INKUBATE
The network for creative collaboration

JUOLA & ASSOCIATES

MarkIt!™ Book Predictability Application

Published by Inkubate and Joula Associates, Fall 2015

The Challenge

Bringing manuscripts to market is time-consuming and expensive. It takes a tremendous amount of expertise by agents, editors and publishers (AEPs) to determine which manuscripts should be acquired, curated and brought to market. This process is not just an "art-form," but also relies on the critical ability of AEPs to make sound financial assessments as to whether a manuscript is a worthy investment, creating, ***"the business of publishing."***

But determining which manuscripts to publish has, up until now, always been a subjective process combining many intangibles. AEPs often try to evaluate a new writer by looking for similarities in their "style" of writing and that of other known and published authors. In parallel, market estimates, surrounding potential **retail unit sales of a new manuscript**, are often extrapolated from old sales data associated with a previously published author or book. This can lead to miscalculations in readers' "tastes," royalty advances, pricing, production costs, inventory controls or missed and squandered opportunities.

In a publishing marketplace that is rapidly evolving and becoming more and more competitive, missed opportunities are never good. But with an historical rate of only approximately 2 out of 10 published books finding commercial success[1] in any given marketplace, it is clear that opportunities are missed. Famously, 20 separate editors rejected Golding's *Lord of the Flies* before it was finally published, and 12 editors rejected Rowling's *Harry Potter and the Philosopher's Stone* before it was finally placed into production. These examples of editorial rejection raise several questions, including:

- What other books did these editors and publishers "green-light" while they passed over these eventual best sellers?
- What did these other "green-lit" books cost to develop?
- What did these other "green-lit" books deliver as a return on investment based upon the material and human capital time spent, i.e., profitability?

What Else Could Help Analyze a Manuscript's Potential Success?

Wouldn't it be great to have additional objective validation metrics that could help determine which manuscripts to choose and how much to invest? That is where the science of *stylometry* (pronounced Sty-LOM-e-try), intersects with artificial intelligence and its ability to identify unique "best-seller" traits of an unpublished manuscript.

Stylometry is the statistical analysis of writing structures, patterns, and words; in essence, a writer's `style.' Inkubate has deployed stylometry as the foundation for a first-of-its-kind tool, MarkIt!™, which is proven to accurately predict the **"unit-sales potential"** of unpublished manuscripts. By combining MarkIt!™ with traditional editorial curation and promotional activities, it is now possible to consistently couple the artistic and creative aspects of publishing with business analytics, to drive success in today's highly competitive marketplace.

The Science of Stylometry

Almost everyone agrees that well-written books sell better. Publishing professionals generally agree that one key factor is "writing style." While this appears to simply replace one ill-defined term with another, and while no one can really tell you what "good style" really means ("use lots of action verbs" or "never split infinitives!") stylometric analysis goes deeper to correlate writing style traits shared by and between writers that could help agents, editors and publishers make more educated "up-front" choices, than they might make using only their own experience.

Stylometry uses cutting edge natural language processing technology backed by 25 years and over 50,000 hours of applied application analyses identifying "authorial fingerprints."[2] These authorial fingerprints identify characteristic patterns of language use by humans in their writing. Researchers[3] have shown that authorial fingerprints are the "human stylome" of a specific set of measurable traits that uniquely identify a given author.

Our technology platform has learned to recognize the characteristics of "saleable writing" by identifying thousands of unique authorial fingerprints contained within published books and assessing those fingerprints against the text that is contained within an unpublished manuscript. Specifically, our science and technological analysis can accurately and consistently predict the anticipated unit sales of a manuscript during the first twelve months following its original publication date.

To understand the science behind our stylometric technology, consider the fact that every word someone writes is the product of personal choice. For example, the English language provides a tremendous number of choices for words to describe something bigger-than-big, words such as "huge," "giant," "enormous," or "colossal." Writers can choose to express an idea with a few precise words or a bunch of common, general ones, and similarly they might choose to break a complicated idea — or not — into bite-sized simple sentences. When writing, authors can put an adverbial phrase at the front of the sentence, at the back, or somewhere in the middle. Most authors are not even conscious of many of these writing style and word selection choices.

As an example of one of the 250,000 algorithms our stylometric analysis platform tracks, consider "function-words." These are the short, common, lightweight words like prepositions, articles, and pronouns used by an author within their writing. The differences in "function-word" word choices one writer makes in their writing versus the "function-word" word choices made by another writer, can be easily illustrated by thinking about how a writer may choose to identify and describe a table place-setting. Writers could choose to describe the place-setting where the fork is "***to***" the left of the plate, "***on***" the left of the plate, or, maybe, "***at***" the left of the plate. Whichever way an author chooses to describe a place setting within their writing, this leaves behind "a pattern" related to their function word choices.

