us speak. We don't know the language, and we haven't practiced it enough."
>
> – Dr. Neal F. Lane, director of the U.S. National Science Foundation (1993-1998), in Hartz & Chappell (1997), p. 38.

Introduction

The scientific community has increasingly recognized the importance of communicating science to non-technical publics (hereafter "science communication"). Greater resources have increasingly been earmarked to teaching scientists best practices for engaging with the public. Among other goals, this educational endeavor aims to teach scientists to communicate clearly with lay audiences, and in particular, to express ideas in their 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 (Author, 2012). In particular, what is the best way to determine how "jargony" (obscure) a given word is? Second, what criteria can best assess the intelligibility of a text as a whole? Answering these questions

requires some benchmarks for comparison. Authentic high quality instances of science communication may be useful for generating such tools.

Although clarity has been characterized and assessed in other disciplines such as medicine and law (e.g., Benson, 1985), few studies have characterized jargon in science communication. This exploratory, data-driven 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. This quantification may help lay the groundwork for assessments in teaching and learning effective science communication.

Literature Review

Definition and Usefulness of Scientific Jargon

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 (Wardhaugh, 2002). In turn, this gives rise to different *varieties* of language. Varieties are sets of human speech patterns uniquely associated with situations, geographical areas or social groups, such as Cockney, legalese, the English of football commentaries, etc. (Biber, 1995). Any speaker of a language must be able to make use of different varieties of language in different situations, a practice called "code switching" (Wardhaugh, 2002).

In linguistics, varieties "associated with *situational* contexts or purposes" are called registers (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 social role relations among them, the topic and purpose of the communication event and more (Biber, 1988). Theorists have assumed that speakers of a language can instinctively intuit the likelihoods of particular words, groups or phrases in given registers (De Beaugrande, 1991). Thus, one can speak of a *scientific* register of English, used primarily by scientists when communicating about science with their colleagues and students. It is characterized by certain features of grammar, discourse, and vocabulary, such as the proliferation of nominalizations (e.g., "the authors performed DNA extraction" instead of "the authors extracted DNA") and of passive constructions (e.g., "solutions may often be obtained") (Biber, 1995). This work will specifically focus on the specialized vocabulary of scientific register, or *scientific jargon*. (This use of the term "jargon" is consistent with standard dictionary definitions of jargon, such as Merriam-Webster's, "the technical terminology or characteristic idiom of a special activity or group".)

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. Jargonness can be related to obscurity, and thus is an antonym to "vocabulary familiarity" (Keselman et al., 2007).

Research suggests that scientific jargon is a necessary mental tool for scientists, which they draw on in the course of their work (Jucks, Schulte-Löbbert, & Bromme, 2007). In particular, jargon can be seen as a useful set of symbols that has developed over time to aid scientists in representing mental schemes, conceptualizing new facts or discoveries and communicating ideas effectively with their peers (Grupp & Heider, 1975).

Clarity as a Learning Goal in Science Communication Training

Scientists are prolific communicators within their own fields, but few of the findings they share with peers reach the public through the media (Suleski & Ibaraki, 2009). To promote dialog and to garner support and legitimacy for scientific endeavors, the scientific community is increasingly encouraging scientists to engage with the public directly in respectful dialog, to achieve mutual understanding and learning (Leshner, 2009; Nisbet & Scheufele, 2009). Science communication scholars agree that bench scientists, engineers, health and science regulators would benefit from more training in science communication (Besley & Tanner, 2011). Nevertheless, little attention has been paid to defining the goals learners should aim for in such training, and how attainment of these goals should be evaluated. One conceptual framework outlines several measurable components of skills a scientist should have to communicate effectively (Author, 2012).

Specifically, to effectively engage with the public, scientists are advised to convey meaningful scientific ideas without scientific jargon (e.g., Dean, 2009; Hartz & Chappell, 1997; Meredith, 2010). In the words of Stableford and Mettger (2007), "[p]lain language embodies clear communication. While some mistakenly believe that the term means just using simple words, or worse, 'dumbing things down,' it actually refers to communications that engage and are accessible to the intended audience" (p. 75). This transition between technical and ordinary speech when discussing science has been deemed an instance of code switching to achieve clear communication (Montgomery, 1989).

This change in speech patterns is important for effective communication of science for several reasons. First and foremost, it is needed to ensure clarity. Scientists are advised to keep words unfamiliar to the audience to a minimum (under 1 in 50) as understanding a spoken or written text in English requires knowing at least 98% of words used in it (Nation, 2006).

No less importantly, jargon should be avoided to promote positive views of science and scientists, since it has been suggested that "[c]ommunication received in one's own language is crucial for

learning, attitude formation, and behavior change" (Hiż, 1975, p. 34). In a medical setting, for example, patients reported they were more satisfied with doctor's appointments and were more willing to comply with the doctor's instructions when physicians used the same vocabulary as the patients (Williams & Ogden, 2004).

If the deficit model views clarity as important due to its role in effective transfer of knowledge, a framework of public engagement with science views clarity as a prior requirement for engagement. Use of jargon excludes those who are not able to decipher it, and thus handicaps the dialog that would allow scientists to understand non-scientists' ideas and perceptions of science-related issues (Burns, O'Connor, & Stocklmayer, 2003).

