Hand annotation and reliability: Corpus Linguistic Approaches to Teaching and Studying Writing

Brian Larson, Texas A&M University School of Law

Available at: https://works.bepress.com/brian-larson/39/
I'm BNL, assoc prof. at TAMU Law. I'll be talking about [title]

From Program Abstract:
Speaker 7 will introduce options for annotating a corpus, using as an example a publicly available corpus of texts written by legal writing students. Annotating a corpus allows the researcher to “bracket” portions of artifacts, like complex legal citations, to exclude them from analysis; or to code portions to allow comparison, for example, between different common sections of a genre, like the fact section of a legal brief and its argument. Because annotation is labor-intensive, it often requires the work of multiple humans, which raises concerns about inter-rater reliability. Attendees will engage in an activity where they code and compare their codes on samples of texts from the corpus.
Throughout this talk, I've used examples from my own work, not because they are particularly good, but because I know them well. And my work could not normally be called corpus linguistics. For example, I used these principles in two studies: one using student-written legal texts, an example of which we will look at today. Published in \textit{Written Communication}, it’s perhaps more of a natural-language processing or NLP study. The other, with Lee-Ann Breuch at University of Minnesota, is an essay in the new methods-focused volume \textit{Points of Departure}. It’s a genre study of student use of research sources in technical writing where we didn’t use any computational tools for coding or assessment.

The examples I used today will be from the \textit{Written Communication} study.


So, why annotate or code texts by hand? [NEXT]
I wanted to ‘bracket’ text for exclusion or coding

For example, in a study I did with law-student texts, I wanted to exclude the caption block from my analysis, because it’s the same in every sample artifact.

I wanted to mark or bracket the “introduction” section in each student memo, because I wanted to compare the introductions to other parts of the memos.

In my project, there were other segments of text I wanted to partially exclude from analysis. For example, in these texts [NEXT]

There are complex legal citations that can ‘break’ the corpus linguistic and NLP tools, because they provide input that those tools are not used to processing. The citations are full of punctuation and abbreviations.

So, the text I wanted my software tools to see was not filled with these complicated things, but… [NEXT]
I wanted tools to see *this text*

Additionally, the Seventh Circuit’s view has gained traction in at least one other federal circuit. In *Gladwell Gov’t Servs., Inc. v. City of Marin*, which dealt with a copyright claim between a county and a contractor, the court acknowledged 17 U.S.C. § 101 by arguing, “[t]he plain language of the statute indicates that a work-for-hire agreement cannot apply to works that are already in existence. Works ‘specially ordered or commissioned’ can only be made *after* the execution of an express agreement between the parties.”

It may be easy for one to argue that the more recent decision produced by the *Playboy* court ought to trump the argument of *Schiller & Schmidt*. It should be noted, however, that the *Schiller & Schmidt* reasoning has not been forgotten in the years since.

Get this at http://tiny.cc/4c18bni

… instead excludes the citations. Once I have “bracketed” the text, I can use software tools to create this clean version. I don’t delete the citation data. I simply hide it from the NLP or corpus tools I’m using. But sometimes you want to do more than just bracket the text… [NEXT]

For details on software solutions to create the version of the text that is citation-free after manual annotation, see Larson (2015, Appendix G).
In this study, I thought I wanted to know when a citation string was a standalone sentence and how many cases or authorities were cited in each. I thought I might test for correlations between certain citation practices and the language I was seeing in the briefs. Here is an instance of a single citation string. It occurs as a standalone sentence and cites four separate authorities.

So, those are some reasons why you might want to annotate your text. Given that...
We need to do two things: ID spans of text to bracket and code them. But these two tasks present challenges, including the fact that humans might not agree at all or at least not perfectly and the tendency to apply category labels without care. Another challenge is the urge to code too much.

