To appear in Proceedings of the 2006 IASTED International Conference on ARTIFICIAL INTELLIGENCE AND APPLICATIONS (AIA 2006), February 13 – 16, 2006, Innsbruck, Austria.

INVESTIGATING ESPERANTO'S STATISTICAL PROPORTIONS RELATIVE TO OTHER LANGUAGES USING NEURAL NETWORKS AND ZIPF'S LAW

Bill Manaris[*], Luca Pellicoro[*], George Pothering[*], and Harland Hodges[&]
[*]Computer Science Department, and [&]Management and Marketing Department
College of Charleston
66 George Street, Charleston, SC 29424
USA
manaris@cs.cofc.edu, ldpellic@edisto.cofc.edu, pother@cs.cofc.edu, hodgesh@cofc.edu

ABSTRACT

Esperanto is a constructed natural language, which was intended to be an easy-to-learn lingua franca. Zipf's law models the statistical proportions of various phenomena in human ecology, including natural languages. Given Esperanto's artificial origins, one wonders how "natural" it appears, relative to other natural languages, in the context of Zipf's law. To explore this question, we collected a total of 283 books from six languages: English, French, German, Italian, Spanish, and Esperanto. We applied Zipf-based metrics on our corpus to extract distributions for word, word distance, word bigram, word trigram, and word length for each book. Statistical analyses show that Esperanto's statistical proportions are similar to those of other languages. We then trained artificial neural networks (ANNs) to classify books according to language. The ANNs achieved high accuracy rates (86.3% to 98.6%). Subsequent analysis identified German as having the most unique proportions, followed by Esperanto, Italian, Spanish, English, and French. Analysis of misclassified patterns shows that Esperanto's statistical proportions resemble mostly those of German and Spanish, and least those of French and Italian.

KEY WORDS

Natural language processing, artificial neural networks, classification, Zipf's law

1. Introduction

Esperanto is a constructed natural language which was developed by Ludovic Zamenhof around 1887 and intended as a simple, easy-to-learn alternative to other natural languages [1, 2]. Given Esperanto's artificial origins and grammatical regularity, one wonders how "natural" this language feels to its speakers, compared to other languages that evolved naturally over thousands of years. In linguistic terms, one wonders how Esperanto's imposed structure and simplicity affects its statistical proportions relative to other natural languages, such as English, French, German, and Spanish.

To explore this question, we collected a corpus of 283 books in electronic form from Project Gutenberg and elsewhere [3, 4]. These books spanned six languages: English (97), French (45), German (35), Italian (35), Spanish (38), and Esperanto (34). We then extracted various statistical proportion measures using Zipf-based metrics. Finally, we used these measures (features) to train artificial neural networks (ANNs) to carry out several classification tasks comparing Esperanto to other natural languages.

The paper spans three areas of scientific inquiry: natural language processing, fractals (Zipf's law), and ANN-based classification. Section 2 provides an overview of Zipf's Law and how it may be used to model the statistical proportions (scaling properties) of natural languages. Section 3 provides an overview of related research. Section 4 presents the experimental methodology. Section 5 provides a brief overview of the data and presents the various classification experiments conducted. Section 6 interprets these results relative to our hypothesis. The last section offers concluding remarks and identifies future research directions.

2. Zipf's Law

George Kingsley Zipf (1902-1950) was a linguistics professor at Harvard, who studied results from various fields demonstrating an intriguing relationship (or statistical proportion) found in many natural phenomena.

Zipf's law models the statistical proportions (scaling properties) of many phenomena in human ecology, including natural language and music [5, 6]. Zipf's law is one of many related laws that describe scaling properties of phenomena studied in the physical, biological, and behavioral sciences. These include Pareto's law, Lotka's law, power laws, Benford's law, Bradford's law, Heaps' law, etc. [7].

Zipf distributions (also known as *1/f* and *pink noise* distributions) have been discovered in a wide range of human and naturally occurring phenomena including city sizes, incomes, subroutine calls, earthquake magnitudes, thickness of sediment depositions, extinctions of species, traffic jams, and visits to websites [8, 9, 10].

