Using MTurk to Improve Content Analyses

Content Analysis PDW

Tim Hubbard, Ph.D.

UNIVERSITY OF NOTRE DAME
Mendoza College of Business

You just collected thousands of media articles about a sample of firms…

How do you feel confident they are actually about the firm?

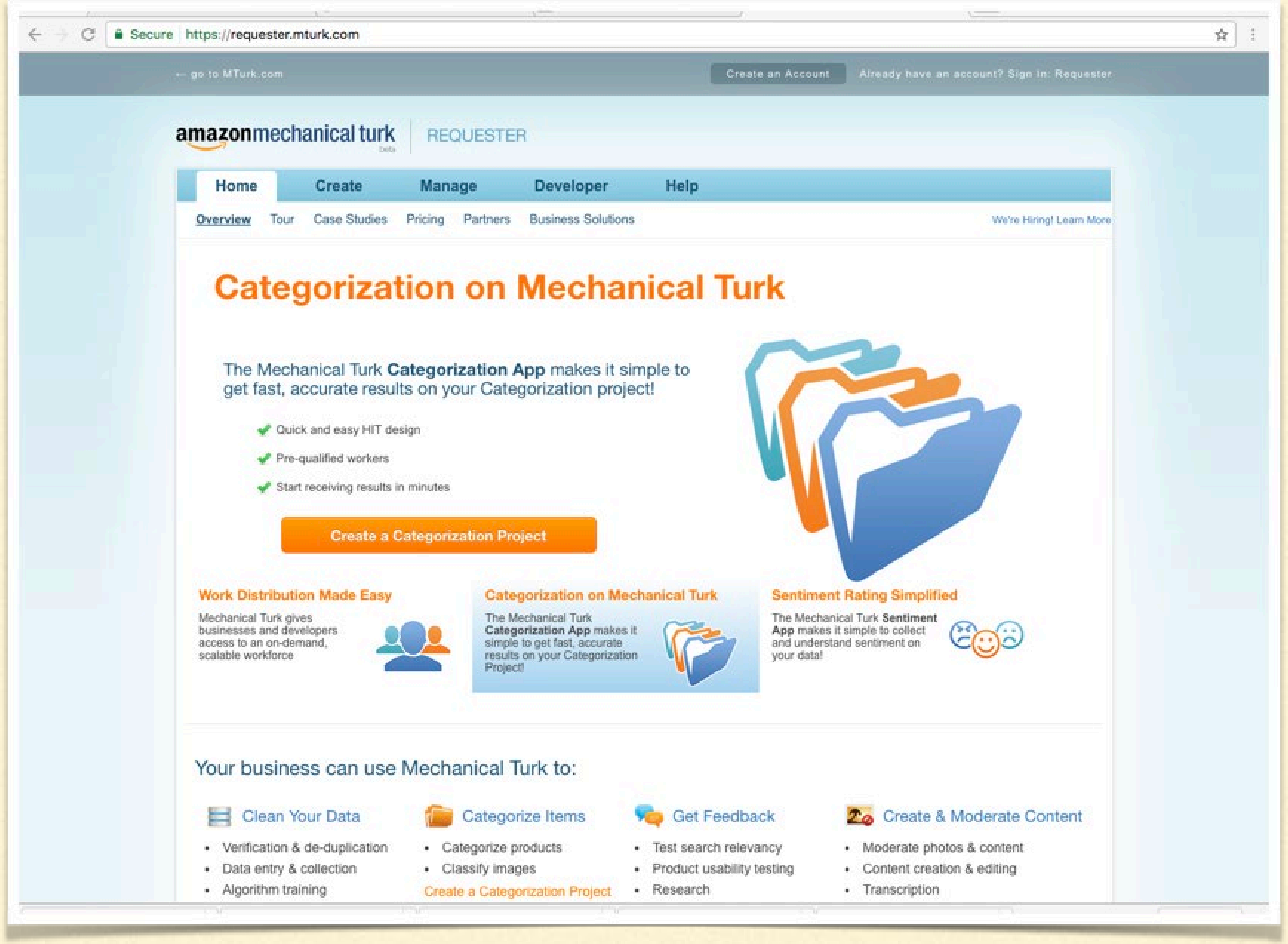
Why Should We Care?

