# Universum Inference and Corpus Homogeneity

Carl Vogel,* Gerard Lynch and Jerom Janssen

**Abstract.** Universum Inference is re-interpreted for assessment of corpus homogeneity in computational stylometry. Recent stylometric research quantifies strength of characterization within dramatic works by assessing the homogeneity of corpora associated with dramatic personas. A methodological advance is suggested to mitigate the potential for the assessment of homogeneity to be achieved by chance. Baseline comparison analysis is constructed for contributions to debates by nonfictional participants: the corpus analyzed consists of transcripts of US Presidential and Vice-Presidential debates from the 2000 election cycle. The corpus is also analyzed in translation to Italian, Spanish and Portuguese. Adding randomized categories makes assessments of homogeneity more conservative.

## 1 Background & Method

Recent research in text classification has applied the assessment of corpus homogeneity to strength of characterization within fictional work [8]. The idea is that a character within a play is a strong character if the text associated with the character is homogeneous and distinct from other characters—the character is strong if a random sample of text of that character is more like the rest of the text of that character than it is like the text of other characters. Another possibility is that random samples of texts of a character reliably are most similar to its play, at least, if not its character. A playwright whose characters "find their author" in this sense, but not their characters or play, while still highly individual as an author, does not construct strong characters. One goal of this paper is to provide a baseline for comparison in which the contributions of individual characters are not scripted by a single author, but whose contributions have to be understood in light of each other's statements, like dialog: we assess homogeneity of contributions to national election debates.

Another focus of this work is in an attempt to improve the methodology for assessing the homogeneity of corpora. The method is related to inference with the universum in machine learning. Random data drawn from the same probability space as the corpora under consideration are considered among the actual corpora and categories within it. Inference with the universum involves approaching classification

Computational Linguistics Group, Intelligent Systems Laboratory, O'Reilly Institute, Trinity College, Dublin 2, Ireland e-mail: vogel,gplynch,janssenj@tcd.ie

* This work is supported by Science Foundation Ireland RFP 05/RF/CMS002.

tasks by supplementing data sets with data points that are not actually part of the categories from which a system is choosing, but which are realistic given the features under consideration [9]. The supplemental data sets, if well chosen, can sharpen the distinction between categories, making it possible to reclassify data points that otherwise fall in between categories. Part of the idea is that clashes with the universum should be maximized. Research in this area includes focus on how best to choose the universum [6]. One can use the same sort of reasoning to quantify the homogeneity of the categories in terms of their propensity to be confused with the universum material. As corpus homogeneity is assessed in part by rank similarity of files within it, the effect of adding random data is to diffuse the homogeneity of texts within a category since it is increasingly likely that randomly constructed data files will be the most similar to some of the actual texts. Thus, a category that is assessed as significantly homogeneous even with the addition of random data can be judged with greater confidence, with a reduction of the possibility of type I error.

In §2 we apply our methods to assess the homogeneity of debate contributions of main the contributors to the US national election debates from 2000.[2] Surprisingly, the transcripts do not reveal Bush or Gore to have provided self-homogeneous contributions in the sense used here (if they had been characters in a play, they would not have been among the strong characters). Adding fabricated contributions drawn from random sampling from the concatenation of the entirety of the actual data set alters the outcome by weakening some of the rank similarity measures within actual categories. The second experiment individuates the same corpus into a larger number of smaller categories: the categories are individuated by speaker and date, rather than aggregating across the debates. Then universum data is added. Finally, using translations of the debates into Italian, Portuguese and Spanish we turn the problem into one of language classification. On inspecting the texts of Bush vs. those of Gore, one might not think them as distinct from each other as texts of Italian are from those of Spanish. Whatever one's prior expectations about the homogeneity of categories individuated by speaker, there are very clear intuitions about categorization by language (and the effectiveness of letter distribution analysis in underpinning language classification generally [1]). Thus, we are able to use the universum method to enhance the assessment of homogeneity in general instances of text classification problems, as well as in computational stylometry.

**The classification method used here** involves several stages of analysis. A corpus of text is split into files indexed by categories. Files are balanced by size. In any one sub-experiment, the number of files in each category considered is balanced. Experiments are repeated hundreds of times, and average results analyzed.

The first stage is to compute the pairwise similarity of all of the files in the sub-experiment. Similarity is based on *n*-gram frequency distributions, for whatever level of tokenization that is settled upon, and for whatever value of *n* [7]. In the experiments reported here, we use letter unigrams. Their efficacy in linguistic classification tasks is perhaps surprising, but they have repeatedly proven themselves

[2] The presidential debates occurred on October 3, 2000, October 11, 2000, and October 17, 2000. The Vice Presidential debate occurred on October 5, 2000. The transcript source was http://www.debates.org/—last verified, June 2008.