Informally, Zipf's law categorizes phenomena in which certain types of events are quite frequent, whereas other types events are rare. For example, in English, short words (e.g., "the") are very frequent, whereas long words (e.g., "anthropomorphologically") are quite rare.

Comparing a word's frequency of occurrence with its statistical rank, Zipf noticed an inverse relationship: successive word counts are roughly proportional to 1, 1/2, 1/3, 1/4, and so on. This is captured by the formula:

$$P(f) \sim 1/f^{n} \qquad (1)$$

where $P(f)$ denotes the probability of a word (or event) of rank f and n is near 1 [5]. Plotting the word counts (frequencies) against their statistical rank on log scale produces a near straight line (see Fig. 1). This line is characterized by two real numbers: the *slope* of the trendline, and R^2, i.e., the proportion of y-variability of data points with respect to this trendline. The slope is the same as exponent $-n$ in (1). This plot is known as *rank-frequency* distribution.

In theory, this slope may range from 0 to –∞. A slope near 0 indicates a uniform distribution (white noise). A slope near –∞ indicates a monotonous phenomenon (i.e., a "book" consisting mostly of one word). It has been suggested that a slope near –1.0, corresponds to a proportion that feels balanced to humans, for certain phenomena, such as music, urban structures, and images [6, 10, 12, 13].

Zipf's main contribution was that (a) he was the first to hypothesize that there is a universal principle connecting the scaling results reported in various disciplines, and (b) he proposed a mathematical formula to describe it. Although his attempts to derive a comprehensive theory were incomplete (and some say misguided), his mathematical formula is pretty accurate. Zipf's work had considerable influence on a young graduate student named Benoit Mandelbrot, who went on to develop the field of fractal geometry [11].

Mandelbrot generalized Zipf's law to account for all types of fractal phenomena in nature, as:

$$P(f) \sim 1/bf^{n} \qquad (2)$$

where b is a real constant.

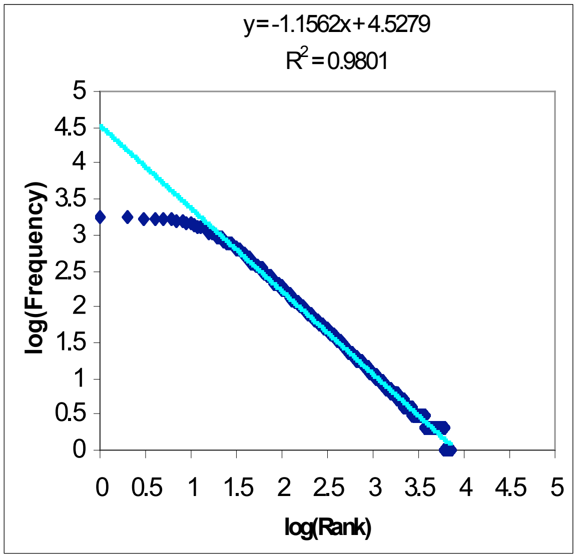

Fig 1. Rank-frequency log-scale plot of the word distribution in the English version of *Robinson Crusoe* – a near-Zipfian distribution (slope is –1.16, R^2 is 0.98).

2.1. Statistical Proportions of Natural Languages

Word rank-frequency distribution slopes vary across languages. For instance, Zipf reports that, whereas English exhibits a word-distribution slope near –1.0, other languages may exhibit different slopes.[1] In particular, he points out that Palestinian Hebrew, among others, exhibits a more chaotic word slope (> –1.0). Palestinian Hebrew is of particular interest, because (similar to Esperanto) it is an "artificially constructed" language, but with a Semitic base. (It is different from Yiddish, which has a Germanic base) [5, p. 129]. Zipf believed that the artificiality of the language might reveal itself in its statistical proportions.

Accordingly, Gelbukh and Sidorov, using a corpus of 78 books of various genres, report a mean word-distribution slope of –0.9738 for English (*std* 0.0190), and –0.8929 for Russian (*std* 0.0227) [14].

