Lexical Semantic Richness in Poe's Essays and Short Stories: Comparing Corpora with Word Smith Tools and Range

Edgar Bernad-Mechó

Department of English Studies

Universitat Jaume I

Castelló de la Plana, Spain

ebernad@uji.es

Abstract

Edgar Allan Poe's Essays and Short Stories have been widely analyzed throughout the decades. Previous research confirms an ample use of varied vocabulary in his short stories. Nevertheless, little emphasis has been put on some of his not-so-famous works: his essays. Thus, the main aim of this paper is twofold: on the one hand, we aim at comparing the lexical semantic richness in Poe's essays and in his short stories; on the other hand, we intend to test the effectiveness of two different analytical tools to check this semantic variation, i.e. WordSmith Tools and Range. In order to achieve these aims, three short stories and two essays by Poe were selected and combined to create two main corpora: one of short stories and one of essays. After separating the corpora intro fragments of 2000 tokens, lexical semantic richness was assessed using the two aforementioned tools. Results show that i) lexical semantic richness is higher in short stories than it is in essays, and ii) both tools have proven to be effective. These results are further discussed and pedagogical applications for language teaching are put forward.

Keywords: Lexical Semantic Richness; Edgar Allan Poe; Word Smith Tools; Range; Lexical Frequency Profile.

I.Introduction

Lexical semantic richness is a concept which is difficult to be defined. It may include a series of characteristics such as lexical variation, sophistication, and density or originality, among others. In short, the aim of lexical semantic richness is to measure to some extent the vocabulary size and use that a speaker makes of their lexicon. This becomes a relevant parameter since it allows researchers to infer the level of proficiency of a speaker (either L1 or L2/FL) as far as vocabulary is concerned. Measures of lexical richness attempt to quantify the degree to which a writer is using a varied and large vocabulary (Laufer & Nation 1995). Measuring lexical semantic richness is quite a difficult task, though. The high number of variables and the difficulty of considering them all individually make the statement of the lexical semantic richness in just one item a very hard undertaking.

The main aim of this paper is to test the effectiveness of two different tools when analyzing lexical semantic richness of texts. The two tools that will be taken into account in this paper are *WordSmith Tools* (Scott, 1998) and *Range*, developed by Nation (1995), which includes the 34.000 more frequent words in the British National Corpus (BNC). WordSmith Tools and Range are two tools commonly used by linguists when conducting any kind of quantitative analysis. Berber-Sardinha (2000), for instance,employs WordSmith to analyze small corpora. He affirms that WordSmith and KeyWords provide facilities for comparing a study corpus to a reference corpus. On the other hand, Range has been developed by Paul Nation and some studies of lexical richness in L2 using it have been conducted (see Laufer & Nation 1995). Moreover, the 34.000 most common words in English according to the British National Corpus were organized in bands of 1000 words by Nation. Furhtermore, Range, unlike WordSmith, is available to be downloaded for free in his web-site and offers an analysis of the Lexical Frequency Profile (LFP) of works.

To test the effectiveness of these two tools, a comparison of the lexical semantic richness between Edgar Allan Poe's short stories and essays will be made. The initial hypothesis is a null one; being the same author, no significant differences are expected to be found when modifying the literary genre.

II. Methodology

Sample

The sample used for this study will be made up of three short stories and two essays written by Edgar Allan Poe. The three short stories are *The Fall of the House of Usher*, *The Cask of Amontillado*, and *The Masque of the Red Death*; they have 7161, 2344 and 2423 words respectively. The two essays are *Eureka* and *The Philosophy of Composition*, having 38053 and 4609 words. All these writings have been selected out of the ten last years of Poe's career, disregarding his first works. In this manner, a more homogenous writing style is expected.

In addition, the short stories have been added together and then separated into fragments of 2000 words approximately in order to analyze the rate at which new word-types are inserted into Poe's vocabulary. Also, seven fragments of 2000 words each have been extracted out of the two essays randomly to create a sample that had a comparable size to the corpus of short stories. In this sense, it was important to divide the works into fragments containing the same number of words so that the comparison could be accurate.

Analysis

In order to analyze the lexical semantic richness in Poe's works, a series of variables were taken into account. The most important one was the lexical semantic variation or type-token ratio of the texts, that is, the number of different words in a text divided by the total number of words in that text. This variable provides us with a good image of what the range of Poe's vocabulary consists of.