- Stronger connection between construct and proxy
- Reduces measurement error
- Increases the likelihood of finding results
- Thoroughness of independent ratings are more valid and easy to report

Our Situation

- Study of celebrity startup firms
- Reviewer requested we collect more media data
- Needed to analyze general media’s tenor, volume, and non-conformity of sample firms
- Needed to do it thoroughly—and fast

Amazon’s MTurk Helped



Categorization on Mechanical Turk

The Mechanical Turk **Categorization App** makes it simple to get fast, accurate results on your Categorization project!

- ✓ Quick and easy HIT design
- ✓ Pre-qualified workers
- ✓ Start receiving results in minutes

Work Distribution Made Easy

Mechanical Turk gives businesses and developers access to an on-demand, scalable workforce

Categorization on Mechanical Turk

The Mechanical Turk **Categorization App** makes it simple to get fast, accurate results on your Categorization Project!

Sentiment Rating Simplified

The Mechanical Turk **Sentiment App** makes it simple to collect and understand sentiment on your data!

Your business can use Mechanical Turk to:

Clean Your Data
- Verification & de-duplication
- Data entry & collection
- Algorithm training

Categorize Items
- Categorize products
- Classify images

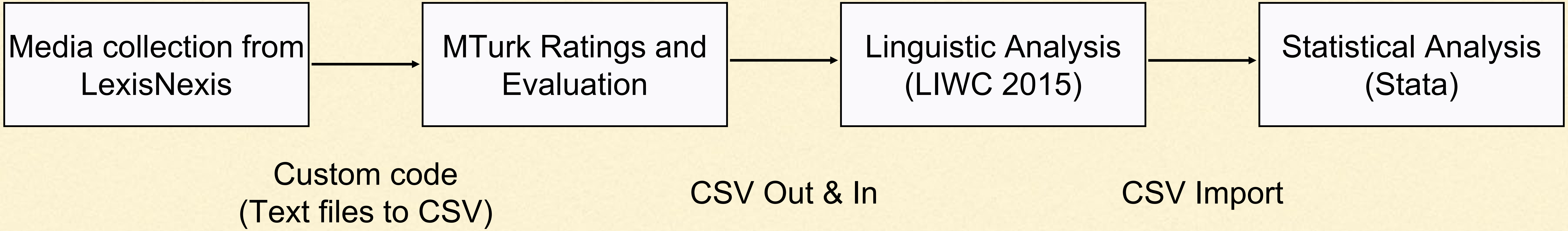
Create a Categorization Project

Get Feedback
- Test search relevancy
- Product usability testing
- Research

Create & Moderate Content
- Moderate photos & content
- Content creation & editing
- Transcription

- Human coding
- Gives access to multiple individual raters per article
- Very fast
- Imports data using CSV files
- Easy payment

General Process



Media collection from LexisNexis

MTurk Ratings and Evaluation

Linguistic Analysis (LIWC 2015)