Thus, for cognitive, emotional and social reasons, scientists should avoid jargon and express themselves in generic terms when engaging with the public. However, this is easier said than done, as experts use jargon excessively for several reasons.

*Lack of Motivation***.** First, in science and in other disciplines, some experts object to expressing themselves in everyday language out of principle reinforced by social norms. Legalese, for example, has been hypothesized to persist among lawyers due to self-interest, supported by inertia, incompetence, status, wariness of change, and the appeal of intimidating and confusing non-lawyers such as juries and witnesses (Benson, 1985). Moreover, both in law and in medicine, it has been argued that jargon is necessary for accurate writing, and that clear, simple writing is necessarily dull and condescending in tone. Rebuttals to these claims can be found in the literature (Benson, 1985; Stableford & Mettger, 2007). Similar motivations may deter a scientist from communicating clearly with the public, especially as some scientists say public outreach may incur professional stigma (Burchell, Franklin, & Holden, 2009).

*Lack of Skill***.** Even well-intentioned experts use jargon when they should not because they fail to assess their addressees' knowledge level. For example, physicians in San Francisco have been shown to use unclear jargon in 81% of patient encounters four times per visit on average (Castro, Wilson, Wang, & Schillinger, 2007). When medical students were asked to answer fictitious patients' medical questions in written form over the internet, they used medical jargon in their answers, even if the question was phrased entirely in everyday words (Bromme, Jucks, & Wagner, 2005). Another study found that although over 85% of science students recognize terms such as "epigenetic" as jargon that should be defined when writing to a non-technical audience, they also make liberal use of advanced jargon when describing their own work (Author, 2012).

Research in the sociology of education and science has suggested that when learning science, people are enculturated into an academic community and learn how to "talk science" like scientists

(Lemke, 1990). Similarly, situated learning theory claims that when engaging in authentic scientific activities, individuals learn the scientific jargon as a necessary tool for the task (Brown, Collins, & Duguid, 1989). Montgomery (1989) theorized that scientific jargon has become so entrenched into scientific practice that it has become inseparable from science itself, which may explain why scientists find it difficult to communicate without jargon.

"Curse of Knowledge". It is difficult to avoid jargon because of a cognitive bias called the "Curse of Knowledge": When individuals assess another person's perspective, they overestimate what the other person knows, because their judgment is impaired by their own knowledge. Thus, for example, when adults know the outcome of an event, they overestimate another person's capability to correctly predict the outcome (Birch & Bloom, 2004). Similarly, if undergraduate students are familiar with a technical term, they overestimate how many other people understand it (Hayes & Bajzek, 2008), and scientists may overestimate public familiarity with scientific jargon.

Thus overall, it is difficult for experts to avoid scientific jargon when discussing their field of expertise with non-experts. Clarity in expert communication with the public is impeded both by sociological and psychological factors. 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). In the words of one communication guide, "[t]here are, in fact, only two ways to beat the Curse of Knowledge reliably. The first is not to learn anything. The second is to take your ideas and transform them" (Heath & Heath, 2008, p. 20). We argue here that research-based strategies to support this transformation should be both explicitly taught in science communication training and rigorously assessed.

Evaluating Clarity in Science Communication

There have been two main approaches to assess the understandability of any text, and in particular, to evaluate the clarity of a scientific text. The first approach uses readability formulas, and the second analyzes the vocabulary used, either based on short word lists or on large bodies of authentic texts.

Readability formulas. Readability formulas are regression equations that utilize parameters such as word length (measured in syllables), and sentence length (in words), to predict the level of difficulty in reading a given text (Ley & Florio, 1996). These formulas are usually validated by performance on comprehension tests provided to students in different grades. Thus, by plugging in the parameters of a new text into the formula, the text's estimated reading grade level can be found.

One common readability metric is the Flesch Reading Ease score (Flesch, 1948). Flesch found scientific journals to be "very difficult" to read, a finding that has been corroborated frequently and

recently for leading medical journals BMJ and JAMA (Weeks & Wallace, 2002) and a geology journal (Hartley, Sotto, & Fox, 2004). Based on the Flesch score and other formulas, it is estimated that only about 5% of the US population can read and understand these medical journals. Also, more alarmingly, even most medical literature intended for *patients* is estimated to be too complex for most patients to read (Ley & Florio, 1996; Stableford & Mettger, 2007).

Readability formulas are convenient, widely employed and based on sound data and methodology. Even so, relying on sentence and word length neglects several facets of the perceived difficulty of a scientific text, e.g.: (1) Short words that can be hard to understand (e.g., "average" vs. "mean"); (2) Short yet confusing sentences (e.g., "These parts store iron ions cells bind") (3) Non-textual features such as numbers and formulas, (4) The text's overall audience appeal, cultural appropriateness, tone, etc.; and, most importantly for this study, (5) The reader's background knowledge of the topic being discussed, and in particular, the reader's familiarity with the vocabulary used (e.g., "Plants fix carbon") (Hartley et al., 2004; Stableford & Mettger, 2007).