Our statistical analyses indicate that word-distribution slopes are also correlated with text length, i.e., shorter texts exhibit slopes closer to zero (i.e., more chaotic distributions) than longer texts. Also, during our corpus selection, we discovered that word distribution slopes also depend on genre (e.g., poem, news article, play, book). It is possible that these two observations are interrelated – poems are usually shorter than news articles, news articles are usually sorter than plays, and so on.

3. Related Research

To the best of our knowledge, Zipf's law has never been used for classification of natural languages. However, it has been used successfully for classification in various other domains.

Burgos and Moreno-Tovar use Zipf's law to differentiate among immune systems of normal, irradiated chimeric, and athymic mice [15]. Kalda *et al.* use it to distinguish healthy from non-healthy heartbeats in humans [16]. Li and Yang use it to distinguish cancerous human tissue from normal tissue using microarray gene data [17]. Taylor *et al.* use a derivative technique to authenticate and date paintings by Jackson Pollock [18]. Machado, *et al.* use Zipf-based metrics and ANNs to classify music pieces according to composer [19]. Finally, Manaris *et al.* use Zipf-based metrics and ANNs to predict human aesthetic responses based on statistical proportions of music pieces [20].

[1] It should be noted that Zipf and his students had to calculate word frequencies and corresponding slopes by hand, since computers were not yet available.

Language	Word		Word Distance		Word Bigram		Word Trigram		Word Length	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Esperanto	-0.9204	0.1239	-0.9446	0.0375	-0.4008	0.0902	-0.1484	0.0623	-2.7943	0.3345
English	-1.1858	0.1018	-1.0169	0.0387	-0.5542	0.0636	-0.2136	0.0411	-3.0040	0.2694
French	-1.0448	0.1081	-0.9738	0.0355	-0.5070	0.0765	-0.2036	0.0568	-2.7428	0.2831
German	-0.9745	0.0974	-0.9709	0.0333	-0.4024	0.0657	-0.1089	0.0400	-3.0679	0.2421
Italian	-0.9947	0.1234	-0.9584	0.0405	-0.4255	0.0885	-0.1317	0.0498	-2.9147	0.4372
Spanish	-0.9255	0.0983	-0.9193	0.0489	-0.4363	0.0787	-0.1607	0.0673	-2.7744	0.3240

Table 1. Mean and standard deviation of each metric slope for all languages.

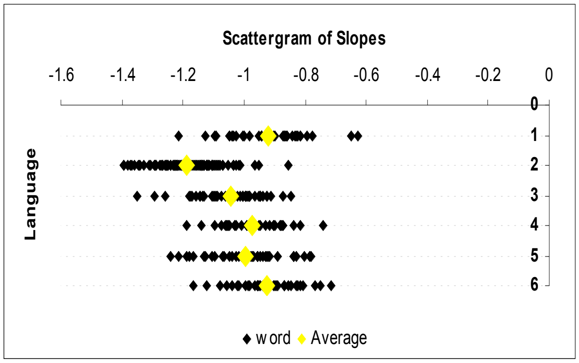

Fig 2. Plot of word distribution slopes. Esperanto is 1, English is 2, French is 3, German is 4, Italian is 5, and Spanish is 6.

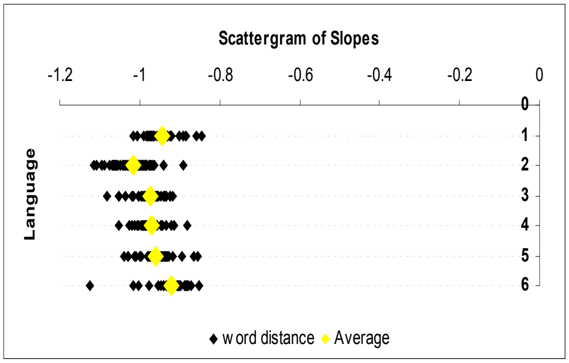

Fig 3. Plot of word-distance distribution slopes. Esperanto is 1, English is 2, French is 3, German is 4, Italian is 5, and Spanish is 6.

4. Methodology

We collected a corpus of 284 books in six languages: English (97), French (45), German (35), Italian (35), and Spanish (38), and Esperanto (34). These books came from Project Gutenberg and other on-line sources of e-texts [3, 4]. Given our observation that distribution slopes depend on genre (e.g., poem, news article, play, book), we collected only novels. The book sizes ranged from 2419 to 338903 words.