In addition to the word choices our software can analyze, we can look at aspects of readability such as word frequency, word length, or psycholinguistic accessibility. We can look at sentence structure and grammar via the use of specific parts of speech or word clusters and we can delve into semantics and morphology by looking at word roots, prefixes, suffixes, and letter clusters, or rhythm and poetry by looking at syllables. We can even explore the more esoteric aspects of writing style such as the specific words a writer uses to open sentences. Once our stylometric analysis platform collects one or more of these choice patterns and applies classifications, such as, linear discriminant analysis[4] and support vector machines[5] to identify the key aspects of the author's writing, our technology can then reliably identify that writer again and again using this "authorial fingerprint" when presented with multiple unknown documents of interest.

Our stylometric technology can also solve other problems when analyzing writing style. A specific use case from 2 years ago was the use of our techniques in the analysis of Robert Galbraith's *A Cuckoo's Calling*. Robert Galbraith was a here-until-now unknown writer that shot to the top of the charts. There was wide suspicion in the British tabloids that he was not a real person at all, but in fact, a pseudonym. His personal story began to quickly unravel as journalists pursued leads into his history. Nobody had ever met him (except his agent) and not one shred of evidence could be found to support many claims in his biography that attempted to establish his existence. Our engineering and mathematics team was contacted to investigate who Robert Galbraith was and, by using our stylometric technology, our platform was able to determine who Robert Galbraith really was in a little over 6 hours from when we began examining his title against other published authors. We analyzed four separate tracking parameters, including: function words; distribution of word lengths; the use of words in context (as word pairs); and groups of consecutive characters like the "*tion*" in the word "action" or "attention" to determine how these patterns could identify writing style within *A Cuckoo's Calling*.

Our stylometric technology identified hundreds of thousands of writing style points using only these four parameters and performed over one million points of analyses to determine that Robert Galbraith's writing style was *exactly* like JK Rowling's. Consequently, Rowling was brought into the light by our technology and begrudgingly and publicly admitted that she was, in fact, writing under the name of Robert Galbraith, a person who, at least in this instance, was a fabrication[6].

Just as individual people write differently from each other, so do groups of people based upon where they are from. With our stylometric platform's analysis technology, we can review writings by American and Commonwealth authors to identify writings "in US style." Similarly, a person's writing style can identify many characteristics and data points that correlate to their nationality, native language, age, education, social class, gender. Our stylometric technology's analysis features will also identify intangible characteristics, such as personality traits or background. As an example, the word "colour" means something different about a person's background than the word "color." This would be one feature among many that can give the researcher a confident indicator of individual or group identity. By analyzing a novel with our stylometric technology, our platform can determine many of the characteristics of the person who wrote it.

Our proprietary technology combines millions of individual indicators of identity into an overall judgment based on observed patterns of how people write.

Our technology has been peer-reviewed and the published research indicates that our stylometric analysis can determine aspects of identity with 95%+ accuracy.

How Stylometry and MarkIt!™ Work Together

What happens when the stylometric analysis is related to "writers who sell well?" We have learned that this is actually just another external trait with a high statistical correlation to a proprietary measure of a **population's writing style**. With the MarkIt!™ tool (for publishing professionals), our population is a group of writers and manuscripts that have performed **similarly** in the retail marketplace. And so, by analyzing books that have sold well and poorly according to the past 10 years of Nielsen's BookScan™ sales rankings, we have built a "corpus" in which our stylometric technology has identified the writing, stylistic, and character-traits within manuscripts that "group" books together into groups that correlate directly to the selling levels of those published works[7]. Those "groups" have been segmented into 11 unit sales level brackets (e.g., sold 0-100 units 101-4,000 units, 4,001-10,000 units, etc., all the way up to 1,000,000s of copies). By analyzing a new manuscript with our stylometric methodologies and comparing its "authorial fingerprint" to our corpus, we can objectively and accurately predict a manuscript's unit sales potential.

To validate our techniques, 1,000s of previously published titles were stripped of all identifying information (e.g., title name, author name, publisher, publication date, genre category, and format) and the results processed to unformatted text files. Each title was then assigned a unique identifying code for later reference. We built thousands of models of book sales, and then validated each model using leave-one-out cross-validation to select only the most accurate models. This scheme assures that new books would be placed by analysis into the correct grouping, and if not, measures how closely they would be placed. By using this methodology, the MarkIt!™ tool's analysis is returning mid 90th percentile accuracy rates. So, while our Stylometric analysis may occasionally miss a potential bestseller, it will still spot it as a close runner-up!

How Our Stylometric Technology Was Developed

We have spent over $2 million in federal research support grants to develop the science of Stylometry into a simple, lightweight, elegant, and novel technology for writing assessment. Beyond unmasking J.K. Rowling's newest pseudonym, Robert Galbraith, our technology has also been used in numerous court cases where determination of document authorship has been fundamental to the facts of the case. Additionally, the Defense Advanced Research Projects Agency (DARPA) is using our technology to develop a new system to secure our nation's computers.