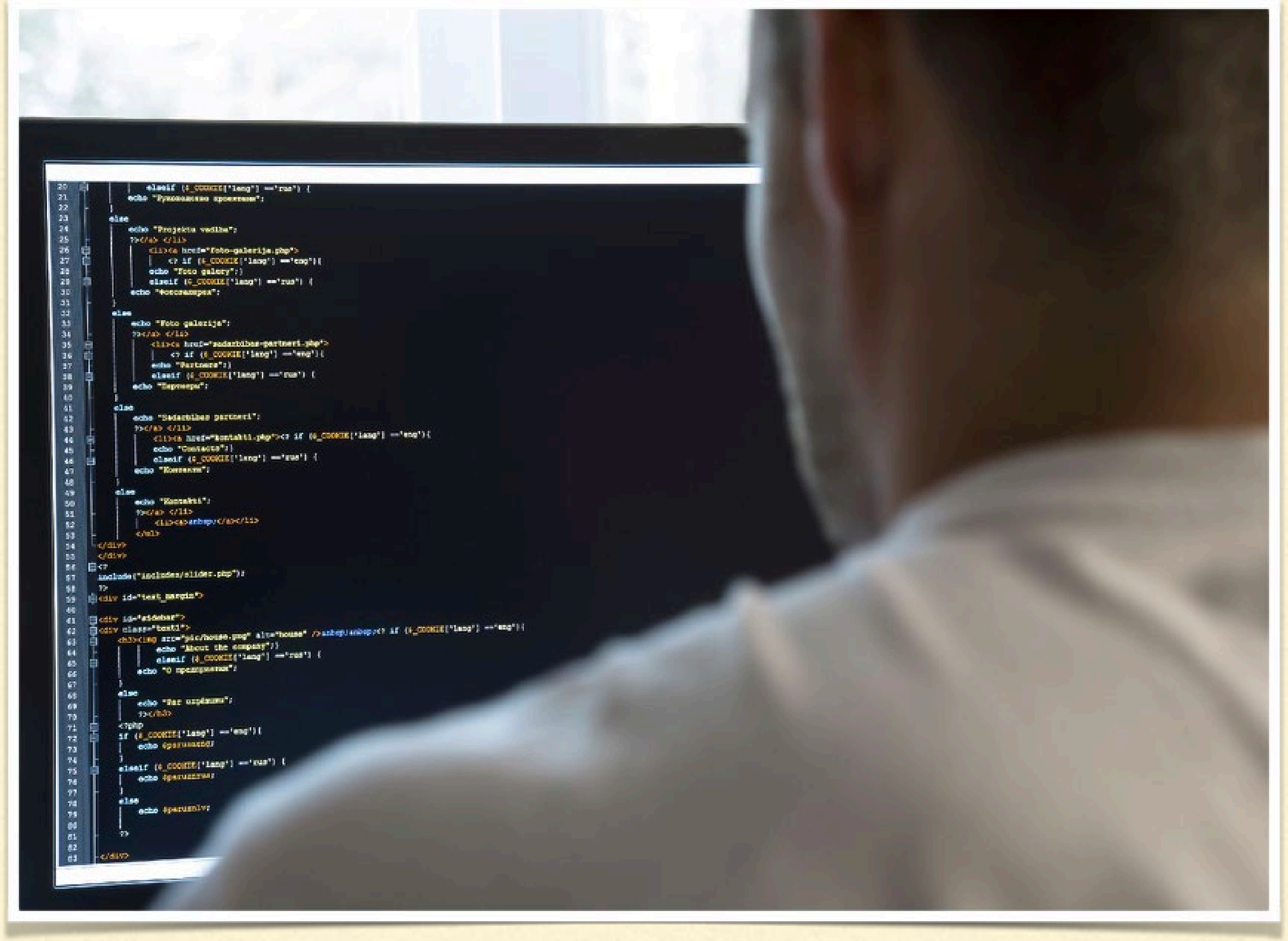
Statistical Analysis (Stata)

Custom code
(Text files to CSV)