4.1. Zipf-based Metrics

In order to measure the statistical proportions of these languages, each book was measured using the following metrics:

- **Word distribution:** this metric counts occurrences (frequencies) of words and plots them against their statistical rank. This is one of Zipf's original metrics [5].
- **Word-distance distribution:** this metric counts occurrences (frequencies) of distances between word repetitions and plots them against their statistical rank. This is one of Zipf's original metrics [5].
- **Word-bigram distribution:** this metric counts occurrences of word bigrams and plots them against their statistical rank.
- **Word-trigram distribution:** this metric counts occurrences of word trigrams and plots them against their statistical rank.
- **Word-lengh distribution:** this metric counts occurrences of word lengths and plots them against their statistical rank.

This generated a feature vector of 11 elements per book: total number of words, and pairs of slope and R^2 values for each of the above metrics.

We also explored other metrics, such as character distribution, characters-per-sentence distribution, and words-per-sentence distribution. However, these metrics did not generate power-law distributions, and thus were excluded from the classification experiments.

Table 1 shows the mean and standard deviation of all metrics for each language in our corpus.

5. Classification Experiments

Here we describe how we used these feature vectors to train ANNs to carry out several classification tasks comparing Esperanto to other natural languages.

5.1. Statistical Analysis

First, we analyzed the data for statistical relationships. Scatter diagrams of the data reveal that the statistical proportions of each language overlap significantly. For example, figures 2 and 3 show the spread of slope values for word and word-distance distributions, respectively. Clearly, from this perspective, Esperanto does not stand out as having unique statistical proportions. Subsequent analyses of variance (ANOVA) verify the statistical overlap of the six languages. Figures 4 and 5 show the

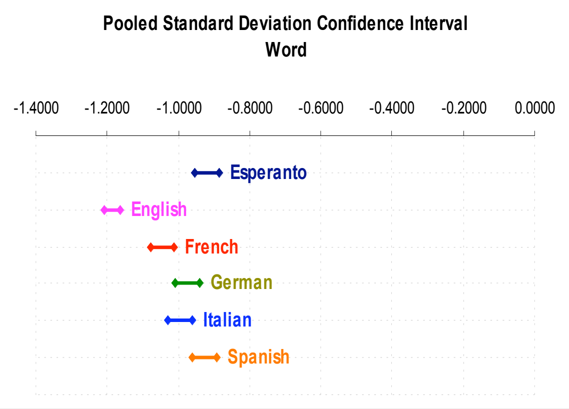

Fig 3. Pooled standard deviation 95% confidence levels for word distribution.

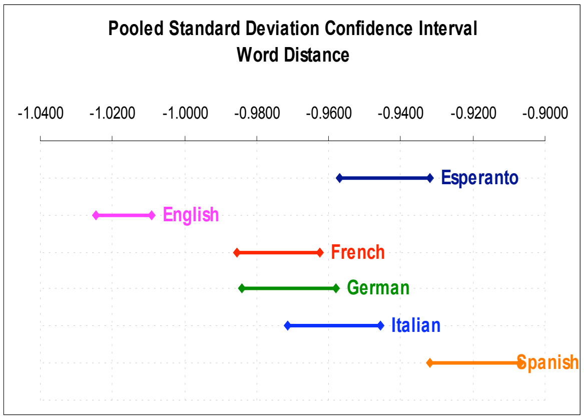

Fig 4. Pooled standard deviation 95% confidence levels for word-distance distribution.

relationships for the pooled standard deviation 95% confidence levels for word and word-distance slope values, respectively. In particular given our corpus,

- Esperanto is statistically equivalent to Spanish in word slope;
- Esperanto is statistically equivalent to Italian in word-distance slope;
- Esperanto is statistically equivalent to German and Italian in word-bigram slope;
- Esperanto is statistically equivalent to Italian, and Spanish in word-trigram slope; and
- Esperanto is statistically equivalent to French, Italian, and Spanish in word-length slope.

In other words, looking only at one feature at a time, Esperanto resembles all other languages in our corpus, except English.

