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 2

Some people are offended by genius. This explains academics who try to prove that some great work of literature was not written by the author credited. For example, every few years we hear about a new theory on how Shakespeare didn't write one of the plays credited to him.

Of course, different people do have different writing styles, which can be identified by various factors. In this article, we conclude our survey of ways of characterizing an author's voice. The third way we can use to identify an author's style is by using transformational grammar.

Transformational grammar (T-grammar) begins with the hypothesis that sentences have a structure. This means that the words in a sentence are hierarchically ordered with respect to each other.

T-grammar has two parts. The first part is a set of patterns called kernel sentences. The second part is a set of over thirty transformations that can be applied to a kernel to produce a new sentence. Transformations can insert, delete or move words within a kernel, and can insert one kernel inside of another.

English has the following kernel sentences:

```
1: NP V-BE NP or ADJ or (ADV)  # being verb
2: NP V-INT (ADV)              # intransitive verb
3: NP V-TRAN NP (ADV)          # transitive verb
4: NP V-SEEM ADJ (ADV)         # verbs of sensing
5: NP V-BECAME NP or ADJ (ADV) # verbs of sensing through time
6: NP V-MID NP (ADV)           # middle verbs convey no action
```

(NP stands for a noun phrase, ADJ for an adjective, (ADV) for an optional adverb, and the V-forms are various types of verbs).

An example transformation converts "He should go" into "He should not go" by inserting a negation.

The T-grammar approach is more powerful than the verb phrase approach. It can identify differences the other approach can't. For example, "Will you be done on time?" is an

application of the T-yesno-question transformation on the kernel sentence “You will be done on time.” The verb phrase approach would just find “will be” in both sentences.

We can use T-grammar in several ways to identify the unique style of an author. First, we can count the fraction of all sentences in the corpus that match each of the kernel patterns. For example, an author who is writing for academic or bureaucratic consumption will have relatively few of these kernel sentences, since the “official style” considers long, complex sentences to be a mark of erudition.

Second, we can determine the fraction of all sentences in the corpus that applied any of the transformations. To compute these percentages, we need a software tool. It takes parse trees as input, which must be produced by another program. It classifies each sentence as a kernel sentences, or a sentence derived from kernel sentences using the transformations.

The final way we can identify an author’s style is by using N-grams. An N-gram model predicts the next word in a sequence based on the previous N-1 words. An N-gram is a sequence of words: a 2-gram is called a bigram and a 3-gram is called a trigram.

N-grams are used for a variety of tasks in natural language processing. These include speech recognition, machine translation, spelling correction, part-of-speech tagging, and authorship identification. This last use is the one we are interested in.

We compute probabilities by counting things. When we count things in natural language processing, we must have a corpus. This is a digital collection of speech or text. When we analyze the voice of a published author, we use their books and papers. When a ghostwriter seeks the voice of the author, he or she must rely on the interview transcripts.

The goal of N-gram analysis is to compute the probability of a word occurring next, given a history of words already seen. We estimate this probability using frequency counts. N-gram models approximate the entire history with the conditional probability of the previous N-1 words, given the current word. Most expositions of N-gram analysis are done with bigrams, but the extension to trigrams isn’t that hard. When we have a large enough corpus, we prefer the extra information provided by trigrams.

What kind of information does an N-gram provide? Some of it is lexical, some it syntactic, and some of it is semantic. Vocabulary can be used to find the unique style of an author. A bigram or trigram analysis finds the most frequently used words and also the most frequently used phrases.

Syntactic information collected by N-grams is distributed in the calculated probabilities. For example, the sum of the probabilities that each of the transitive verbs is followed by a specific noun reflects the rule that direct objects follow a transitive verb.

Semantic information is also distributed in the calculated probabilities. The probability that a form of the verb 'to see' is followed by a color name is likely to be non-zero. In contrast, forms of the verb 'to hear' would be unlikely to be followed by a color name, unless the author has synesthesia.

These two articles have shown that there are a variety of ways to characterize the voice or style of an author. To employ them, a ghostwriter needs the right software tools, and a corpus of sufficient size. With the output of these tools, a ghostwriter can write a book that sounds like the way the author speaks.

The ghostwriter can also produce objective evidence that the new work meets this important requirement. The ghostwriter can run the same analysis tools on the interview transcripts and the completed manuscript. The results should show the similarities in speech patterns.

References

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing, 2nd ed.* Pearson Education.
Ouhalla, J. (1999). *Introducing Transformational Grammar.* Arnold Publishers.
Williams, J. B. (1973). *Style and Grammar: A Writer's Book of Transformations*. Dodd, Mead & Co.