Vocabulary analysis – Word list based. Vocabulary analysis determines how much vocabulary a person needs to understand a text. Analyses have assessed how many words in a text belong to (1) a short list of common words, such as in the Dale-Chall formula (Dale & Chall, 1948), (2) a database of words familiar to students at different school grade levels, such as Dale and O'Rourke (1981), (3) a list of common words in academic texts (Coxhead & Hirsh, 2007; Coxhead, Stevens, & Tinkle, 2010), or (4) a business jargon database (Business Idiots, LLC, 2005; Ley & Florio, 1996). These databases are often difficult to obtain, insufficiently documented or outdated.

Most importantly, this approach is limited in scope: Relying on short, closed and predetermined word lists is not flexible enough to capture the wide range of scientific words in a text that are *not* included on these lists. Instead, larger samples of the language can be used, which is the approach we take in this study. This method is called *corpus-based* linguistic analysis.

Vocabulary analysis – Corpus-based. A *corpus* is a large collection of natural texts, written or spoken, in machine-readable form, which may be annotated with various forms of linguistic information (McEnery, Xiao, & Tono, 2006). A corpus includes authentic texts, which adequately represent a particular language or language variety: General corpora are used for an overall description of a language or language variety, and specialized, unbalanced corpora tend to be domain- or genre-specific, such as a newspaper text corpus, a corpus of film subtitles or a legal text corpus (McEnery et al., 2006).

Corpora have been used to study vocabulary, often by relying on *word frequency*, defined as the number of occurrences of a word in a given text or corpus (Paquot & Bestgen, 2008). Researchers

have used word frequency data to infer "what a text is really about", and to learn about language variation in different groups and contexts (Scott & Tribble, 2006, pp. 55–56). Some studies have focused on comparing the high-frequency words in texts (e.g., the, of, and). Other studies have focused on medium-to-low frequency words, such as scientific jargon. For a comprehensive review on comparing corpora, see Kilgarriff (2001).

Jargon has been evaluated using corpora in at least two ways: (1) A machine learning approach to estimating the level of a health information text based on the frequencies of its words in large corpora of medical information (Leroy, Miller, Rosemblat, & Browne, 2008), and (2) Using internet news websites as a corpus to gauge the familiarity of scientific words through *Google News* hits (Author, 2012). The last study was limited in its accuracy, transparency and stability, as the *Google News* corpus is constantly changing, and hit numbers displayed to the user are only approximations.

Here we present a new, flexible and transparent method, based on large, freely available corpora, to assess the extent of use of scientific jargon in science communication. In this study, we put our new method to the test, attempting to quantify jargon use in light of our hypotheses.

Hypotheses

1. Jargon is *less pervasive* in popular science communication than within communication among scientists.
2. Effective science communication uses *jargon that is less obscure* than the jargon found in 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, up to 18 minutes long, featured in the TED conferences. TED (originally "Technology, Entertainment, Design") is a nonprofit organization that holds two annual conferences in California and Scotland to promote "ideas worth spreading," on topics such as entertainment and design but also economics, science, and education. The TED website has made over 1,200 videos of TEDTalks freely available online, and over a quarter (n = 330) have been tagged under "science" (TED Conferences, LLC, 2012). Many scientific TEDTalks are delivered by scientists. Online TEDTalks are extremely popular, accumulating over half a billion views in total to date (Kessler, 2011). Given their popularity, the high proportion of science videos and their high

quality, we drew on science-related 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). About 68% of these TEDTalk transcripts were by scientists and engineers (e.g., physicists and marine biologists) and the rest were by other professionals (e.g., historians of science). TEDTalk transcripts in English are professionally transcribed and approved by TED.

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 nearly 200 hours of recordings by the English Language Institute (ELI) at the University of Michigan. Designed for research in contemporary English university speech, MICASE spans various settings, including lectures, classroom discussions, lab sections, seminars, and advising sessions (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).

External validity. To examine the external validity of the method, we applied it to samples of transcripts of two events communicating the discovery of a Higgs-like particle that took place on July 4, 2012 at CERN: (1) Two scientific seminars about the findings, by Prof. Joe Incandela from the CMS collaboration and Dr. Fabiola Gianotti from the ATLAS collaboration, both delivered to an audience of fellow particle physicists; and (2) statements by the same two spokespeople at the press conference immediately following the seminars, addressing an audience of non-specialists. Approximately 10 minutes of Incandela and Gianotti's utterances from each of the events were sampled randomly and transcribed by the first author, yielding two roughly equal sized transcripts, with 1,572 tokens for the seminars and 1,645 tokens for the press conference.

Preparation for Analysis. For each type of transcripts, measures were taken to omit metadata, partially uttered words and text annotations, such as "(Applause)".

Data Sources – Limitations

Data set sizes. The data sets analyzed (TEDTalks, MICASE, etc.) are rather limited in size for a corpus-based study, making statistical inference difficult. These small corpora were used bearing in mind the exploratory nature of this study and due to practical limitations of data availability for these spoken registers. Further replications in larger corpora may shed more light on the method's efficacy.

Different settings and transcription standards. 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. Moreover, the transcripts employed different transcription standards. Words partially uttered and transcribed (e.g., "You can't understand how somebody thinks, in ano- in another society") were extremely prevalent in MICASE, and all those that were two characters long were omitted from the transcripts. It is believed that most of the remaining partially transcribed words were classified as Unknown (Category E) and not as jargon (see below).