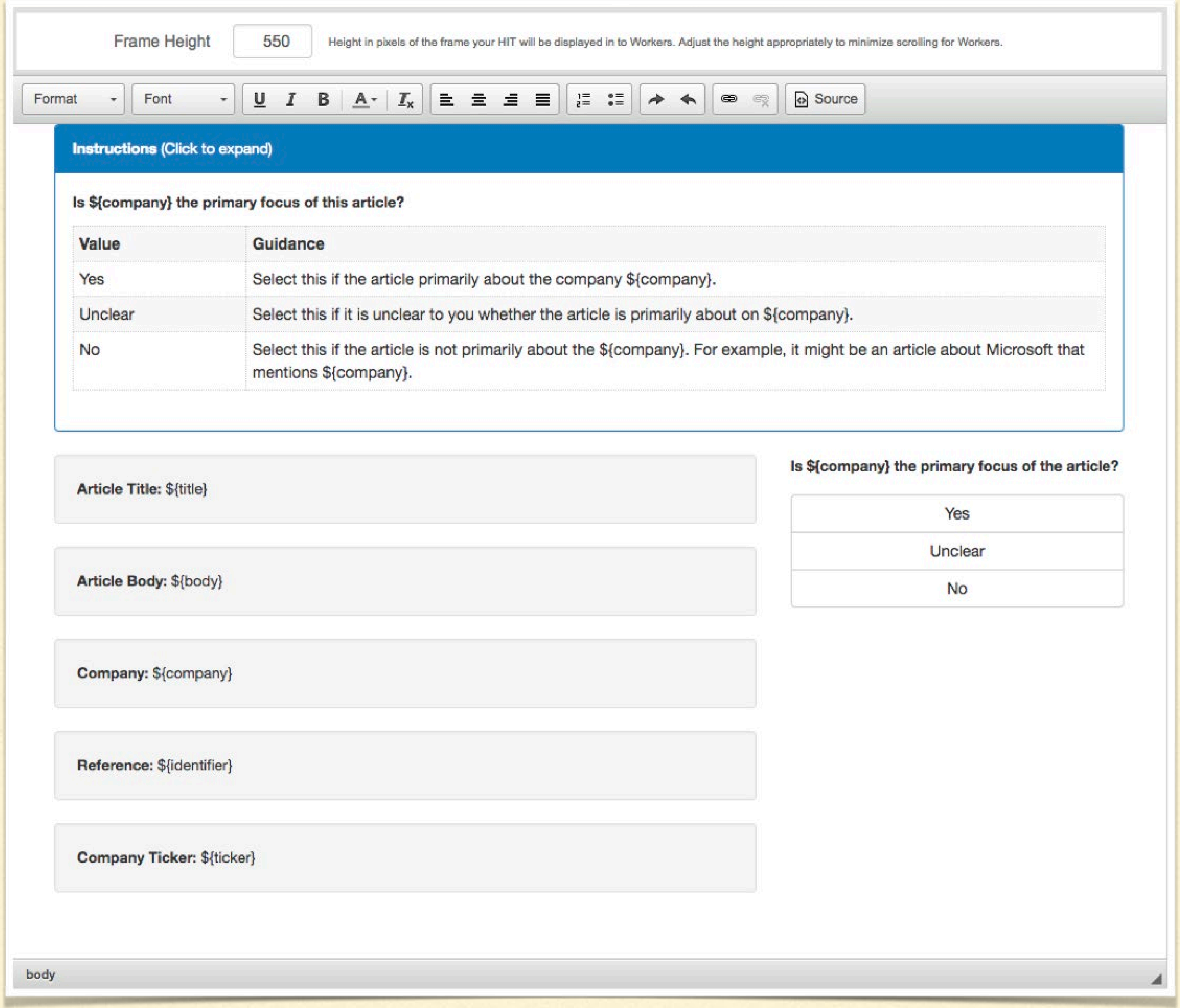
CSV Out & In

CSV Import

Parameters



- 6,260 articles from LexisNexis
- Presented only the first page of the articles
- 2 raters per article (12,520 ratings)
  - Disagreement or unclear, added one more (795)
- 7 hours per run
- $0.08/rating



Frame Height 550 Height in pixels of the frame your HIT will be displayed in to Workers. Adjust the height appropriately to minimize scrolling for Workers.

Instructions (Click to expand)

Is ${company} the primary focus of this article?

Value	Guidance
Yes	Select this if the article primarily about the company ${company}.
Unclear	Select this if it is unclear to you whether the article is primarily about on ${company}.
No	Select this if the article is not primarily about the ${company}. For example, it might be an article about Microsoft that mentions ${company}.