This lexical semantic variation was analyzed using WordSmith Tools. First, we analyzed the type/token ratio in each of the fragments of both genres and each of the literary works as a whole. Then, we dealt with the progression and the rate at which new types are inserted into the works; to do so, we progressively added new fragments of each genre and noted down the results for each addition. For instance, we analyzed the lexical variation of Short Stories-Fragment 1 and then Short Stories-Fragment 1 + 2, Short Stories-Fragment 1 + 2 + 3, and on. The lexical semantic variation was measured by means of the formula types/tokens*100, that is, number of types divided number of tokens and then multiplied by 100 to obtain a percentage.

Finally, we analyzed the lexical semantic frequency profile of these works with the software Range. This allowed us to look at the proportion of frequency of words used by Poe compared to the British National Corpus. This corpus is organized into the 1000 most frequent words in English, the 2000 most frequent words, 3000, and all the way to the 34.000 most common words. However, for our purposes, we employed a simpler version of Range which only includes the 3.000 most common words.

Range divides the types used by an author into ranges of frequency. With this tool we were able to check whether Poe uses a high or low percentage of 'uncommon' words in English. It is important to note that proper nouns as well as numbers or symbols were previously removed from the fragments in order to obtain objective results.

By considering these two variables together, an objective overview of Poe's lexical semantic richness was obtained. Results are presented and discussed in the next section.

III. Results and Discussion

After analyzing Poe's full corpus with WordSmith Tools, we found that the shorter the work is the higher the lexical variation is (see figure 1).

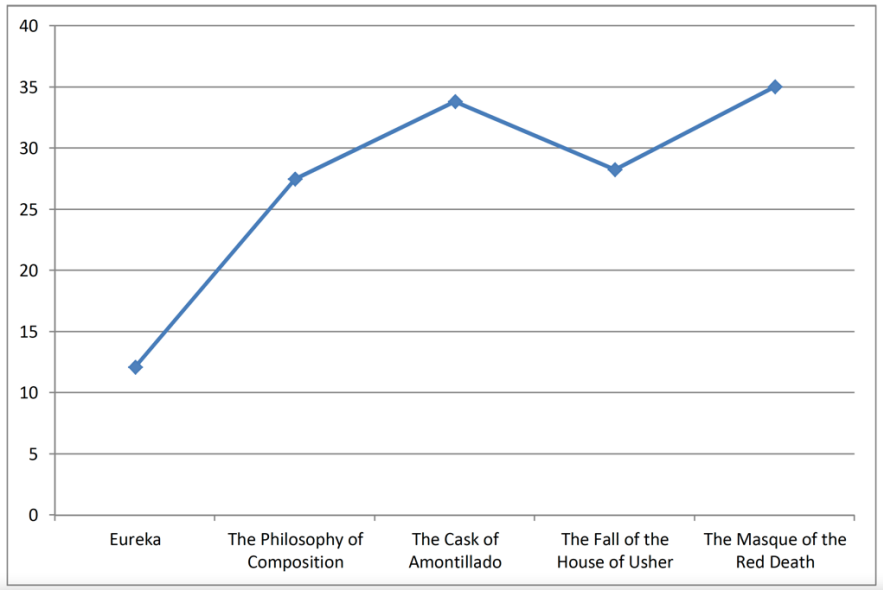

Figure 1 – Ratio type/token for full works

As we can see, the highest points in the table correspond to *The Masque of the Red Death* and *The Cask of Amontillado*; these are the shorter texts (2344 and 2423 tokens respectively). The lowest type/token ratio corresponds to *Eureka*,which is the longest work (38053 tokens). This proves the existence of a point in the writers' works where the input of new types is rarer. Still, the analysis of works of different lengths seems to show different results. This is evidenced in the second part of the analysis in which we looked at different fragments of 2000 tokens each (see figure 2).

This figure shows the analysis of six fragments extracted from the short stories, of 2000 words each, and seven fragments of 2000 words extracted from the essays. An average number of types as well as the ratio for 2000 tokens have been worked out.