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 96,986,707 orthographic words, and was designed to represent a wide range of British English, as it was used between 1960 and 1993. Written texts comprise 90% of the corpus, including samples of newspapers, academic books, popular fiction and unpublished letters, and the remaining 10% are transcripts of spoken data, including radio shows, formal government meetings and informal conversations from respondents of various ages, social classes and regions in the UK. The BNC was compiled by the BNC Consortium, an industrial/academic group led by Oxford University Press, and is publicly accessible via web interfaces such as BNCweb[2] (Hoffman & Evert, 2006).

General corpus – Limitations. While the BNC is generally accepted as being a balanced corpus (McEnery et al., 2006), it has three major limitations for this study: (1) It is largely written, British, formal and adult, and this affects the distribution of the words in the lists (Nation, 2006). Particularly, it raises a possible dialect problem when comparing word frequencies with non-UK data sets; (2) As its most recent parts are from 1993, many words that have come into common usage since then, such as "website", are conspicuously absent; and (3) The BNC is not a "science-

free" corpus, nor was it designed to accurately represent public familiarity with science terminology. Rather, it contains some transcripts of university lectures about science and other similarly academic sources.

Scientific corpus. To represent the scientific variety 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 the journals with the top 20% impact factor in 22 fields of science, engineering, technology and other fields. The PERC Corpus was compiled by the Professional English Research Consortium (PERC), a Japan-based association of scholars, educators, and related professionals and organizations, and is also publicly accessible 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, to focus on uncommon words, 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)[4]. 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). Also excluded were words appearing in pre-assembled lists of (1) proper names (e.g., "Galapagos," "Einstein") and (2) interjections, exclamations, and hesitations (e.g., "Umm," "Oh"; BNC Word Family Lists 15 and 16, respectively).

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 their 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 in the *general corpus* and (2) Its frequency in the *scientific corpus*. As these corpora had different sizes (~100 million and ~17 million, respectively), we compared *normalized* frequency values, namely the frequency of that word's appearance *per million words* in each corpus (McEnery et al., 2006). 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 (Dunning, 1993). Words appearing more frequently in the scientific corpus, and whose log-likelihood statistic was above the 95th percentile (i.e., $p < 0.05$), were considered *significantly* more frequent in the scientific corpus (Category B1; critical value 3.84). 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), were classified as scientific jargon. (Granted, it is possible that some specific jargon may be found in high frequencies in BNC, perhaps sometimes at higher frequencies than in PERC, but these were ignored to err on the side of caution.)

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, akin to the *weirdness ratio* value from computational linguistics (Ahmad, 1992).

The common logarithm was then 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, e.g., by tens (10^1), hundreds (10^2) or thousands (10^3). This happens because word frequencies have a very skewed distribution, described by Zipf's law (Kilgarriff, 2001). By extracting the common (base-10) logarithm of the quotient of frequencies, one easily notices the order of magnitude of this quotient. For example, the word "solubilities" is over 213 times more common in the scientific corpus than in the general corpus. This is a difference in the hundreds, or of *two* orders of magnitude. Accordingly, "solubilities" has a jargonness score slightly above two, at 2.33 ($\log_{10}(213) \approx 2.33$); By comparison, "agroecosystem" is 1,091 times more common in the scientific corpus than in the general one, or *three* orders of magnitude greater – hence its jargonness is 3.04.

(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:

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

[Fig. 1 about here]

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"; "sulphur" vs. "sulfur"; and "vapor" vs. "vapour", although both words in each pair are probably equally "jargony" to non-technical publics. Hence, the method can be improved by grouping words by their root forms, or lemmata, and comparing the frequencies of those, rather than of the word types. However, this requires more technical expertise from the researcher/evaluator. Next, this analysis ignores the context in which scientific jargon appears, treating a word equally whether if it

was explained in everyday words, or without clarification. Also, the method ignores different meanings of homographs (e.g., "kitchen sink" vs. "carbon sink"). Finally, it breaks up multiword phrases, counting each word separately, ignoring the difficulty of understanding phrases that have unique meanings in science (e.g., "the big bang"). These discrepancies were not remedied in this study, but future work should seek to lemmatize words and standardize transcription styles before analysis, and account for multiword units.

Results

*Identification of Uncommon Words***.** To pinpoint jargon, we identified uncommon English words from three sets of transcripts, assuming that part of these uncommon words would be jargon. The proportions of uncommon words from the total word counts were compared. 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.

[Table 1 about here]

*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 & Fig. 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, $p < 0.01$).

[Table 2 about here]

[Figure 2 about here]

*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 (Fig. 3). 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). Thus in academic scientific speech, jargon had a much higher jargonness score than the jargon extracted from the science communication transcripts from TEDTalks, regardless of whether the TEDTalks were about science or design.

[Figure 3 about here]

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 ($n_{seminars,\ tokens}$ = 1,572; $n_{seminars,\ types}$ = 473), versus statements in the press conference on the same topic by the same two spokespeople at CERN ($n_{press\ conf.,\ tokens}$ = 1,645; $n_{press\ conf.,\ types}$ = 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 (e.g., "topologies" (jargonness 1.84) and "calibration" (jargonness 1.68)).

