Verifying Authorship When There Is No Handwriting, Paper, or Ink

In 1992, linguist Carole Chaski was an assistant professor at North Carolina State University when she was contacted by a detective in Raleigh's Major Crime Unit, investigating the sudden death of a recent college graduate who was found in his apartment, in his own bed, with no signs of struggle or wounds.

The toxicology report, however, showed evidence of three common, legal drugs—together, a lethal combination—in the victim's system, raising suspicion of foul play. On the home computer, there were multiple drafts of typed suicide notes. Proof enough of suicide, you would think, but the detective on the case had no hard evidence that the victim had actually authored these letters. How could he prove authorship of the letters when there was no ink, paper, or handwriting to analyze?

> "How could he prove authorship of the letters when there was no ink, paper, or handwriting to analyze?"

Chaski ran a syntactic analysis of the suicide notes, comparing them to other documents the victim was known to have authored. Her analysis showed a one in 10,000 chance the victim had authored the suicide notes. She then ran the same analysis using documents authored by the victim's two roommates. The analysis eliminated one as the author, but for the other roommate, it was a match. There was no statistically significant difference between his syntactic patterns and those of the suicide notes. There was a high probability this roommate authored the suicide notes.

Chaski's revelation gave the detective the data he needed to focus his investigation, which led to charges against the roommate, who later admitted at trial that he had indeed authored the suicide notes. He was charged and convicted of voluntary manslaughter.

This experience led Chaski to further develop her methodology and paradigm for valid forensic linguistics. In 1998, she founded the Institute for Linguistic Evidence, the first venue for forensic linguistics research. In 2007, she founded ALIAS Technology, which provides police detectives, security teams, intelligence analysts, and attorneys with forensic linguistics services and software.

Using Chaski's patent-pending technology, the ALIAS platform has the power to detect linguistic forgery and reliably classify documents for authorship as well as linguistically assess suicidal language, threats, relationships between texts, and linguistic profiling.

The Challenge of Finding a Multilingual Language Parser

ALIAS Technology detects authorship using a linguistic methodology developed by Chaski. Unlike other forensic linguistics methodologies that rely on word frequency, spelling errors, and other obvious and easy-to-imitate features, Chaski's method is based on syntax—how we combine words into phrases and sentences.

"In order for us to produce language, we must combine elements into phrases and sentences. This ability is highly automatic and unconscious, so that we can focus our communication on what we are saying instead of how we are saying it. We automatically choose options that produce simple or complex structures. Our memories of how we actually say something degrades in milliseconds. What makes individuals' use of language distinctive," Chaski says, "is the pattern of simple and complex phrases used." Because the methodology relies on the analysis of sophisticated and normally unconscious linguistic features, it can be difficult for document authors to consciously manipulate or disguise these patterns.

> Chaskis method is based on syntax...the only method that's transferrable...from one language to another.

Chaski's methodology is about 95% accurate in identifying authorship, based on experiments run on different gold-standard or ground truth data in which the real authors are known. To date, it's also the only forensic linguistics methodology reliable enough to be successfully admitted as scientific testimony in the U.S. and internationally, under current legal standards for an expert to state a scientific conclusion. Because all languages have syntax, it's also the only method that's transferrable, with high accuracy, from one language to another.

While commercial and academic parsers do well with formal texts like novels and legal documents, they fail to understand the nuances of more casual writing, posing a problem for the analysis of documents like suicide notes, text messages, and blog comments.

To solve this problem, Chaski developed her own proprietary parser, which worked to suit her needs. Yet with this success, a new problem presented itself: the ability to parse non-English languages.

Says Chaski, “We were very successful with the work we’d been doing in English. Because of that success, we were attracting more and more international attention and cases. We had research associates from Pakistan, Spain, Korea, and China. As a linguist, I wanted to go multilingual.”

“Because I’m not a specialist in those languages,” she says, “developing a parser for them would have been a nightmare.”

How Rosette and the ALIAS Parser Work Together to Detect Linguistic Forgery

Since ALIAS Technology was still in its early stages, the company took advantage of BasisTech’s Startup Program, which partners with high-potential, growing startups to offer them access to technology that might otherwise be cost-prohibitive.

“The Startup Program has been wonderful,” Chaski says. “It says a lot to me about BasisTech, that they’re willing to help people who are still small. To me, it was an ethical issue that showed the company had corporate aspects I liked and valued.”

> “For the purposes of what we’re doing, Rosette is the only game in town.”

Because Chaski needs a parsing solution that departed from the traditional capabilities of commercial or academic parsers, the ALIAS parser and Rosette work together to fulfill her requirements. For new data being run through the ALIAS platform, Rosette acts as the first layer. The output from Rosette is then run through Chaski’s proprietary parser, which relaxes the grammar rules and lets her parse according to the rules people use in more casual writing.