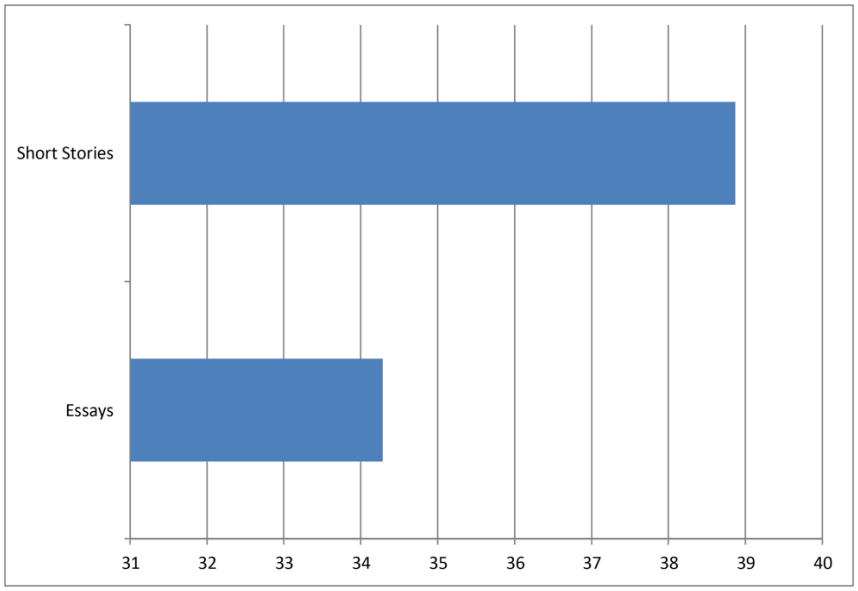

Figure 2 – Average ratio type/token for short stories and essays

This analysis shows that the ratio type/token for short stories is higher in the case of short stories, that is, more types are present in Poe's short stories per 2000 words. This could possibly be due to the higher literary focus of short stories. Whereas essays can be considered as semi-plain language discussions on a given topic, short stories require a series of literary techniques and semantic variation in order to create a proper atmosphere. This could even be higher in the case of American Romanticism where very dark and obscure environments were portrayed. This claim, however, should be tested out by comparing works belonging to the American Romanticism with works from other periods.

When considering individual works separately, it is interesting to mention that the ratio type/token of *The Fall of the House of Usher* (41,03) was higher than the rest of the short stories (36,1) –whose ratio was already higher than the ratio for essays (34,28). *The Fall of the House of Usher* is probably one of Poe's darkest works and it is full of very detailed descriptions which imply a high mastery of the language. This fact could explain the slight difference in the number of types in this work.

In our next step, we analyzed the progression of the number of types as we kept adding fragments of 2000 tokens to each genre. We observe that the short stories are prone to the

creation of more types than the essays (see figure 3). As stated before, this could be explained if we understand essays as texts where few topics are discussed while short stories include conversations, descriptions and different situations which require a higher number of types.

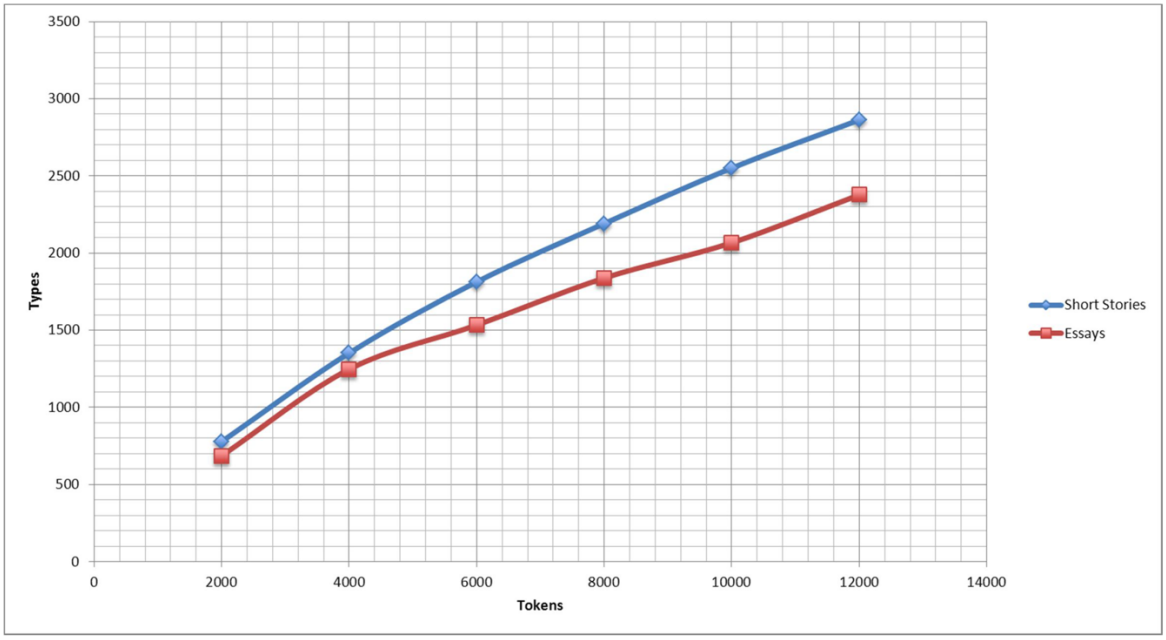

Figure 3 – Type/token progression

Finally, the last aspect that we analyzed is the LFP of both short stories and essays (see tables 4 and 5). In this case, we processed two samples of approximately 12.000 tokens each (one of a compendium of the short stories and the other including fragments of the two essays) with the program Range developed by Nations (2005). The results of this processing show the following data:

In both, the short stories and the essays there is a considerable amount of tokens and types which is off the lists, that is, words which are not among the 3.000 most frequent words in English. In the case of the short stories, these words make up the biggest group of types (42% of the types belong to this category). In the case of the essays this group is the second biggest group after the most common words (those included among the 1.000 most common

words in English). Once again, the presence of a higher elaborated language is generally higher in the short stories.

Even though uncommon words play an important role in the case of types, their presence in the text (tokens) is way lower; only 15% in the short stories and 12% in the essays. In any case, the presence of such a high number of 'rare' types in Poe's texts is a good sign of Poe's lexical richness.

WORLD LIST	TOKENS/%	TYPES/%	FAMILIES
One	8680/73.52	985/33.06	604
Two	775/6.56	435/14.60	310
Three	558/4.73	307/10.31	2336
Not in thelists	1794/15.19	1252/42.03	?????
TOTAL	11807	2979	1150

Table 1 – LFP of Poe's short stories

WORLD LIST	TOKENS/%	TYPES/%	FAMILIES
One	9319/78.14	952/39.39	548
Two	878/7.36	413/17.09	257
Three	291/2.44	167/6.91	128
Not in thelists	1438/12.06	885/36.62	?????
TOTAL	11926	2417	933

Table2 – LFP of Poe's essays

IV.Conclusions

In this paper we have analyzed a series of texts by Edgar Allan Poe: three short stories and two essays. The aim of the project was to determine the differences in lexical semantic richness in Poe's different genres, as well as the validity of two tools commonly used among linguists (WordSmith Tools and Range).

We have seen that these two tools can be useful when determining an author's lexical semantic richness since they provide us with many data such as type/token ratios, and frequency of words. Also, with Range, this frequency is compared against the British National Corpus and the data are organized in frequency bands.

All in all, we can say that Edgar Allan Poe was a writer with a high lexical semantic richness given the fact that he there is an elevated presence of uncommon words among his writings. The initial null hypothesis presented in this paper has been rejected; the lexical semantic richness in short stories is slightly higher than in essays. This is probably due to the fact that short stories require a higher amount of literary vocabulary than essays, which are mere reflections on a certain matter.

Nevertheless, further studies comparing these data to data obtained from other authors are necessary to establish a more exact degree of lexical semantic richness. An analysis of other genres such as poetry might also be a good starting point to compare. Also, it would be interesting for further research to analyze the use of the less frequent words and hapax legomena in Poe's writings in order to obtain a better image of the author's lexical semantic richness; that is, analyzing his lexical semantic competence.

These two tools have been proved to be effective when measuring lexical richness. Analysis of L2/FL lexical richness could be conducted following the methodology presented in this paper. These analyses would show how the process of lexical enrichment takes place in learners and which words in the frequency bands still need to be learned. In this sense, the collection of real samples of learners' productions would offer clear direction towards the needs for vocabulary teaching in L2/EFL.

References

LAUFER, B. & NATION, P. 1995. Vocabulary Size and Use: Lexical Richness in L2 Written Production, in *Applied Linguistics*, Vol. 16, No. 3. Oxford, England: Oxford University Press.

SCOOT, M. 1998. *WordSmith Tools Version 3*. Oxford, England: Oxford University Press.

NATION, P. 2005. Range. Available on-line at: http://www.vuw.ac.nz/lals/staff/Paul_Nation

BERBER-SARDINHA, T. 2000. *Comparing Corpora with WordSmith Tools: How Large Must the Reference Corpus Be?*. Sao Paulo, Brazil: Catholic University of Sao Paulo.

POE, E. A. 1848. *Eureka*. Project Gutenberg. Available on-line at: www.gutenberg.org

POE, E. A. 1846. *The Cask of Amontillado*. Project Gutenberg. Available on-line at: www.gutenberg.org

POE, E. A. 1839. *The Fall of the House of Usher*. Project Gutenberg. Available on-line at: www.gutenberg.org

POE, E. A. 1842. *The Masque of the Red Death*. Project Gutenberg. Available on-line at: www.gutenberg.org

POE, E. A. 1846. *The Philosophy of Composition*. Project Gutenberg. Available on-line at: www.gutenberg.org