Thus overall, most words used in the transcripts were common words, and only less than 3.5% of the tokens were uncommon (not found in the 9,000 most common word families), which is consistent with Zipf's law and other previous works on word frequency (Brossard & Shanahan, 2006; Nation, 2006). Among the uncommon words in each group, academic scientific speech contained significantly more jargon than science communication, by a factor ranging from 1.15 to 2.44. 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 at CERN vs. Higgs boson press conference at CERN) 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.

Apart from the control group, most speakers sampled in the transcripts were scientists or science students, with a maximum of only 32% non-scientist speakers in one of the groups (TED Science). The observed shift in lexical choice might partly be explained as a result of speakers tailoring their utterances to suit a general audience, which is an instance of code switching. When scientists address the public they may sometimes opt to use *less* jargon, but not always *less obscure* jargon. The use of relatively obscure words at the CERN press conference (median jargonness 1.65) may suggest that the speakers' skill at code shifting was poorer than the "gold standard" of TEDTalks (median jargonness 0.97). This could be explained by the fact that it was unrehearsed speech, a type of speech that may be most at risk of incurring the "Curse of Knowledge" (Keysar, Barr, & Horton, 1998). The method suggested here appears to be sensitive to such differences in the use of jargon in speech tailored for different audiences and rehearsed to different degrees.

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, and may serve to evaluate both per-word and per-text "jargonness" based on word usage patterns that are empirically measured, rather than based on intuition alone.

To measure jargonness, one only needs access to several computer applications and datasets, all available free of charge, mentioned here in order of use:

(1) *AntWordProfiler.* This program can receive any text, and copy uncommon words from it into a file (Anthony, 2009);

(2) *BNC Word Family Lists 1-9, 15-16.* AntWordProfiler needs these lists to identify uncommon words; Packaged with Range (Heatley & Nation, 1994);

(3) *A spreadsheet application.* E.g., LibreOffice Calc (free), or Microsoft Excel (non-free); This retrieves uncommon words from the AntWordProfiler output file;

(4) *FreqGrabber.* This script receives a list of (uncommon) words, retrieves each word's frequencies in BNC and in PERC, and records this data in spreadsheets (Halwany, 2011).

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.

Overall, while this method takes mere minutes to apply, it can be automated further and consolidated into a single software package. Ideally, it would also highlight technical terms and provide the user with an opportunity to revise accordingly, as in Jucks et al. (2007). Perhaps a future development could also suggest alternative words, just as some medical databases associate "consumer-friendly display" names such as "kneecap" with technical terms and concepts such as "patella" (Zeng & Tse, 2006).

Concluding Remarks

The main contribution of this paper is its application of linguistics to the assessment of clarity in science communication, as well as integrating separate threads of studies in linguistics, science, medicine and law to paint a broad picture of jargon, public literacy and the assessment of clarity. This study was able to quantify salient differences in the use of jargon in different types of scientific communication. The ecological validity of this study is based on the analysis of authentic speech of real scientists and science communicators addressing real audiences, rather than on subjects' speech in a laboratory setting.

Future research in the evaluation of science communication skills could develop in many directions.

First, the data generated by the method should be put to the test of human evaluation. If one word has a jargonness score of 1.5 and another scored 1.75, can members of non-technical publics usually tell the difference? Also, are they usually less familiar with the more jargony word? Answering these questions would require a systematic analysis of human ratings of the words' perceived jargonness and human performance on vocabulary tests. Also, it is worth assessing how well non-technical publics understand entire texts which have different overall jargonness statistics. These studies could help develop ways to predict public familiarity with a scientific term or public comprehension of a scientific text.

Second, how does the measure of "jargonness" of a word compare to other measures? More statistical measures for "jargonness" should be tested, perhaps in combination with higher-stringency thresholds for the inclusion of uncommon words, such as a minimum number of appearances in the target corpora (Paquot & Bestgen, 2008).

Third, several interactions between vocabulary choice and situational and personal variables merit further investigation. 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 heed Dr. Neal F. Lane's call to learn to talk about science with the public fluently and clearly, and with less mumbo jumbo.

Acknowledgements

We thank X for contributing the FreqGrabber program, Y for his expert statistical advice and the two anonymous reviewers for their insightful comments.

References

Ahmad, K. (1992). What is a term? The semi-automatic extraction of terms from text. In M. Snell-Hornby, F. Poechhacker, & K. Kaindl (Eds.), *Translation studies: an interdiscipline* (1st ed., pp. 267–278). Amsterdam: John Benjamins Pub Co.

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

Author. (2012).*Science Communication*.

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

Benson, R. W. (1985). The End of Legalese: The Game is Over. *New York University Review of Law and Social Change*, *13*(3), 519–573.

