We continue our series on using Natural Language Processing tools in the writing and editing process.

Software-Assisted Ghost Writing
Characterizing the Author's Voice, Part 1

Like many people, I enjoy reading mysteries. I prefer challenging cases, with stylish writing and memorable characters. So, it should be no surprise that two of my favorite authors are Rex Stout (Nero Wolfe) and Dorothy L. Sayers (Lord Peter Wimsey). Both of them passed on over four decades ago. Even so, their readers have been delighted with the renewal of their series in recent years.

Twenty years ago, the trustees of the estate of Sayers' son authorized Jill Paton Walsh to complete an unfinished manuscript that Sayers left behind. After the success of that venture, she has produced three more original volumes, all of which have been well-received.

Over thirty years ago, the Stout estate selected Robert Goldsborough to continue the Nero Wolfe series. To the delight of Nero Wolfe fans, he created seven novels to successfully extend the franchise. After a hiatus of eight years, he has written another six Nero Wolfe novels in this decade.

Both Walsh and Goldsborough faced problems similar to those encountered by ghostwriters. The characters created by the original authors are beloved by their readers. The character can mature and develop (as Lord Peter has), but they can't act inconsistently with their core beliefs. Perhaps more importantly, they can't speak inconsistently with their past conversations. The personalities of real people sometimes mature, but their speech patterns rarely change in adulthood.

In the same way, the prose style of the original author must be maintained. Both readers and reviewers expect that the narrative to read as if written by the original author. Part of the appeal of these series is the universe created by the author. The narrative style builds that universe.

A major challenge faced by a ghostwriter is to write a book that sounds the way the credited author speaks. There are both differences with and similarities to the problems faced by authors extending a franchise. The extending author has the benefit of a large body of written work. This can be mined for the narrative style of the original author, as well as the speech patterns of the characters. On the other hand, an author working with a ghostwriter probably doesn't have a large body of written work. If they did, they would be unlikely to have hired a ghostwriter.

This is why it is important to record and transcribe the research interviews that the ghostwriter conducts with the author. The author's recorded words can serve the same purpose as the written body of work does for an author extending a series. Successful ghostwriters listen both to what their authors say, and how they say it.   A written record assists the memory, and offers a chance to notice things that get missed initially.

So, what exactly does it mean to write in a style like someone else?   There are a quite a few indicators. First, there is distinctive vocabulary. If an author writes about two very different subjects, there should be a difference in vocabulary. But when one person writes about the same subject, during the same time period, there should be a common vocabulary. In addition, the author will use some terms more frequently than others.

If an author writes about related subjects, but several decades pass between writing each work, there often is a difference in vocabulary. Over 25 years have passed since the series *Star Trek: The Next Generation* was completed. You can tell this by the computer jargon the characters employ. They regularly use terms like "mainframe", "subroutine", and "memory bank". These were common jargon in the late 1980's and early 1990's, but seem very dated now. If the same script writers were asked to "re-boot" that series, we would expect them to use today's terminology.

A ghostwriter needs to identify and use the unique vocabulary of the author, and of the subject being addressed. Vocabulary includes not only individual words, but also phrases. A relatively simple program can extract the unique words from a set of interview transcripts, ranked by frequency of use. We'll address phrases later.

Second, there is distinctive grammar. There are several ways to identify a person's style using grammar. Bureaucrats, attorneys and academics tend to use the passive voice more often than the general public. Many experts say that the passive voice characterizes English that is hard to understand. Fortunately, people who use it often are unlikely to hire ghostwriters, so we don't need to mimic it.

People of different classes use different grammar. The Dean of an imaginary Oxford college had this bit of conversation with the aforementioned Lord Peter Wimsey:
"Then I am instructed to convey to you the fellows' invitation to high table tomorrow evening. Shall you come?"

“An invitation to dinner, and in the future interrogative mood. A most challenging form. And I shall. Come, that is.”
You can tell a member of the British aristocracy or intelligentsia by the verb phrases they use, even if you can’t hear their distinctive accent on the printed page.

Extracting the verb phrases from a set of transcripts requires a much more sophisticated program. There are open source English parsers available. They use statistical parsing, which is the most effective method currently known. However, they only achieve an accuracy rate of about 90%. To achieve accuracy above 98%, the output needs to be post-processed by programs that use different technology.

The English grammar sitting by my desk lists 41 compound tense verb forms, not including imperatives and infinitives. Most speakers of American English only use a small subset of them in conversation. We tend to use more in writing, and verb form usage can be used to identify an author’s unique style.

Here is a list of patterns that detect complex verb phrases. A program can apply these rules to identify the complex verb phrases used by an author.

```
present perfect:            (VBP have) (VBN _ )
past perfect:               (VBD had) (VBN _ )
future:                     (MD {shall,will}) (VB _ )
present indicative:         (VBP {am,are,is}) (VBG _ )
past indicative:            (VBD {was,were}) (VBG _ )
future perfect:             (MD {shall,will}) (VB have) (VBN _ )
present perfect indicative: (VBP {have,has}) (VBN been) (VBG _ )
past perfect indicative:    (VBD had) (VBN been) (VBG _ )
future indicative:          (MD {shall,will}) (VB be) (VBG _ )
```

The items in CAPS are tags for parts of speech, the lower case items are literals, and the __ indicates any verb of the specified form matches. (VB stands for verb base form, MD for modal auxiliary, VBG for present participle, VBN for past participle, VBD for past tense, VBP for non-3rd person singular present.)

In the next issue, we shall conclude our explanation of computational methods for characterizing the author’s voice.

References

Curme, G.O. (1947). *English Grammar.* Barnes & Noble.
Sayers, D.L. (1964) *Gaudy Night*, Harper & Row. BBC edition 1987.