But what if we examine all the features together? To answer this question, we conducted six ANN binary classification experiments (to see how well the features can distinguish each language from the rest combined), a multi-classification experiment (to see how well the features can distinguish all the languages at the same time), and a metric filtering assessment (to discern the relevance of each feature for language classification). These experiments were carried out using the Weka data-mining environment [21].

5.2. Binary Classification Tasks

This section describes the six binary classification experiments we conducted. The first experiment tested Esperanto against all other languages; the second, English against all other languages; and so on.

For these experiments we used a feed-forward ANN trained via back-propagation. The ANN architecture consisted of three layers (see figure 5). The input layer contained 11 nodes, one for each of the features introduced in section 4; the hidden layer contained 6 nodes; and the output layer contained 2 nodes, one for the language of interest, and one for the other five languages.

We conducted six 10-fold cross-validation experiments. Each experiment tested the ability to classify one language against the rest. Each ANN was trained for 10000 cycles, with a learning rate of 2.0 and momentum of 1.0.

The 10-fold cross-validation nature of the study divided the set of feature vectors (each vector representing one book) into 10 subsets of approximately equal size. Each ANN was then trained and tested 10 times, wherein each time one of the subsets was used for testing and the remaining nine subsets were used for training. The average of these 10 trials gave the results for one of the six classification experiments. Table 2 summarizes the results for all six languages. The following definitions are used:

- *Success %* = number of correctly classified books / number of all books * 100.
- *RMS* = root mean square error of actual ANN outputs compared to true outputs.
- *TP (True Positive)* = number of books classified as language X / true number of books in language X.
- *FP (False Positive)* = number of books classified as language X / number of books not belonging to language X.
- *Precision* = number of correctly classified books of language X / number of books classified as belonging to language X.
- *Recall* = number of correctly classified books of language X / number of books in language X.
- *F-Measure* = (2 * Precision * Recall) / (Precision + Recall); this combines precision and recall into one measure.

As shown in Table 2, the binary-classification experiments yielded an average success rate of 95.3%.

This means the ANN was able to differentiate each language from all others combined with high accuracy. Most significant for our purposes, however, is that while the success rate of the ANN is high in distinguishing Esperanto from the other languages based on the Zipf-based metrics being used, in fact Esperanto fares neither better nor worse than the other natural languages. In other words, no language emerges as unique in any pronounced way.

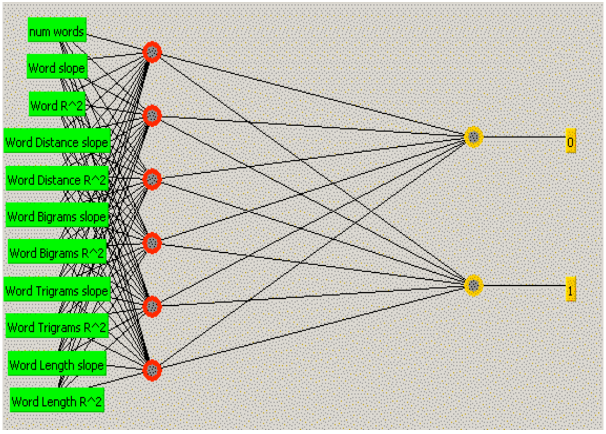

Fig 5. ANN architecture for binary-classification tasks.

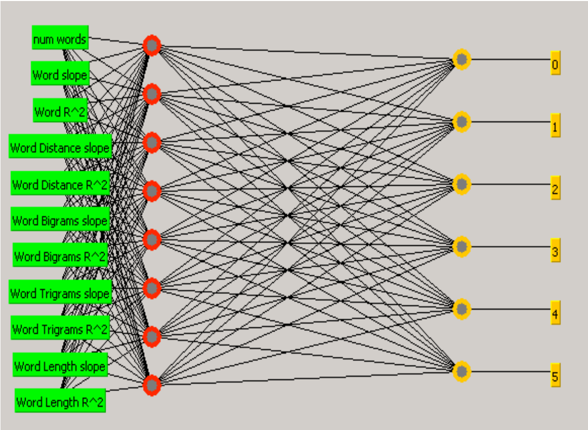

Fig 6. ANN architecture for multi-classification task.