Article Title: ${title}

Article Body: ${body}

Company: ${company}

Reference: ${identifier}

Company Ticker: ${ticker}

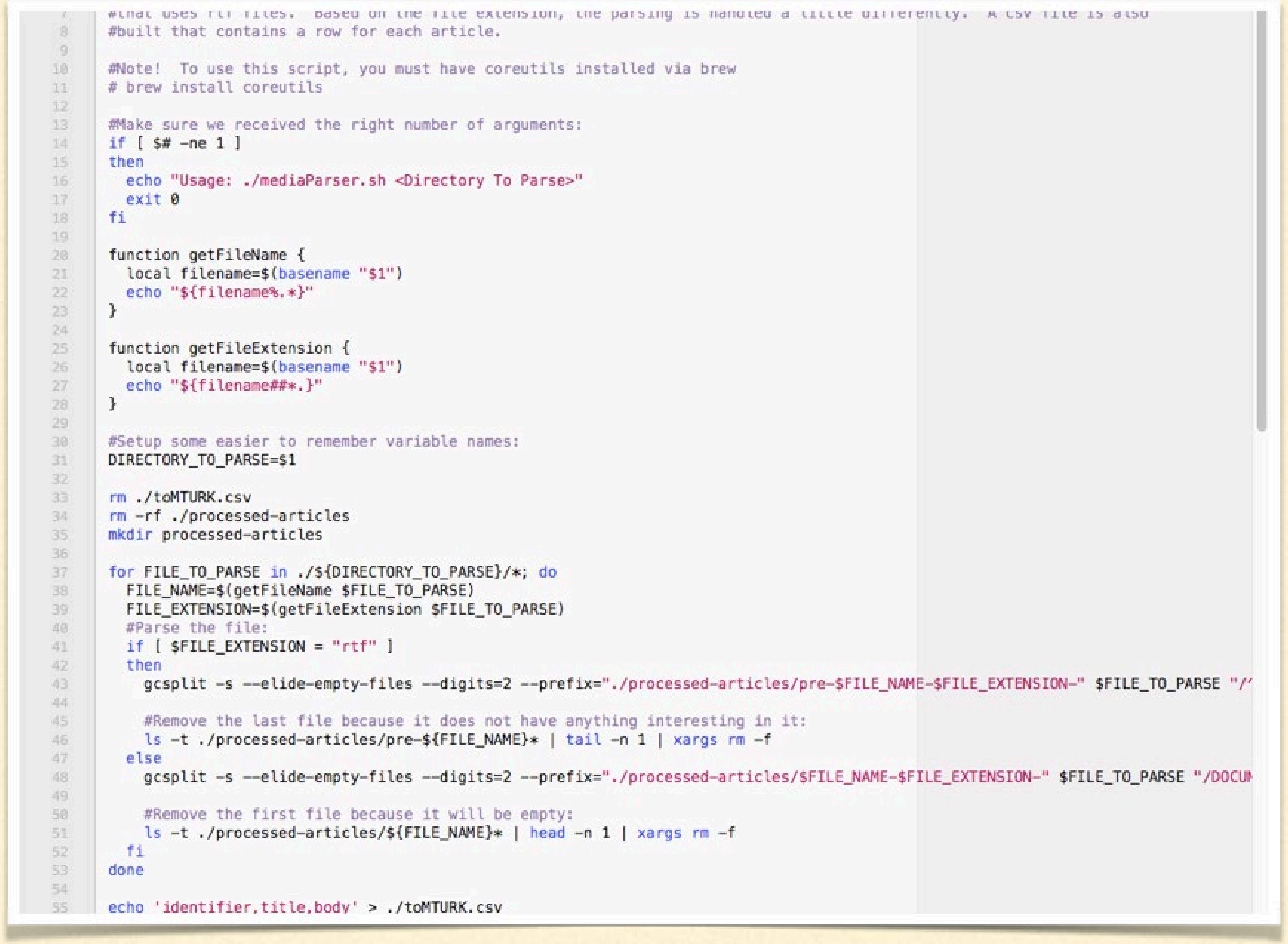
Is ${company} the primary focus of the article?