Besley, J. C., & Tanner, A. H. (2011). What Science Communication Scholars Think About Training Scientists to Communicate. *Science Communication*, *33*(2), 239–263. doi:10.1177/1075547010386972

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

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

Birch, S. A. J., & Bloom, P. (2004). Understanding children's and adults' limitations in mental state reasoning. *Trends in Cognitive Sciences*, *8*(6), 255–260. doi:10.1016/j.tics.2004.04.011

Bromme, R., Jucks, R., & Wagner, T. (2005). How to refer to "diabetes"? Language in online health advice. *Applied Cognitive Psychology*, *19*(5), 569–586. doi:10.1002/acp.1099

Brossard, D., & Shanahan, J. (2006). Do They Know What They Read? Building a Scientific Literacy Measurement Instrument Based on Science Media Coverage. *Science Communication*, *28*(1), 47–63. doi:10.1177/1075547006291345

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated Cognition and the Culture of Learning. *Educational Researcher*, *18*(1), 32–42. doi:10.3102/0013189X018001032

Burchell, K., Franklin, S., & Holden, K. (2009). *Public culture as professional science: final report of the ScoPE project – Scientists on public engagement: from communication to deliberation?* BIOS, London School of Economics and Political Science. Retrieved from http://eprints.kingston.ac.uk/20016/1/ScoPE_report_-_09_10_09_FINAL.pdf

Burns, T. W., O'Connor, D. J., & Stocklmayer, S. M. (2003). Science Communication: A Contemporary Definition. *Public Understanding of Science*, *12*(2), 183–202. doi:10.1177/09636625030122004

Business Idiots, LLC. (2005). Fight the Bull. Retrieved October 9, 2012, from http://www.fightthebull.com/bullfighter.asp

Castro, C. M., Wilson, C., Wang, F., & Schillinger, D. (2007). Babel babble: physicians' use of unclarified medical jargon with patients. *American journal of health behavior*, *31 Suppl 1*, S85–95. doi:10.5555/ajhb.2007.31.supp.S85

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliquée*, *XII*(2), 65–78.

Coxhead, A., Stevens, L., & Tinkle, J. (2010). Why Might Secondary Science Textbooks be Difficult to Read? *New Zealand Studies in Applied Linguistics*, *16*(2), 37–52.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin*, *27*, 11–20.

Dale, E., & O'Rourke, J. (1981). *The Living Word Vocabulary: A national vocabulary inventory*. Chicago: World Book-Childcraft International.

De Beaugrande, R. (1991). *Linguistic theory: the discourse of fundamental works*. London; New York: Longman.

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

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, *19*(1), 61–74.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. doi:10.1037/h0057532

Grupp, G., & Heider, M. (1975). Non-Overlapping Disciplinary Vocabularies. In S. B. Day (Ed.), *Communication of scientific information.* (pp. 28–36). Basel: Karger.

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

Hartley, J., Sotto, E., & Fox, C. (2004). Clarity Across the Disciplines: An Analysis of Texts in the Sciences, Social Sciences, and Arts and Humanities. *Science Communication*, *26*(2), 188–210. doi:10.1177/1075547004270164

Hartz, J., & Chappell, R. (1997). *Worlds Apart: How the Distance Between Science and Journalism Threatens America's Future* (No. 98-F02). Nashville, TN: First Amendment Center. Retrieved from http://www.freedomforum.org/publications/first/worldsapart/worldsapart.pdf

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

Heath, C., & Heath, D. (2008). *Made to stick: Why some ideas survive and others die*. New York: Random House.

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

Hiż, H. (1975). Specialized Languages of Biology, Medicine and Science, and Connections between Them. In S. B. Day (Ed.), *Communication of scientific information.* (pp. 37–43). Basel: Karger.

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.

Jucks, R., Schulte-Löbbert, P., & Bromme, R. (2007). Supporting Experts' Written Knowledge Communication Through Reflective Prompts on the Use of Specialist Concepts. *Zeitschrift für Psychologie / Journal of Psychology*, *215*(4), 235–245.

Keselman, A., Tse, T., Crowell, J., Browne, A., Ngo, L., & Zeng, Q. (2007). Assessing Consumer Health Vocabulary Familiarity: An Exploratory Study. *Journal of Medical Internet Research*, *9*(1), e5. doi:10.2196/jmir.9.1.e5

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/

Keysar, B., Barr, D. J., & Horton, W. S. (1998). The Egocentric Basis of Language Use: Insights From a Processing Approach. *Current Directions in Psychological Science*, *7*(2), 46–50. doi:10.1111/1467-8721.ep13175613

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, *6*(1), 1–37.

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

Leroy, G., Miller, T., Rosemblat, G., & Browne, A. (2008). A balanced approach to health information evaluation: A vocabulary-based naïve Bayes classifier and readability formulas. *Journal of the American Society for Information Science and Technology*, *59*(9), 1409–1419. doi:10.1002/asi.20837

Leshner, A. I. (2009, July 9). AAAS News Release - "Alan I. Leshner: Commentary on the Pew/AAAS Survey of Public Attitudes Toward U.S. Scientific Achievements." *AAAS*. Retrieved September 30, 2012, from http://www.aaas.org/news/releases/2009/0709pew_leshner_response.shtml

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.

Montgomery, S. L. (1989). The cult of Jargon: Reflections on language in science. *Science as Culture*, *1*(6), 42–77. doi:10.1080/09505438909526248

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

Paquot, M., & Bestgen, Y. (2008). Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In A. H. Jucker, D. Schreier, & M. Hundt (Eds.),

Corpora: Pragmatics and Discourse (pp. 247–269). Presented at the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29), Ascona, Switzerland.