The things we want to do are (1) have a coding guide that describes how to annotate and resolves common problems (2) test inter-rater reliability or inter-annotator agreement and (3) limit what we require folks to code.
First, you need a coding guide, because humans don’t always agree. You have a very limited and rudimentary version in the handout titled “Workshop exercise coding guide.” It can help resolve common problems. So, we had two annotators, and in the first example, you can see a spot where during our development process one annotator thought there was a citation. Another annotator thought there was not. The second example shows the same authority cited in a separate “citation sentence.” This disagreement precipitated an instruction in the coding guide that a citation “is always set off from the grammatical portion of a sentence by commas or other punctuation. It is sometimes bracketed by parentheses.”
For these kinds of problems, the annotation instructions have to be more granular because computers are literal. From the computer’s perspective, the addition of a space at the beginning or end of a span makes it not the same span. Here are three different annotations of the same text span, the first including both leading and trailing spaces, the second including neither, and the third including only the leading space. These look different to the computer. The second page of the handout titled “Workshop exercise coding guide” provide specific instructions about these spaces and punctuation for our project.
Inter-rater reliability or inter-annotator agreement. Both terms refer to the same thing, the assessment of whether annotators agree. Generally, you’ll want 5 to 20% of texts annotated by two annotators so you can determine how much disagreement remains between annotators after they have studied your coding guide. One way to measure this difference is with the F-score. Software like the General Architecture for Text Engineering can do these comparisons for you. It can count matches strictly, where these two examples would not be counted as a match. Or it can count them leniently, where they would because they overlap. There are conventional thresholds about what the F-score needs to be to be reliable. The Cunningham text in the references discusses how this test works and how to do it with GATE. (See particularly chapter 10.)

Often you’ll want your annotators to categorize text they are annotating. The Cohen’s Kappa test is one way.... So, for example, if we ask two coders to identify the color of a dress as white blue or black, we assess the places where they agreed, in the solid black blocks here. But we can also see the places where they disagreed, which can be informative.... Here, for example, the annotators agreed pretty well on white dresses, but.... Note that this test is better than just counting instances where the two parties agree or disagree—what’s called “observed agreement” because this test accounts for the risk of random agreement.... So if Annotator 1 had simply coded every dress as white, she would have achieved 80% observed agreement with Annotator 2, who coded 80 dresses that way. See the references for further information on this test. Finally, you need to be careful when you define your categories [NEXT]

<table>
<thead>
<tr>
<th></th>
<th>Annotator 1</th>
<th>Annotator 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>White</td>
<td>Blue</td>
</tr>
<tr>
<td>White</td>
<td>70</td>
<td>5</td>
</tr>
<tr>
<td>Blue</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Black</td>
<td>1</td>
<td>9</td>
</tr>
</tbody>
</table>

- Use Cohen’s Kappa (or similar tests) to assess agreement on category codes.
- Cohen’s Kappa uses a “confusion matrix” or “contingency table” to ID agreement.

Get this at http://tiny.cc/4c18bnl

Image: https://www.wired.com/2015/02/science-one-agrees-color-dress/
Cunningham et al. (2014)
Two observations…. First, don’t have too many categories. This slide depicts the confusion matrix from two coders on the project I did with Lee-Ann Breuch. We had 10 possible codes, which means that just as a statistical matter, our annotators were unlikely to agree as often as we would have liked. The utility of the confusion matrix should be evident here…. The chapter explains our design woes further. Second, avoid casual application of social categories like gender and race. I’ve written and presented on the propensity of researchers in writing studies and natural language process to misuse gender as a variable in research….


One final observation. If you can make your annotations available to other researchers, that helps to advance the research world. Other scholars can examine your findings, test them, and build on them. Of course, you need to plan for that from the start in your study design. If you are gathering student writing, for example, you need your consent form to include a license of the student’s copyrights in order to distribute the texts.

Images: Parallel lines © 2013 theilr. License: [CC BY-SA 2.0](https://flic.kr/p/gDqtze). Copyright © 2013 Brian Mather. License: [CC BY-SA 2.0](https://flic.kr/p/gT1aVQ).
Activity

- For 10 mins, working by yourself
- Read “Workshop exercise coding guide”
- Identify and mark citation ‘spans’ in “Exercise for Hand annotation and reliability” according to coding instructions
- For 5 mins, convene with a neighbor and compare your results

Get this at http://tiny.cc/4c18bnl
Overview

You will read and mark citation segments in paper memos. You will identify the number of authorities for each cite. (In the actual study’s second phase, coders recorded the annotations on a computer using software called “GATE: General Architecture for Text Engineering,” which is not used in this C’s workshop.)

Identifying and marking citations

You will mark all citations you find, but only in the “Argument” section of the artifact.

