Data Documentation

Computer-aided Text Analysis on News Reporting regarding the United States (USCATA)

The USCATA data set was constructed by Daniela Stockmann, Assistant Professor in the Department of Political Science of Leiden University with the help of Wang Mingde, Cao Manwen, Ai Dan, Cai Jingyi, Zhou Moli, and Andrew Miller. [1] The data set sampled 10 constructed weeks for 1999 and 2003, deriving a sample of 2,280 articles.[2] Constructed-week sampling is a combination of simple random sampling and stratified random sampling. According to this technique, all weeks during a period of time of interest to the researcher are numbered, and subsequently one Monday, Tuesday, and so on is randomly selected until one (or several) weeks are constructed (Stempel, 1952). Constructed week sampling is especially attractive for research that involves multiple years because it relies on a small sample size while at the same time retaining representative results. Ten constructed weeks retain representative results for one year of Chinese news reporting about the United States (Stockmann, 2010a).

To analyze the tone of news reporting in these articles we relied on computer-aided text analysis using Yoshikoder, according to my knowledge the only content analysis software program that can handle Chinese characters. This technique is often criticized for the lack of in-depth and detailed understanding of meaning that qualitative research can extract, its main advantage lies in encouraging transparency and consistency, thus producing replicable, reliable, and generalizable results (Neuendorf, 2002). Elsewhere I have argued that the extraction of qualitative meaning should always be the standard against which to judge the quality of the measurement when using computer-assisted content analysis (Stockmann, 2010a). Measurement of the dictionaries used in computer-aided text analysis was developed based on extensive qualitative reading of texts and pre-tested several times. Below we discuss a number of problems and solutions during data collection, which we may be useful to researchers who intend to use Yoshikoder to analyze texts in Mandarin Chinese.

1. How to improve speed

Usually, new coders first encounter a problem of speed. A beginner roughly takes two hours for analyzing around 15 articles only (e.g. with 18 variables or more). Some are overwhelmed. However, the speed can be increased to about 25 – 30 articles per hour by practice and the application of several tips:

[1] Address correspondence to Daniela Stockmann at dstockmann@fsw.leidenuniv.nl

[2] The Beijing Evening News was sampled from CDroms published by the Beijing Daily Group. The People's Daily was sampled using the Renmin Ribao Archive. In order to exclude tangential articles I did not collect those that only mentioned the United States once in the text.

a. Use “Yoshikoder–0.63–preview.3” version, which allows you to load all articles intended for analysis at once (change “Default encoding” of “Preferences” setting into “UTF-8”; then, one can click “Add Document” to load a large number of articles).

b. Code all cells of a numeral variable as “0” before formally starting analysis. This saves a lot of time, since most variables in each observation, in fact, will be coded as “0” as the result.

c. Change order of variables (column). You may find that keywords measuring some variables appear more frequently than others for a specific content analysis. If the data set contains many variables, move the columns in an Excel spreadsheet of those variables to the beginning of the spreadsheet). After finishing analysis, all changed columns can be restored to the former order (if this is important).

d. You may find it easier to add “reminder columns” before important variables to to temporarily record statistics and make it easier to transfer content from Yoshikoder reports to targeted datasets. Don’t forget to delete the columns after finishing a dataset. Also, assigning important variables different colors allows you to locate variables more quickly.

e. You may find it helpful to first fill in all titles for each individual article before formally starting analysis. That makes some tips above (e.g. tip b) possible.

2. How to deal with inconsistencies between coders

a. Tangential variable: The US time series dataset includes a variable called “tangential,” which measures whether an article only mentions keyword “美国” once (coded as 1). Originally, we assumed that the variable should be rather straightforward. Nevertheless, the results turned out to be pretty different between coders. We found out that the problem was caused by an incorrect segmentation process. For example, Yoshikoder segments 美国华盛顿 as two words: 美国 and 华盛顿. Since both words are under the keyword category of “美国,” the frequency for terms referring to the US is statistically counted as 2. But if one counts the connected words as 2, the number of the same set of positive/negative words surrounding the words in concordance will be multiplied by 2, because the same group of words will be counted twice. During our coding, some noticed the problem and corrected it, but some did not. That caused inconsistencies between coders. We ended up solving it by counting “connected keywords” as single keywords and suggesting a more careful qualitative reading in order to identify “connected words.”

Apart from that, we suggest that once coders find articles that contain two or several connected keywords for the same variable, they should treat them as one word. In contrast, if keywords which measure different variables are connected, they should be counted separately.

b. A similar problem caused by incorrect segmentation is that sometimes Yoshikoder incorrectly recognizes part of a sentence as a match to keywords. For example, keywords of 美国 are not isolated from the sentence containing them. And the problem often ends up with the measurement of "zero" US frequency, which can only be solved by coders' qualitative reading. This problem can largely be solved by means of segmentation of the text. If Yoshikoder did not properly segment from a phrase or sentence, we suggest that coders open the file and add spaces in between the words manually.

c. USCATA includes a variable measuring the number of keywords mentioning 美国. Though this seems straightforward, we still encountered that coding from different coders slightly differed. We figured out that the problem was partially caused by different coding with regard to keywords of美国 contained in news report titles. Some coders didn't count them. We solved the inconsistency problem by including all keyword matches in the titles.

d. "Missing articles": When we first checked the data set for completeness, we found that some articles appeared to be missing. Later we found that apparently missing articles were given different titles assigned to the same article, which were coded twice by different coders. We solve the problem by changing all those articles titles into the same ones.

e. It is helpful to have different coders overlap in their coding using Yoshikoder to check consistency between coders. If we found inconsistencies, we asked coders to inquire about the root of the problem and made sure that rules were consistency applied across coders.

f. USCATA also includes variables that measure the article length and whether the articles constituted a Xinhua article "Xinhua." We also detected inconsistencies between coders when coding these variables. Reasons for such mistakes were mostly accidental. But a few errors regarding variable "Xinhua" were caused by the segmentation mistakes of Yoshikoder. For keyword 新华 was sometimes not properly segmented and Yoshikoder could not detect it. The problem can only be solved by qualitative reading by coders. To avoid inconsistencies, coders were not longer asked to code the variable "xinhuanum," which traces sensitivity over time (Stockmann, 2010b). Coding of this variable for all observations was left to one coder to deal with after the collection of results by all coders.

Bibliography

Neuendorf, K. A. (2002). *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage Publications.

Stempel, G. H. (1952). Sample Size for Classifying Subject Matter in Dailies. *Journalism Quarterly, 29*(3), 333-334.

Stockmann, D. (2010a). Information Overload? Collecting, Managing, and Analyzing Chinese Media Content. In A. Carlson, M. Gallagher & M. Manion (Eds.), *Sources and Methods in Chinese Politics*. New York: Cambridge University Press.

Stockmann, D. (2010b). Information Overload? Collecting, Managing, and Analyzing Chinese Media Content. In A. Carlson, M. Gallagher & M. Manion (Eds.), *Sources and Methods in Chinese Politics* (pp. 107-125). New York: Cambridge University Press.