Scott, M., & Tribble, C. (2006). *Textual patterns: key words and corpus analysis in language education*. Amsterdam; Philadelphia: John Benjamins Pub.

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

Suleski, J., & Ibaraki, M. (2009). Scientists are talking, but mostly to each other: a quantitative analysis of research represented in mass media. *Public Understanding of Science*, *19*(1), 115–125. doi:10.1177/0963662508096776

TED Conferences, LLC. (2012, June 11). TED Talks. Retrieved June 11, 2012, from http://www.ted.com/talks

Wardhaugh, R. (2002). *An introduction to sociolinguistics*. Malden, MA: Blackwell Pub.

Weeks, W. B., & Wallace, A. E. (2002). Readability of British and American medical prose at the start of the 21st century. *BMJ (Clinical research ed.)*, *325*(7378), 1451–1452.

Williams, N., & Ogden, J. (2004). The impact of matching the patient's vocabulary: a randomized control trial. *Family Practice*, *21*(6), 630–635. doi:10.1093/fampra/cmh610

Zeng, Q. T., & Tse, T. (2006). Exploring and Developing Consumer Health Vocabularies. *Journal of the American Medical Informatics Association*, *13*(1), 24–29. doi:10.1197/jamia.M1761

[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/

[4] The BNC Word Family Lists contain sets of 1,000 word families ranked by descending frequency in the BNC. The first 1,000 families contain common words, such as "red" and "story", whereas the 9,000th most common families contain rarer words, such as "slumber" and "tornado".

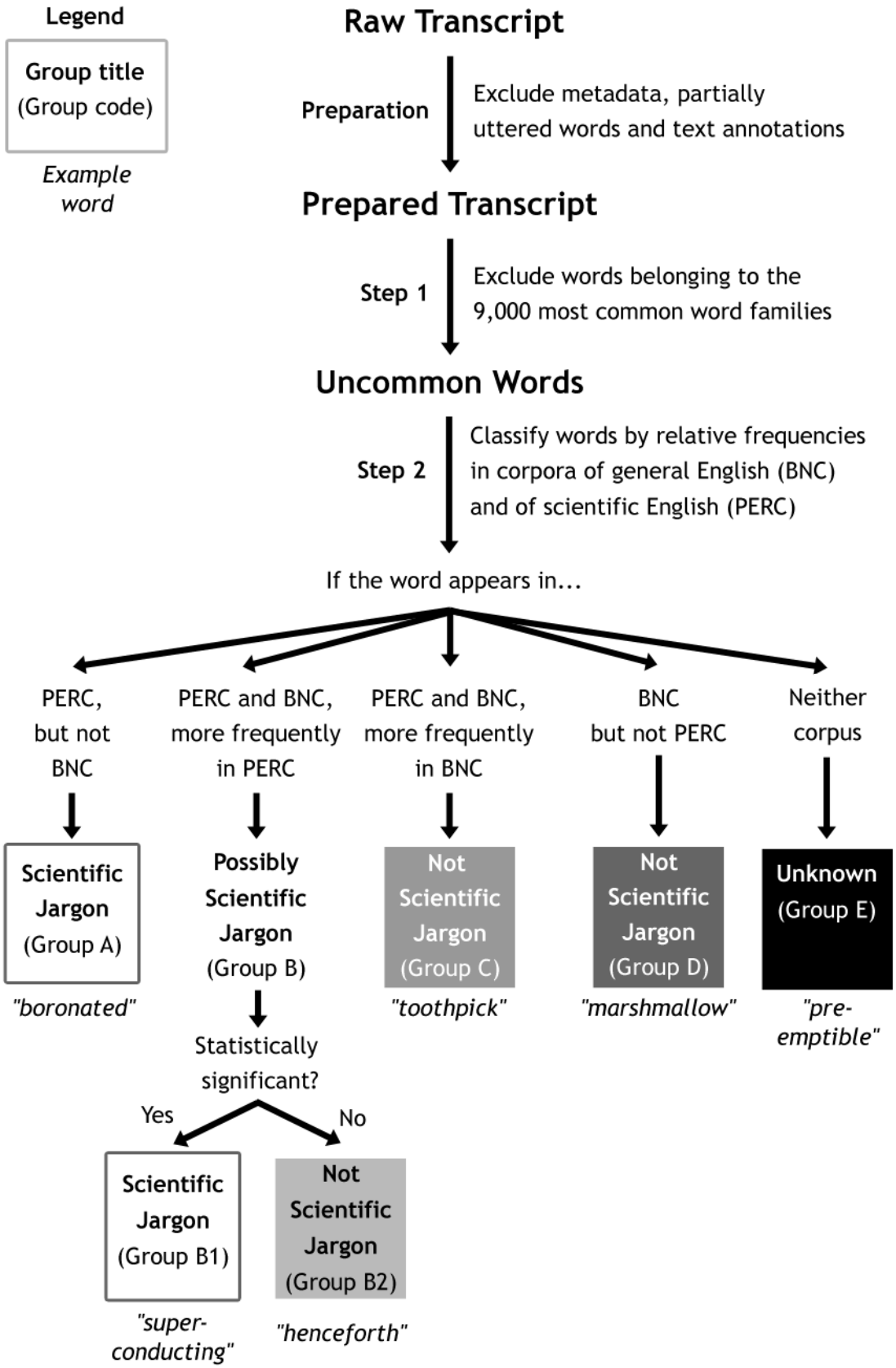

Figure 1. Methodology for classification of words in a spoken text to jargon and non-jargon.