Use whatever hand annotations are convenient for you to mark the paper copy, but be sure you can tell which adjacent space characters (shown as “_” in the handout) you are including and not including in your annotation.

A citation is a reference to a text or authority outside of the memo. It may be a reference to a case, statute, or other authority. It is always set off from the grammatical portion of a sentence by commas or other punctuation. It is sometimes bracketed by parentheses. For example, the cites in the following sentences are highlighted:

As children reach adolescence, courts recognize that the process of graining independence is an important consideration in determining duty and reasonable care. Restatement (Third) of Torts: Affirmative Duty § 42 (Tentative Draft No. 4, 2004).

Lime is a well known screenwriter with fifteen years of experience in television writing, (Compl. ¶ 11.), and OGS sought to commission Lime to write an episode of Lawless Love, (Id. ¶ 8.).

The two preceding examples also illustrate two kinds of cites:

Sentence: A Sentence citation is punctuated as a complete sentence set off from the rest of the author’s text. (Like the Restatement cite in the previous examples.) A sentence cite can be very short, like (Id. ¶ 24.) or (Id.). It is also possible for two citation sentences to appear in a row.

Clause: A Clause citation appears within one of the author’s prose sentences but is set off from it by commas (or sometimes a comma and a semi-colon or two semi-colons). (Like the Compl. and Id. cites in the previous examples.)

A citation may include explanatory material in parentheses or an explanatory clause. Parenthetical and explanatory information is included in the cite and should be marked as part of the cite.
When annotating citations, follow these conventions for covering the surrounding punctuation and spaces:

- For sentence citations, start your span with the first number, letter or parenthesis of the citation and end your span with the space before the next sentence begins:
  ...form. (Compl. ¶ 9). Then...

- For clause citations, start your span with the comma that begins the citation. Include the comma that ends it only if that comma is not necessary for the sentence to be properly punctuated absent the citation. Do not include a sentence-ending period in your citation span. Example:

  Indeed, deference to the urgings of Congressional agencies has been a long-standing tradition recently reaffirmed by the Supreme Court in U.S. v. Mead Corp., 533 U.S. 218 (2001). Quoting from the historic Skidmore v. Swift & Co., 323 U.S. 134 (1944), the court reiterated that “an agency’s interpretation may merit some deference whatever its form, given the “specialized experience and broader investigations and information” available to the agency, id. at 139, and given the value of uniformity in its administrative and judicial understandings, id., at 140.”

- Generally, with citations, whether sentence or clause, try to select the citation span so that what would be left if the citation were deleted would be a grammatical sentence, properly punctuated.

- Next to each citation in the margin of the paper, write a large “S” for a sentence citation and a large “C” for a clause citation.
INTRODUCTION

In March of 2010, after a meeting with plaintiff OGS_TV, INC. (“OGS”), defendant Elizabeth Lime (“Lime”) created a television screenplay entitled “Anticipatory Repudiation.” After the conclusion of the television season, OGS had not incorporated Lime’s screenplay into its television program Lawless Love, despite specifically asking her to create it.

* * *

STATEMENT_OF_FACTS

OGS is the producer of the television program Lawless Love, which depicts the lives of young attorneys. (Pl.’s Compl. ¶ 5). Lime is a veteran television writer of more than fifteen years. (Id. ¶ 11). In March of 2010, OGS made contact with Lime, asking her to create a screenplay for an episode of Lawless Love. (Id. ¶ 9). That day,
Lime met with Dave Nelson (“Nelson”), executive producer of Lawless Love, to work out the details of a possible agreement. *(Id. ¶ 12).*

* * *

**STANDARD OF REVIEW**

In adjudicating a Fed. R. Civ. P. 12(b)(6) motion to dismiss, courts can only allow complaints to survive if they “contain sufficient factual matter, accepted as true, to state a claim to relief that is plausible on its face.” Ashcroft v. Iqbal, 556 U.S. 662, 129 S.Ct. 1937, 1949 (2009) (citing Bell Atl. Corp. v. Twombly, 550 U.S. 544, 570 (2007)). ... This Memorandum asserts that Plaintiff’s claim must be dismissed, as it is not adequately supported by factual allegations to warrant surviving a motion to dismiss.