[8], and perform well with respect to word-level tokenization [5]. However, other levels of tokenization are obviously also effective. An advantage of letter unigrams is that there is no disputing their individuation, and this renders it very easy to replicate experiments based on letter unigrams. This is important if the text classification task involves authorship attribution for forensic purposes [2]. The similarity metric used is the chi-by-degrees-of-freedom statistic suggested for the calculation of corpus homogeneity in the past by Kilgarriff, using word-level tokenization [3]. This essentially means calculating the $\chi^2$ statistic for each token in the pair of files under consideration, and averaging that over the total number of tokens considered. Normally, $\chi^2$ is used in inferential statistics to assess whether two distributions are significantly different; however, here we are using the value in the other direction, as a measure of similarity. With all of the pairs of files evaluated for their similarity, files within categories can be ranked for their overall similarity. For each file in a category, the Mann-Whitney rank ordering statistic is used to assess the goodness of fit of the file with respect to its own category (its *a priori* category), and with respect to all other categories under consideration on the basis of the ranks of pair-wise similarity scores. The best-fit alternative categories are recorded.

Homogeneity of a category of files is measured with Bernoulli Schema. This is akin to tossing a coin in repeated experiments to assess whether the coin is fair. Here, the coin has *c* sides, one for each category that could be the best fit for a file. In any one fairness experiment, the *c*-sided coin is tossed *n* times, once for each file in the category. With hundreds of *n*-toss experiments, it is possible to assess whether the coin is fair: when the same side comes up often enough relative to those parameters, it is safe to reject the hypothesis that the coin is fair (that the category of files is randomly self-similar) and to accept that the category is significantly homogeneous.

## 2 Experiments

The debates from the 2000 US presidential election cycle involved three debates between George Bush and Al Gore, and one Debate between Joseph Lieberman and Dick Cheney. The transcripts were parsed using PlayParser [4] into the contributions of each of the speakers, including a moderator and a member of the audience. The resulting data files were approximately 40K bytes for each of the candidates for each date, and approximately 10K for the moderator. These files were processed using the Unix split command to generate sub-files balanced at approximately 2K each. The data associated with the candidates are indexed in two ways—one is in a broad category which includes all of the speaker's files, and the other uses more categories, treating the different dates for each candidate as a separate category. The files of the moderator are always held within just one category. There was not enough data from members of the audience to consider in the experiments.

The universum data was constructed as follows. The entire data set of the debates was concatenated into a seed file. This was used as a representative distribution of characters to sample from. Depending on the exact configuration of the experiment, a matching number of categories was selected, and those categories were used for a set of files balanced in size and number with the actual data. However, the files were constructed by random sampling from the entire corpus construed as a bag of letters. Thus, the universum is a random data set shaped by the overall distribution of the actual data under consideration. If different levels of tokenization were employed or a different value of $n$ in the $n$-gram, the seed bag would be shaped accordingly. In the experiments that follow, we first run the experiment with the underlying data that we wish to analyze, and then again with universum categories.

### Experiment 1—Speakers Define Categories

In this experiment, the categories were construed from the four candidates and the moderator. The files from all four of the debates were considered together. In each sub-experiment, ten files from each category were considered from the approximately 20 available for each speaker in each debate, and 1000 such experiments were run with independent random samples from the superset in each. In each subexperiment, we assessed the homogeneity of each category. This meant considering how many of the ten files in the *a priori* category had that category as its best fit in terms of overall similarity. The results are averaged across all 1000 subexperiments. Given these parameters (five categories, ten files each), six out of ten files must be assigned to a category on the basis of similarity for it to be deemed significantly homogeneous.[3] Only the categories associated with Cheney (9.473), Lieberman (7.357) and the Moderator (7.723) are significantly homogeneous. The confusion matrix associated with the assignment of files that did not fit into its *a priori* category is can be summarized as follows: the Cheney and Lieberman categories attract the files associated with the each of other categories (the Moderator is nearly equivalent in homogeneity to Lieberman, but is not an attractor at all).

Next we consider the data associated with another 1000 experiments, but with the additional files generated randomly according to the constraints discussed above (here, five constructed categories with randomly generated files, seeded by the distributions in the concatenated corpus, balanced in size at 2K, with ten files to each category). With ten categories and ten files, five files is the threshold for category homogeneity. The significant homogeneity values are reduced to: Cheney, 8.864; Lieberman, 5.080; Moderator, 7.199. In another universum variation, there is a single large (containing 60 files) randomly generated category, and sampled ten files to any one subexperiment, just like the categories of the actual participants, in each of 1000 random samplings over all of the categories are run. With six categories of ten files each, the critical homogeneity level is six files assigned to the category: Cheney, 9.268; Lieberman, 6.450; Moderator, 7.513. These results suggest that when adding universum categories, it is more conservative to add as many categories as there are categories in the underlying dataset than to draw upon a single large universum.

[3] The significance threshold is $p < 0.05$ throughout this article. A longer version of this article, with full tables and confusion matrices is available (http://www.cs.tcd.ie/Carl.Vogel/vlj.sgai2008.pdf).