Yes

Unclear

No

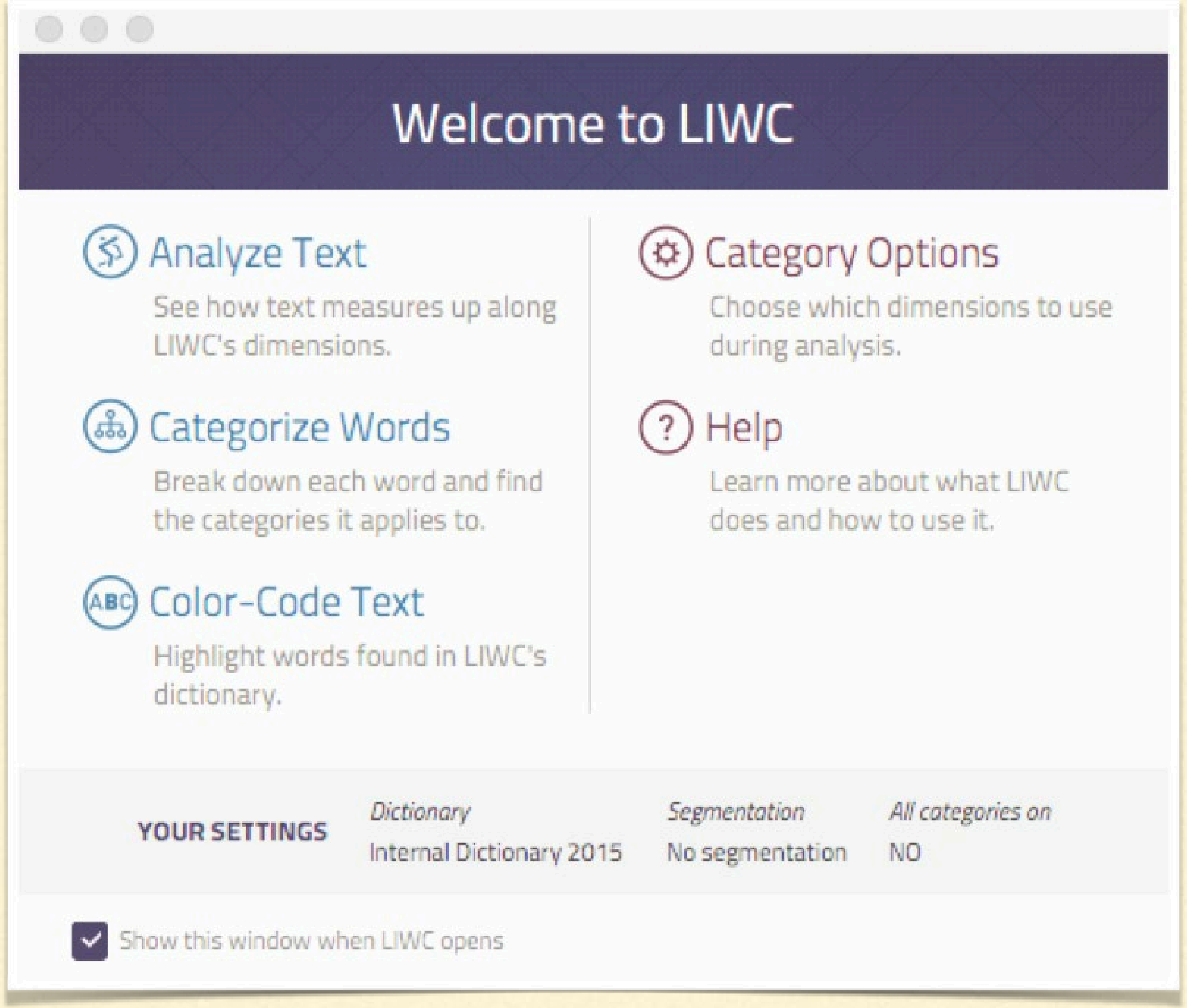
Moving from LexisNexis to MTurk



```
#that uses rtf files.  Based on the file extension, the parsing is handled a little differently.  A csv file is also
#built that contains a row for each article.

#Note!  To use this script, you must have coreutils installed via brew
# brew install coreutils

#Make sure we received the right number of arguments:
if [ $# -ne 1 ]
then
  echo "Usage: ./mediaParser.sh <Directory To Parse>"
  exit 0
fi

function getFileName {
  local filename=$(basename "$1")
  echo "${filename%.*}"
}

function getFileExtension {
  local filename=$(basename "$1")
  echo "${filename##*.}"
}

#Setup some easier to remember variable names:
DIRECTORY_TO_PARSE=$1

rm ./toMTURK.csv
rm -rf ./processed-articles
mkdir processed-articles

for FILE_TO_PARSE in ./${DIRECTORY_TO_PARSE}/*; do
  FILE_NAME=$(getFileName $FILE_TO_PARSE)
  FILE_EXTENSION=$(getFileExtension $FILE_TO_PARSE)
  #Parse the file:
  if [ $FILE_EXTENSION = "rtf" ]
  then
    gcsplit -s --elide-empty-files --digits=2 --prefix="./processed-articles/pre-$FILE_NAME-$FILE_EXTENSION-" $FILE_TO_PARSE "/'

    #Remove the last file because it does not have anything interesting in it:
    ls -t ./processed-articles/pre-${FILE_NAME}* | tail -n 1 | xargs rm -f
  else
    gcsplit -s --elide-empty-files --digits=2 --prefix="./processed-articles/$FILE_NAME-$FILE_EXTENSION-" $FILE_TO_PARSE "/DOCUM

    #Remove the first file because it will be empty:
    ls -t ./processed-articles/${FILE_NAME}* | head -n 1 | xargs rm -f
  fi
done

echo 'identifier,title,body' > ./toMTURK.csv
```