**ARGUMENT**

I. **Plaintiff May Not Be Afforded Copyright Protection For “Anticipatory Repudiation” Because It Cannot Be Considered A Work Made For Hire Under 17 U.S.C. § 101.**

OGS alleges that Defendants’ creation, performance, and broadcast of “The Love Below” represent infringement of the copyright that OGS supposedly owns in “Anticipatory Repudiation.” ... Defendants assert through several arguments, however, that OGS is not legally afforded copyright ownership of “Anticipatory Repudiation” because the requirements set forth in 17 U.S.C. § 101 were not actually met.

* * *

A. **The plain language of the Copyright Act of 1976 requires an express written agreement be executed before a work is commenced in order for it to be recognized as one made for hire.**

In interpreting statutory language, courts should first look to the plain meaning of the pertinent provisions. The most significant provision of the Act in terms of this
discussion_is_17_U.S.C. § 101, which_lays_out_the_definitions_to_be_considered_for_ copyright_issues. The_statute_provides_that_a_work_made_for_hire_is “a work specially_ ordered_or_commissioned_for_use_as_a_contribution_to_a_collective_work . . . if the_ parties_expressly_agree_in_a_written_instrument_signed_by_them_that_the_work_shall_be_ considered_a_work_made_for_hire.” 17_U.S.C. § 101. Additionally, the_statute_recognizes_that_in_works_made_for_hire, the_person_for_whom_the_work_was_made_is_ considered_the_owner_of_the_copyrign. 17_U.S.C. § 201(b).

* * *

Admittedly, there_is_a_federal_circuit_split_on_the_issue_of_when_the_writing_ requirement_must_be_met. The_Second_Circuit_has_taken_a_more_flexiblestance, stating_ that_if_a_writing_is_executed_after_a_work_has_commenced, it can_be_judged_to_meet_the_ writing_requirement_if_it_confirms_an_earlier_agreement. Playboy_Enters. v. Dumas, 53_ F.3d 549 (2d Cir. 1995). The_Seventh_Circuit, however, takes_a_strict_approach, arguing_ that_a_writing_must_precede_the.commencement_of_the_work_in_order_to_be_considered._ Schiller & Schmidt, Inc. v. Nordisco Corp., 969 F.2d 410 (7th Cir. 1992). The_ Supreme_Court_is_silent_on_this_dispute, as it has “den[jied] certiorari when given_the_ opportunity.” Carolyn_M. Salzmann, You Commissioned It, You Bought It, But Do You_ Own It? The Work for Hire: Why is Something So Simple, So Complicated?, 31_U. Tol. L. Rev. 497, 515 (2000).

The_Seventh_Circuit’s_logic_is_more_compelling, however. In Schiller & Schmidt, 969 F.2d at 410, the_Seventh_Circuit_interpreted_the_whole_of_the_Act_as_requiring_a_ written_agreement_to_be_executed_prior_to_the_creation_of_a_work_made_for_hire_in_order_ to_satisfy_the_statute. In that_case, which_involved_photographs_taken_to_illustrate_an_
office_supply_catalog, a written agreement was signed long after the photographs had been taken and used. _Id._ Citing the necessity of clarifying copyright law, the court stated, “[t]he writing must precede the creation of the property in order to serve its purpose of identifying the (noncreator) owner unequivocally.” _Id._ at 413.

* * *

Additionally, the Seventh Circuit’s view has gained traction in at least one other federal circuit. In _Gladwell Gov’t Servs., Inc. v. Cnty. of Marin_, which dealt with a copyright claim between a county and a contractor, the court acknowledged 17 U.S.C. § 101 by arguing, “[t]he plain language of the statute indicates that a work-for-hire agreement cannot apply to works that are already in existence. Works ‘specially ordered or commissioned’ can only be made after the execution of an express agreement between the parties.” 265 Fed.Appx. 624, 626 (9th Cir. 2008). Specially commissioning a work logically precedes the work itself, so if parties are required to memorialize their agreement of a commission, it must necessarily come before the commissioned work is completed.

It may be easy for one to argue that the more recent decision produced by the _Playboy_ court ought to trump the argument of _Schiller & Schmidt_. It should be noted, however, that the _Schiller & Schmidt_ reasoning has not been forgotten in the years since.