### Experiment 2—Individual Debates

We also considered each debate in isolation. The moderator files are considered as a monolithic category, however. Again, 1000 experiments were constructed each with a sample of ten files from each of the categories. With nine categories and ten files per category, a group of at least five files achieves significant homogeneity. In this experiment, Cheney (9.202), Lieberman (6.230) and the Moderator (7.435) are associated with significantly homogeneous categories as before, as is the Gore category for October 11 (6.275). A confusion matrix of the participants' distribution in similarity to each other across the debates can be summarized thusly: Cheney is an attractor for nearly all the categories; Bush's best alternatives are other Bush files and Cheney; Gore's best alternatives are other Gore files, Cheney and Lieberman.

With an additional nine random categories drawn on the same underlying frequency distribution and containing 20 files each, balanced at 2K in size, and 1000 experiments selecting 10 files from each category. With 18 categories and 10 files per category in each experiment, four files for a category achieves significant homogeneity. The Cheney data from October 5 (8.262), the Moderator data (6.833), and the Gore data from October 11 (5.571) remain significantly homogeneous. The Lieberman data loses significance. None of the individual universum categories are significantly homogeneous. This is because, as before, their files are randomly constructed from an underlying distribution and the files within the categories are often most similar to others of the random categories.

### Experiment 3—Translation Filter Added

The debate transcripts are translated into French, German, Italian, Japanese, Portuguese, and Spanish.[4] The source indicates that the translations were constructed using an MT system from the English source data, but does not name which one. The effect of using the translated data is that it introduces noise, presumably uniformly, to each of the natural categories. These languages are sufficiently close to each other that there should be some classification of files to the wrong languages, but equally, the languages should for the most part form clearly separate categories. 1000 experiments were run, with 10 files for each category defined by language. With four categories, seven out of ten is the threshold for significant homogeneity. Each category is homogeneous, as expected (English, 10; Spanish, 9.994; Italian, 9.962; Portuguese, 9.947). The 1000 experiments re-run with four universum category diminishes the homogeneity values, but they retain strong significance (English, 9.949; Spanish, 9.388; Italian, 9.475; Portuguese, 9.788).

[4] See http://www.debates.org/pages/trans_trans.html—last verified, June 2008.

## 3 Final Remarks

These experiments have shown that adding categories of files randomly constructed from the same underlying distributions as the data set under scrutiny has the effect of making judgements of category homogeneity more conservative. Adding the same number of universum categories as underlying categories enhances conservatism. This is certainly the right direction for conclusions associated with positive authorship attribution in forensic settings in which standards of certainty must be made extremely strict. For other tasks, this may be less advantageous. This work is part of a larger project in extending and validating methods of text classification. We are particularly interested in computational stylometry and the use of these methods to assess the strength in linguistic terms of characters in plays. Here we have applied the same reasoning to contributors to political debates. The results provide some baseline data about expected levels of homogeneity among individuals providing sequences of short monologues in a shared setting which can be used to illuminate the results that emerge when the monologues are all scripted by the same author. We are considering debates from the 2004 and 2008 election cycles as well. With a larger corpus of data, and more temporal calibration to the corpus, it should be possible to expand the analysis of linguistic homogeneity of real players over time.

## References

1. W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994. Las Vegas, NV, UNLV Publications/Reprographics.
2. Carole Chaski. Who wrote it? Steps toward a science of authorship identification. *National Institute of Justice Journal*, 233:15–22, 1997.
3. Adam Kilgarriff. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133, 2001.
4. Gerard Lynch and Carl Vogel. Automatic character assignation. In Max Bramer, editor, *AI-2007 Twenty-seventh SGAI International Conference on Artificial Intelligence*, pages 335–348. Springer, 2007.
5. Cormac O'Brien and Carl Vogel. Spam filters: Bayes vs. chi-squared; letters vs. words. In Markus Alesky et al., editor, *Proceedings of the International Symposium on Information and Communication Technologies*, pages 298–303, 2003.
6. Fabian H. Sinz, Olivier Chapelle, Alekh Agarwal, and Bernhard Schölkopf. An analysis of inference with the universum. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS 2007)*, pages 1–8, 2008.
7. Carl Vogel. N-gram distributions in texts as proxy for textual fingerprints. In Anna Esposito, E. Keller, M. Marinaro, and M. Bratanic, editors, *The Fundamentals of Verbal and Non-Verbal Communication and the Biometrical Issue*, pages 189 – 194. Amsterdam: IOS Press, 2007.
8. Carl Vogel and Gerard Lynch. Computational stylometry: Who's in a play? In A. Esposito, N. Bourbakis, N. Avouris, and I. Hatzilygeroudis, editors, *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. Berlin: Springer, 2008. To Appear.
9. Jason Weston, Ronan Collobert, Fabian Sinz, Léon Bottou, and Vladimir Vapnik. Inference with the universum. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 127–134, 2006.