5.3. Multi-classification Task

Now we describe the multi-classification experiments conducted to see how well the Zipf-based features might help classify all the languages simultaneously. For this experiment, we modified the ANN architecture to have six output nodes rather than two – one for each of the six languages (see figure 6). We also reduced the sample of size of English from 97 to 48, so that it was similar to that of other languages.

We conducted five different, n-fold cross-validation ANN experiments. For each of the five runs we changed some of the parameters. For example, we varied the number of folds in the cross validation from 10 to 18; and/or we changed the random seed used to generate the data sets for different folds. Each ANN was trained for 500 cycles, with a learning rate of 3.0 and momentum of 2.0.

Table 3 summarizes the accuracy achieved by the ANN in each of the five runs. Table 4 summarizes the accuracy achieved by the ANN for each language across all five runs. Overall, the multi-classification experiments yielded an average success rate of 87.3%.

Of the six languages in our corpus, German exhibited the most unique signature (98.59%), with Esperanto second (97.18%), and French last (91.54%). Once again it is important to note that there is nothing in this data that makes Esperanto stand out.

5.4. Relevance of Metrics

We evaluated the relevance of each of our metrics for the classification tasks performed. To do this, we used the Correlation-based Feature Subset Selection algorithm [22]. This algorithm considers the individual predictive ability of each feature, along with the degree of redundancy relative to other features. It "identifies and screens irrelevant, redundant, and noisy features." It also "identifies relevant features as long as their relevance does not strongly depend on other features." [22, p. 3].

Table 5 shows the identified subset of features. The relevance value provides a measure of the corresponding feature's significance in classification; higher percentage values are more significant.

Based on this result, we repeated the mutli-classification ANN experiment using a reduced feature set. This set consisted of the five features identified as most relevant in Table 5, namely Word R^2, Word Distance slope, Word Distance R^2, Word Trigrams slope, and Word Length R^2. This was a 10-fold cross-validation run, with training of 500 cycles, learning rate of 3.0 and momentum of 2.0.

The ANN achieved an overall success rate of 81.97%. This validates the above result. It also demonstrates that ANNs not only can handle, but usually can benefit from redundancies in the feature set. Table 6 shows the ANN accuracy statistics per language using this reduced feature set.

Finally, analysis of misclassified patterns within the ANN confusion matrices indicates that the combined statistical proportions of Esperanto resemble mostly those of German (59.1%) and Spanish (27.3%), followed by English (11.4%) and Italian (2.3%). In other words, 59% of all classification errors where between German and Esperanto books being misclassified for each other.

6. Conclusion

Our hypothesis was that Esperanto's imposed structure and simplicity would be reflected in its statistical proportions, relative to other natural languages.

Confidence intervals analysis shows that Esperanto has proportions similar to one or more other languages from our corpus, when examining a single proportion at a time. As mentioned in section 5.1, English is the only language that is statistically different from Esperanto, across all measured individual proportions.

Various ANN classification experiments demonstrate that, when combining all proportions (features) together, each language in our corpus exhibits its own, unique combination of features – its own "signature". This signature allows ANNs to achieve high success rates (averaging 87.3% for the multi-language classification task, and 95.3% for the binary classification task).

Language	*Success %*	*RMS*	*TP Rate*	*FP Rate*	*Precision*	*Recall*	*F-Measure*
Esperanto	97.18%	0.1602	0.912	0.020	0.861	0.912	0.886
English	94.36%	0.2231	0.887	0.027	0.945	0.887	0.915
French	91.54%	0.2520	0.800	0.063	0.706	0.800	0.750
German	98.59%	0.1267	0.914	0.004	0.970	0.914	0.941
Italian	95.42%	0.2078	0.829	0.028	0.806	0.829	0.817
Spanish	94.71%	0.2178	0.763	0.024	0.829	0.763	0.795
Average	*95.30%*	*0.1979*	*0.851*	*0.028*	*0.853*	*0.851*	*0.851*
Std	*0.0223*	*0.0419*	*0.058*	*0.018*	*0.0881*	*0.058*	*0.068*

Table 2. Combined accuracy statistics of the six ANN binary-classification experiments.

