How to Best Write Research Papers: Basic English? Sophisticated English?

Martine Ceberio, Christian Servin, Olga Kosheleva, and Vladik Kreinovich

Abstract Instructors from English department praise our students when they use the most sophisticated grammatical constructions and the most appropriate (often rarely used) words – as long as this helps better convey all the subtleties of the meaning. On the other hand, we usually teach the students to use the most primitive Basic English when writing our papers – this way, the resulting paper will be most accessible to the international audience. Who is right? In this paper, we analyze this question by using a natural model – inspired by Zipf's law – and we conclude that to achieve the largest possible effect, the paper should be written on an intermediate level – not too primitive, not too sophisticated (actually, on the level of the middle school).

Martine Ceberio
Department of Computer Science, University of Texas at El Paso, 500 W. University
El Paso, TX 79968, USA, e-mail: longpre@utep.edu

Christian Servin
Computer Science and Information Technology Systems Department
El Paso Community College (EPCC), 919 Hunter Dr., El Paso, TX 79915-1908, USA
e-mail: cservin1@epcc.edu

Olga Kosheleva
Department of Teacher Education, University of Texas at El Paso, 500 W. University
El Paso, TX 79968, USA, e-mail: olgak@utep.edu

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, 500 W. University
El Paso, TX 79968, USA, e-mail: vladik@utep.edu

1 Formulation of the Problem

Tension between English classes and what we teach. There seems to be a systemic tension between what our students learn in their English classes and what we teach them when describing how to write scientific papers:

- In English classes, use of rare words and complex constructions is strongly encouraged if it provides a more adequate description of the message. If the English teacher says that the essay was written on the 5-th grade level, this is *not* a compliment, such an essay would not get an Excellent grade.
- On the other hand, when we teach students, we tell them to write in Basic English – since science is an international endeavor, and many foreign researchers do not know English that well to know rare words and rare constructions.

Problem. So, what is the optimal level? If we write in too complex a language, we will miss most of the audience, and the impact of the paper will be small. On the other hand, if we write in too simple a language, we do not convey many subtleties of the meaning – and thus, decrease the impact as well.

What we do in this paper. In this paper, we analyze this problem, and we show what is the optimal level of language complexity.

2 Towards Formulating the Problem in Precise Terms

Levels of complexity. Even native speakers of English are not born with the knowledge of all the language's words and constructions, they acquire is as they study. This provides a natural scale for the language complexity used by linguists: we can be on the level of corresponding to the average language level of kindergarten students, we can be on the level of the 1st grade, ..., level of the 12th grade, of the 1st year of college, etc. Overall, there are about 20 different levels, all the way to PhD level.

For simplicity, we will simply mark them by numbers from 1 to 20, so that Level 1 corresponds to the most basic use of language, and Level 20 to the most sophisticated use of the language.

How widely spread are different levels. Clearly, many folks around the world have a very basic knowledge of English – and are thus on Level 1, a little fewer are on Level 2, ..., all the way to very complex Level 20 on which there is a small minority. How many people are on each level?

A reasonable idea is to use Zipf's law (see, e.g., [1, 3, 4]) for estimating the relative number of people on each level. This law was first observed in linguistics, where it turned out that if we sort all the words from a language in the reverse order of their frequencies f_i, so that

$$f_1 \geq f_2 \geq f_3 \geq \ldots,$$

then we have

$$f_n = \frac{c}{n}. \tag{1}$$

for some constant c. So, the second most frequent word is twice less frequent that the most frequent one, the third most frequent word is three times less frequent, etc.

It turned out that a similar formula (1) is ubiquitous not only in linguistics, it is ubiquitous in many other application areas (see, e.g., [2, 5]) – and there are good explanations for its ubiquity; see, e.g., [1, 3].

Because of this ubiquity, it makes sense to apply this law to our situation as well, and to assume that the number of people of the i-th level of knowledge is proportional to $1/i$.

What is the impact of different readers. The overall impact of a paper comes from combining the impacts on different readers. Intuitively, it is clear that the most sophisticated – thus, the most learned – readers can provide the largest impact, both in terms of the effect on their own work and in terms of them spreading the word around, while readers who have just started doing research will have, on average, the smallest impact.

Here, readers on the last – n-th level ($n = 20$) have the largest impact, readers on the $(n-1)$-st level have a slightly smaller impact, etc., all the way to people on the 1st level who have, on average, the smallest impact. It makes sense to use Zipf's law to describe how this impact decreases: folks on the n-th level have the highest impact I, folks on the next $(n-1)$-th level have impact $\frac{I}{2}$, folks on the $(n-2)$-nd level have the impact $\frac{I}{3}$, etc., and, in general, folks on level i have the impact $\frac{I}{n+1-i}$.

The overall impact-per-unit-of-information of all the folks on level i can be obtained if we multiply the number of people on this level – which is proportional to $\frac{1}{i}$ – and the impact of each of these folks, which is proportional to $\frac{1}{n+1-i}$. Thus, this overall impact I_i is proportional to the product

$$I_i \sim \frac{1}{i \cdot (n+1-i)}. \tag{2}$$

How much information is conveyed on each level. A big portion of information can be conveyed already on the very first Level 1. If we allow Level 2, then an additional portion of the original information can be conveyed, etc., and if we go from Level $n-1$ to Level n, a few very subtle places can finally be conveyed. Intuitively, as we go to a higher and higher level, the portion of new information conveyable by this new level decreases. It is therefore reasonable to us Zipf's law to describe these portions as well: if we denote the portion that can be conveyed on Level 1 by p, then the new portion whose conveyance becomes possible on Level 2 is approximately equal to $\frac{p}{2}$, the new portion whose conveyance has become possible on Level 3 is approximately equal to $\frac{p}{3}$, etc.