Conclusion and What Does this Mean for Publishing?

As a creative arts industry, publishing is unique. The other creative arts industries (music, film and television) deliver far less total content choice and most of that is delivered passively. Reading still demands a high level of active consumer interaction to choose amongst literally millions of possible books and in the consumption of that material. Reading is also intellectually demanding. It is almost impossible to multitask while reading. Imagine trying to "book surf" or, trying to read more than one book simultaneously, like some people "channel surf," their TVs? Likewise, most people have no problem singing or humming along with the car radio, yet imagine trying to read a book out loud while driving?

The myriad of content and the **choices** consumers must make to devote hours and sometimes days or weeks of **active** time to the reading of just one book seemingly sets the publishing industry apart. Yet, the competitive pressures and impact of new technologies and new marketplaces creates the same challenges all other creative arts industries face. Other industries have adapted their content acquisition strategies to include many different objective data streams. Pre-screening content, running focus groups, testing content "on tour," or carefully monitoring the effectiveness of content to attract viewers or listeners and, thus, deliver advertising revenue; are all strategies now commonly employed early on in the content acquisition process of the others.

Inkubate believes that the use of our stylometric analysis techniques can fundamentally change the economics of publishing and deliver opportunities to publishers across all levels of unit sales. If a publisher has reliable objective intelligence that can predict unit sales **before** any investment has been made in human and capital resources, then that publisher can theoretically scale each project appropriately to more likely obtain a success in the marketplace. Publishing **does not** have to remain a largely "hits driven business" in order to accommodate the "2 for 10" success rate reality we introduced at the beginning of this whitepaper. What if stylometry could consistently change that metric to “3 for 10” or “4 for 10 or, perhaps, better?”

The President of one of the prominent top-25 international trade publishers said: “This would revitalize our space and I could feel more comfortable supporting the riskier projects . . . I would be able to know going in what kind of financial result to expect and could plan my budgets accordingly . . . I could find opportunities in niche marketplaces and support more young or unproven authors . . . I’d be more competitive . . . For the better, it would fundamentally change the business of publishing.”[8]

About Inkubate and Juola

Inkubate’s SaaS platform and analytics technology has been in development for the last 4 years in an effort to provide content creators and publishing professionals with state of the art collaboration, communication and content analysis tools. Inkubate’s technology combines “content matchmaking” with “B2B social networking” to enable publishing professionals to efficiently search for and analyze manuscripts and determine the potential for their commercial success.

Our engineering and math sciences team has over 40 years of combined experience applying these types of stylometric analyses within university research settings. The leader of our analytics efforts is Dr. Patrick Juola, who spun off of the cutting-edge research of the “Evaluating Variations in Language Laboratory” (EVL Lab) of Duquesne University, in Pittsburgh, PA to form Juola & Associates in 2010. The EVL Lab has developed one of the most widely used and most accurate systems for attributing authorship in the world. Dr. Juola is internationally recognized for his research and outcomes, and his team of PhD’s includes two former members of the intelligence community who bring their expertise in both software development and text analysis within the national security marketplace.

By combining Joula Associates’ expertise with Inkubate’s publishing industry domain expertise, our companies are solving very complex text analysis outcomes for authors, agents, editors, and publishers using industry and world leading, state-of-the-art, artificial-intelligence driven investigative stylometric techniques.

Contact Information

For further information about our Stylometric applications, please contact, Jay Gale, CEO of Inkubate, jdgale@inkubate.com (603) 491-1168, David Bass, CMO of Inkubate, david@inkubte.com, (415) 710-7775 or Don Seitz, SVP of Sales and Business Development, don@inkubate.com (908) 642-0071

[1] Top 25 International Trade Publishing House Internal Statistics, 2000-2014

[2] Foundations and Trends in Information Retrieval, Vol. 1, No. 3 (2006) 233–334 Copyright, 2008, P. Juola, DOI: 10.1561/1500000005

[3] Hans van Halteren, University of Nijmegen, The Netherlands Copyright 2004

[4] Linear discriminant analysis (LDA) is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events.

[5] Support vector machines (SVM) are a group of supervised learning methods that can be applied to classification or regression. Support vector machines represent an extension to nonlinear models of the generalized portrait algorithm developed by Vladimir Vapnik.

[6] Patrick Juola, “How a Computer Program Helped Reveal J. K. Rowling as Author of *A Cuckoo’s Calling:* Author of the *Harry Potter* books has a distinct linguistic signature,” Scientific American, August 20, 2013**.**

[7] MarkIt!™ normalizes our data for the first 12 months of unit sales following the initial and unique publication date of a new title.

[8] Interview conducted at BEA 2015, Javits Center, NY, NY