Run	*Success %*	*RMS*
1	86.70	0.1869
2	87.55	0.1838
3	87.98	0.1888
4	87.98	0.1781
5	86.27	0.1897
Avg	*87.30*	*0.1855*
Std	*0.697*	*0.0042*

Table 3. Accuracy statistics for each ANN multi-classification experiment.

Language	*TP Rate*	*FP Rate*	*Precision*	*Recall*	*F-Measure*
Esperanto	0.912	0.020	0.861	0.912	0.886
English	0.887	0.027	0.945	0.887	0.915
French	0.800	0.063	0.706	0.800	0.750
German	0.914	0.004	0.970	0.914	0.941
Italian	0.829	0.028	0.806	0.829	0.817
Spanish	0.763	0.024	0.829	0.763	0.795
Average	*0.851*	*0.028*	*0.853*	*0.851*	*0.851*
Std	*0.058*	*0.018*	*0.0881*	*0.058*	*0.068*

Table 4. Detailed accuracy by language for all ANN multi-classification experiments combined.

In other words, Esperanto can be identified easily from the other languages based on the Zipf-based metrics used, but the same holds for the other languages in our corpus. The following two interpretations emerge:

(a) Esperanto, in spite of its short, 120-year lifecycle, has evolved enough to exhibit "natural" statistical proportions.

(b) Esperanto is "artificial", but our metrics cannot differentiate between natural and artificial languages.

We carried out a quick, follow-up experiment with 49 C++ programs, used as a control group. The results demonstrated that our metrics can differentiate between natural and artificial languages. Specifically, the C++ "texts" differed substantially from the rest of the corpus, in terms of, at least, word-trigram slopes. In particular, *average* for C++ was –0.36265 (*std* 0.1100), whereas for the other languages was –0.1651 (*std* 0.0643).

Esperanto was envisioned as a means for facilitating peaceful coexistence of different cultures and nations. Consequently, it integrates characteristics from many other natural languages. It is likely that Zamenhof's linguistic background may have affected Esperanto, not only at the surface level (i.e., morphology, vocabulary, and grammar), but also at a deeper level – in the proportions, the flow, and balance of the language.

It is worthwhile to note that Zamenhof's native languages were Russian and Yiddish. Additionally, he spoke fluent Polish and German. Later, he learned French, Latin, Greek, Hebrew and English. He also explored Italian, Spanish and Lithuanian [1].

As mentioned in section 5.4, analysis of misclassified patterns indicates that Esperanto's statistical proportions resemble mostly those of German (59.1%) and Spanish (27.3%), followed by English (11.4%) and Italian (2.3%). It would be interesting to repeat our experiments extending our corpus with Russian and Polish texts, as well as "texts" from additional artificial languages.

In closing, our results demonstrate that Esperanto exhibits "natural" statistical proportions, similar to those of other European languages. Also, our results demonstrate that Zipf's law and related metrics are effective for classification of natural languages. This is not surprising, given how effective Zipf's law is for classification in other domains.

Acknowledgements

The authors would like to acknowledge Brian Murtagh, Hector Mojica, John Emerson, Jason Trinklein, and Charles Shapiro for contributing to corpus collection and programming for the experiments.

References

1. Wikipedia (2005). "L. L. Zamenhof," accessed Sep. 28, 2005. [http://en.wikipedia.org/wiki/Zamenhof].
2. Boulton, M. (1960). *Zamenhof, Creator of Esperanto*. London: Routledge, Kegan & Paul.
3. Project Gutenberg (2005), accessed Sep. 25, 2005. [http://www.gutenberg.org].
4. Harlow, D. (2005). Literaturo, en la reto, en Esperanto, accessed Sep. 28, 2005. [http://donh.best.vwh.net/Esperanto/Literaturo/literaturo.html].
5. Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort*, New York: Addison-Wesley Press.