	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.

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.

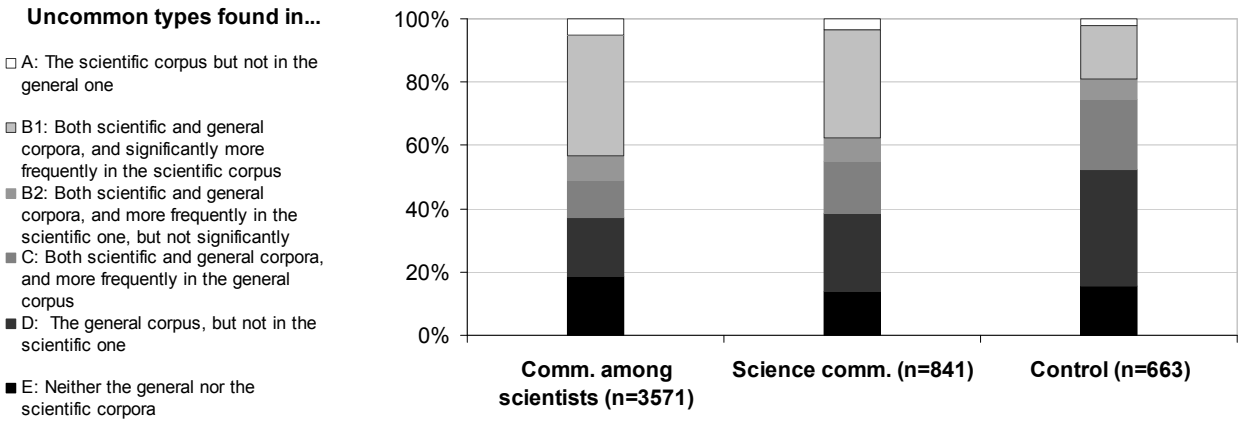

Figure 2. Pervasiveness of scientific jargon in uncommon types from three groups of transcripts: Scientific academic speech (MICASE), science communication (TED Science) and a non-scientific control group (TED Design). Words classified in categories A or B1 are considered scientific jargon.

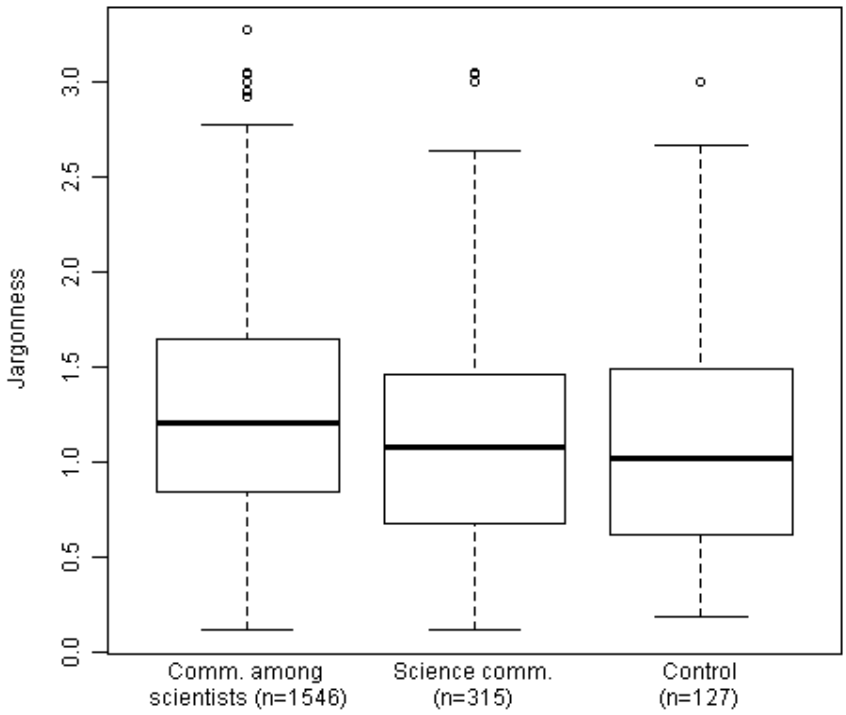

Figure 3. Jargonness (obscurity) of jargon (Categories A & B1). Uncommon types were extracted from three groups of transcripts: Scientific communication among scientists in academic settings ("Comm. among scientists," MICASE), science communication ("Science comm.," TED Science) and a non-scientific control group ("Control," TED Design). Within each box-and-whisker plot, the black band signifies the median, the hinges mark the lower and upper quartiles, and the whiskers span all data points within 1.5 times the inter-quartile range. Communication among scientists has higher jargonness than science communication (Wilcoxon-Mann-Whitney Test. $p < 0.001$).