ALIAS uses Rosette for part-of-speech tagging, lemmatizing, and entity extraction. When Chaski initially looked for a solution, she found that other text analysis platforms offered fewer languages or less advanced entity extraction. “For the purposes of what we’re doing,” she says, “Rosette is the only game in town.”

With Rosette automating the verification of “radio silent ever” applications only since Jan. 1, 2020, Palmer estimated in April 2020 that Rosette “has saved in the 100,000s of minutes this period. The reality is [without Rosette], we wouldn’t be able to apply automation.”

Increased Speed and Accuracy and Expansion to Multiple Languages with Rosette

The integration of Rosette into ALIAS Technology's stack has had two important results: the ability to expand to multiple languages, and increased overall speed and accuracy.

"The incredible gift of Rosette is the multilingualism and breadth of functional coverage," Chaski says. "I no longer have to worry about dealing with an open source solution that may never be worked on again after the person who wrote it finishes his master's. I don't want to have to go out every few years and find a totally different open source solution for every language."

Using Rosette allows ALIAS Technology to function in multiple languages, including Arabic, Korean, Spanish, Italian, and Russian.

Additionally, an unexpected benefit has arisen: the increased speed and accuracy that have accompanied ALIAS Technology's adoption of Rosette.

> "Rosette's English offers better speed and is extremely accurate."

While Chaski originally only used her proprietary parser for English cases, she now uses a combination of her original parser and Rosette's English parser. "As an industry-grade parser," Chaski says, "Rosette's English offers better speed and is extremely accurate. It works in tandem with the ALIAS parser."

Today, ALIAS provides linguistic evidence in cases including homicide, kidnapping, rape, trademark and patent infringement, custody agreements, witness tampering, plagiarism, accident investigations, defamation, and international arbitrations. The kinds of texts in these cases range from text messages, emails, essays, blog posts, accident reports to dissertations, novels and legal documents.

ALIAS algorithms have proven just as reliable on microtexts like instant messaging as on average-sized texts such as emails, and macrotexts like dissertations and judicial rulings.

Multilingually, so far ALIAS has been called on to support cases in the U.S. involving Spanish data, and in Canada with French data. In both cases, ALIAS was not required to give evidence in court, as often the opposing side will agree to the evidence when the method is very strong.

As ALIAS Technology's adoption continues to grow in multiple international locations, Chaski believes Rosette will be an integral part of the platform. "Across languages, almost every culture is now facing issues of electronic authorship, suicide, threatening communication, malware relationships, insider threats, conspiracy, to name a few ways that ALIAS provides reliable linguistic evidence."

"Our plans for growth are to keep using Rosette—specifically the multilingual parts of Rosette—to transfer the algorithms that are successful in English into other languages, and then be able to provide our services for those languages and cultures," she says.

Spotlight: Digital Fingerprinting of Text Messages

Text messages don't play by the same rules as other types of written speech. Unlike emails or more traditional documents, text messages rarely use punctuation or contain anything more substantial than a sentence fragment. Often, text messages are little more than a single word or letter.

With that little to work with, can a parser like ALIAS identify the author of a text message? In a case Chaski was called in to help with, a defense attorney was determined to find out just that.

The defense attorney was investigating a case in which an estranged husband and wife met to discuss reconciling. During this meeting, the wife claimed the husband kidnapped and raped her. He was arrested, and faced 25 to 30 years in prison.

What stuck out to the defense attorney was that 24 text messages had been sent from the wife's phone during the timeframe attributed to the crime. When questioned about it, the wife claimed her husband stole her phone and sent the messages himself. The defense attorney wanted to find out if she was telling the truth—if she wasn't, it could turn the entire case around.

Chaski set out to determine who really wrote those 24 texts, First, she used ALIAS to develop a statistical model, by analyzing over 3,000 text messages the husband and wife agreed they had sent to one another in the past. But because many of the texts in question contained short phrases like "On my way" or "K," there was simply not enough syntax in many of the texts to get a numerical profile.

Chaski's solution was to bundle the texts together so they more closely resembled a document that could be analyzed using ALIAS' core algorithms. When bundled into groupings of 100 in chronological order from each author, ALIAS could differentiate the bundled texts as being from the husband or wife with 96% accuracy.

Armed with a sound statistical model, Chaski applied it to the 24 text messages sent during the three-hour period when the crimes were supposed to have taken place. Her model attributed 23 of the 24 text messages to the wife, proving she did indeed have access to her phone during that time period.

Because of ALIAS, the charges of kidnapping and rape were dropped. Amazingly, the linguistic analysis of text messages proved the husband's innocence—and prevented the justice system from making what would have been a tragic mistake.