Relevance %	*Feature Name*
0%	Number of words
10%	Word slope
40%	Word R^2
50%	Word Distance slope
70%	Word Distance R^2
0%	Word Bigram slope
0%	Word Bigram R^2
30%	Word Trigram slope
0%	Word Trigram R^2
0%	Word Length slope
10%	Word Length R^2

Table 5. Predictive ability of features relative to language.

Language	*TP Rate*	*FP Rate*	*Precision*	*Recall*	*F-Measure*
Esperanto	0.758	0.030	0.806	0.758	0.781
English	0.875	0.054	0.808	0.875	0.840
French	0.867	0.043	0.830	0.867	0.848
German	0.857	0.035	0.811	0.857	0.833
Italian	0.676	0.030	0.793	0.676	0.730
Spanish	0.842	0.026	0.865	0.842	0.853
Average	0.813	0.036	0.819	0.813	0.814
Std	0.072	0.010	0.023	0.072	0.044

Table 6. Detailed accuracy by language for ANN multi-classification experiment using reduced feature set.

6. Voss, R.F., and Clarke, J. (1975). "1/f Noise in Music and Speech", *Nature*, 258: 317-318.
7. Li, W. (2005). Zipf's Law, accessed online Sep. 28, 2005. [http://www.nslij-genetics.org/wli/zipf/].
8. Adamic, L. A. and B. A. Huberman. 2000. "The Nature of Markets in the World Wide Web." *Quarterly Journal of Electronic Commerce*, 1(1): 5–12.
9. Bak, P., C. Tang and K. Wiesenfeld. 1987. "Self-organized Criticality: An Explanation for 1/f Noise." *Physical Review Letters*, 59: 381-384.
10. Schroeder., M. (1991). Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise, New York: W. H. Freeman and Company.
11. Mandelbrot, B. (1977). Fractal Geometry of Nature, New York: W.H. Freeman and Company.
12. Salingaros, N.A., and B.J. West. (1999). "A Universal Rule for the Distribution of Sizes", *Environment and Planning B: Planning and Design*, 26: 909-923.
13. Spehar, B., Clifford, C.W.G. , Newell, B.R., and Taylor, R.P. (2003). "Universal Aesthetic of Fractals." *Computers & Graphics*, 27: 813-820.
14. Alexander Gelbukh, Grigori Sidorov. *Zipf and Heaps Laws' Coefficients Depend on Language*. Proc. CICLing-2001, Conference on Intelligent Text Processing and Computational Linguistics, February 18–24, 2001, Mexico City. *Lecture Notes in Computer Science N 2004*, ISSN 0302-9743, ISBN 3-540-41687-0, Springer-Verlag, pp. 332–335.
15. Burgos, J.D. and P. Moreno-Tovar (1996). "Zipf-scaling Behavior in the Immune System." *Biosystems*, 39(3): 227-232.
16. Kalda, J., M. Sakki, M. Vainu, and M. Laan (2001). "Zipf's Law in Human Heartbeat Dynamics." Available online [http://arxiv.org e-print , physics/0110075.
17. Li, W. and Y. Yang (2002). "Zipf's Law in Importance of Genes for Cancer Classification using Microarray Data." *Journal of Theoretical Biology*, 219: 539-551.
18. Taylor, R.P. A.P. Micolich, and D. Jonas (1999). "Fractal Analysis Of Pollock's Drip Paintings." *Nature*, 399: 422.
19. Machado, P., Romero, J., Santos, M.L., Cardoso, A., and Manaris, B. (2004). "Adaptive Critics for Evolutionary Artists," EvoMUSART2004 – 2nd European Workshop on Evolutionary Music and Art, Coimbra, Portugal, *Lecture Notes in Computer Science, Applications of Evolutionary Computing, LNCS 3005*, Springer-Verlag, pp. 437-446.
20. Manaris, B., Romero, J., Machado, P., Krehbiel, D., Hirzel, T., Pharr, W., and Davis, R.B. (2005). "Zipf's Law, Music Classification and Aesthetics." *Computer Music Journal* 29(1): 55-69, MIT Press.
21. Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*, 2nd ed., Morgan Kaufmann, San Francisco.
22. 1. Hall, M. A. (1998). Correlation-based Feature Subset Selection for Machine Learning. Ph.D. Thesis, University of Waikato.