- Custom developed software to move from LexisNexis text files to a fully populated CSV file
- That file can then go up to MTurk
- Developed by Rampant Strategy OÜ

LIWC 2015



Welcome to LIWC

Analyze Text

See how text measures up along LIWC's dimensions.

Categorize Words

Break down each word and find the categories it applies to.

Color-Code Text

Highlight words found in LIWC's dictionary.

Category Options

Choose which dimensions to use during analysis.

Help

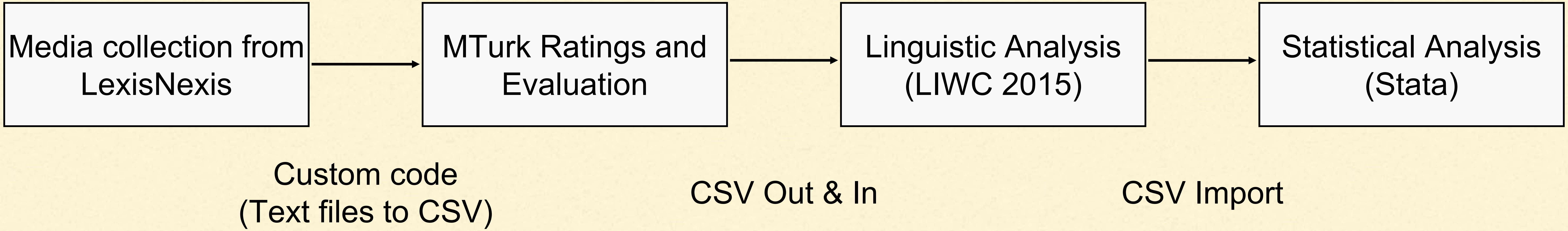
Learn more about what LIWC does and how to use it.

YOUR SETTINGS	*Dictionary*	*Segmentation*	*All categories on*
	Internal Dictionary 2015	No segmentation	NO

☑ Show this window when LIWC opens

- Can import CSV files now
  - Straight from MTurk
- Can export to CSV which is easy to import into Stata

General Process



Media collection from LexisNexis

MTurk Ratings and Evaluation

Linguistic Analysis (LIWC 2015)

Statistical Analysis (Stata)

Custom code
(Text files to CSV)

CSV Out & In

CSV Import

Questions & Discussion

thubbard@nd.edu