So, if we use Level k to write our paper, then the portion of information conveyed by this paper can be obtained by adding up all the portions corresponding to Levels 1 through k and is, thus, proportional to the sum

$$1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{k}. \tag{3}$$

So what is the overall impact of the paper: towards the final formula. If we write a paper on Level k, then the portion of information that we convey is limited by folks on this level or higher. The overall impact-per-piece of information of all these folks can be obtained by adding the impacts (2) corresponding to Levels k through n:

$$\frac{1}{k\cdot(n+1-k)}+\frac{1}{(k+1)\cdot(n-k)}+\ldots+\frac{1}{n\cdot 1}. \tag{4}$$

Thus, the overall effect E of the paper can be obtained by multiplying the amount (3) of conveyed information and the impact (4) per piece of information:

$$E=$$
$$\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{k}\right)\cdot\left(\frac{1}{k\cdot(n+1-k)}+\frac{1}{(k+1)\cdot(n-k)}+\ldots+\frac{1}{n\cdot 1}\right). \tag{5}$$

What we will do. We will find the level k for which the effect E of the paper is the largest.

3 So Which Level Is Optimal: Towards the Answer

Simplification. To simplify the expression (3), let us introduce a special notation for the first factor in the expression (5):

$$S_k \stackrel{\text{def}}{=} 1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{k}. \tag{6}$$

The second factor in the expression (5) can also be represented in terms of the values S_i if we take into account that for every i, we have

$$\frac{1}{i}+\frac{1}{n+1-i}=\frac{n+1}{i\cdot(n+1-i)}.$$

Thus,

$$\frac{1}{i\cdot(n+1-i)}=\frac{1}{n+1}\cdot\left(\frac{1}{i}+\frac{1}{n+1-i}\right).$$

So, the sum (4) can be reformulated as

$$\frac{1}{n+1}\cdot\left(\frac{1}{k}+\ldots+\frac{1}{n}+\frac{1}{1}+\ldots+\frac{1}{n+1-k}\right)=\frac{1}{n+1}\cdot(S_n-S_{k-1}+S_{n+1-k}).$$

So, the expression (5) takes the form

$$E=\frac{1}{n+1}\cdot S_k\cdot(S_n-S_{k-1}+S_{n+1-k}).$$

Maximizing this expression is equivalent to maximizing the same expression but multiplied by $n+1$. So, we arrive at the following conclusion.

Resulting simplified problem. To find the optimal level k, we must maximize the expression

$$M_k\stackrel{\text{def}}{=}S_k\cdot(S_n-S_{k-1}+S_{n+1-k}),\tag{7}$$

where S_i is described by the formula (6).

Examples. When we write on the most basic level, we get $S_1=1$, $S_n\approx 3$ and thus,

$$M_1\approx 6.$$

When we write on the most sophisticated level, we get

$$M_n=S_{20}\cdot\frac{1}{n}\approx 3.0\cdot\frac{1}{20}=0.15.$$

Computations show that the value M_k is the largest for $k=5$, in which case $M_k\approx 8.4$. This effect is 40% higher than when writing on the most primitive Level 1, and more than 50 times higher than writing on the most sophisticated level.

Discussion. Of course, Zipf's law is only approximately true, so the actual optimal level may be $k=4$ or $k=6$. However, in all these cases, we can make the following conclusion.

Conclusion. To achieve the largest possible effect, a research paper must be written on the level $k\approx 5$, crudely speaking corresponding to the middle school. This will drastically increase the effect in comparison with using the most sophisticated level.

Comment. In other words, in an argument between us and folks from the English department, both are wrong: if we want maximal efficiency, we should not use the most primitive level and we should use the most sophisticated level. Instead, we should use an appropriate level in between. A consolation for us is that since this optimal Level 5 is closer to the most primitive Level 1 than to the most sophisticated Level 20, we were kind of closer to the truth :-)

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

References

1. D. Cervantes, O. Kosheleva, and V. Kreinovich, "Why Zipf's law: a symmetry-based explanation", *International Mathematical Forum*, 2018, Vol. 13, No. 6, pp. 255–258.
2. O. Kosheleva and V. Kreinovich, "Zipf's law and 7 ± 2 principle lead to a possible explanation of Daniel's law", *International Mathematical Forum*, 2014, Vol. 9, No. 8, pp. 391–396.
3. O. Kosheleva, V. Kreinovich, and K. Autchariyapanikul, "Commonsense explanations of sparsity, Zipf law, and Nash's bargaining solution", In: Nguyen Ngoc Thach, Doan Thanh Ha, Nguyen Duc Trung, and V. Kreinovich (eds.), *Prediction and Causality in Econometrics and Related Topics*, Springer, Cham, Switzerland, to appear.
4. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, San Francisco, California, 1983.
5. F. Zapata, O. Kosheleva, and V. Kreinovich, "Several years of practice may not be as good as comprehensive training: Zipf's law explains why", *Mathematical Structures and Modeling*, 2020, Vol. 54, pp. 145–148.