

# Gender Annotator

## Table of contents

Two approaches to gender annotation .....	1
How is this different from <i>Gender Guesser</i> ? .....	1
Stanford CoreNLP gender annotator .....	1
Dictionary-based approaches .....	2
US Census names .....	2
US Social Security .....	3
Carnegie Mellon .....	3
Proper noun first names .....	4
Personal pronouns .....	4
INPUT .....	4
OUTPUT .....	4
References .....	4

## Two approaches to gender annotation

The NLP Suite takes two different approaches to annotating the gender of proper names mentioned in a text. Thus, if “John” is mentioned, is “John” a male or female first name? One approach is based on the gender annotator of Stanford CoreNLP. The other uses different databases where first names are tagged as male or female first names.

## How is this different from *Gender Guesser*?

*Gender Guesser* is an online tool (<http://www.hackerfactor.com/GenderGuesser.php>) that aims at identifying the gender of the author of a text (see Argamon et al. 2003). The tools described here aim at identifying the gender of an individual mentioned in a text by its proper name (e.g., John as a male name and Alice as a female name).

## Stanford CoreNLP gender annotator

The algorithm uses the Stanford CoreNLP gender annotator, part of the Stanford CoreNLP toolkit (Manning et al. 2014).

CoreNLP relies on its own name list file, tmp-stanford-models-expanded/edu/stanford/nlp/models/gender/first\_name\_map\_small, contained in the stanford-corenlp-3.5.2-models.jar file. This file can be extracted for editing. You can extract the default gender file from that jar in this manner:

- mkdir tmp-stanford-models-expanded
- cp /path/of/stanford-corenlp-3.5.2-models.jar tmp-stanford-models-expanded
- cd tmp-stanford-models-expanded
- jar xf stanford-corenlp-3.5.2-models.jar

- there should now be tmp-stanford-models-expanded/edu

You can also make your own custom gender file. The file should be in this format:

JOHN\MALE

With one NAME\GENDER entry per line.

## Dictionary-based approaches

The lib subdirectory of the NLP Suite comes with a number of csv files that contain people's first names (e.g., John) with their gender (e.g., male). Some of these files provide probabilistic measures about the gender attribution of a name (i.e., how likely is John to be male name? How about Jamie?).

For more name lists not explicitly used here (e.g., African American names), see <http://mbejda.github.io/>

## *US Census names*

The 1990 United States Census provided data on the distribution of first names in America, with statistics about the likelihood of a first name being male or female.

[https://www.census.gov/topics/population/genealogy/data/1990\\_census/1990\\_census\\_namefiles.html](https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html)

The 1990 census distributed three files: dist.female.first, dist.female.first, and dist.all.last for last names.

For every name listed, the files provide four items of data:

A "Name"  
Frequency in percent  
Cumulative Frequency in percent  
Rank

For example, in the file (dist.all.last) one entry appears as:

MOORE      0.312   5.312      9

What do these values mean? Let's unpack the information:

the last name MOORE (Name) is possessed by 0.312 percent of our population sample (Frequency in percent);

5.312 percent of the sample population is covered by MOORE and the other 8 names occurring more frequently than MOORE (Cumulative Frequency in percent);

the last name MOORE ranks 9th in terms of frequency (Rank).

For a methodological explanation of the US Census gender files, see



















[https://www2.census.gov/topics/genealogy/1990surnames/nam\\_meth.txt?#](https://www2.census.gov/topics/genealogy/1990surnames/nam_meth.txt?#)

## *US Social Security*

The United States Social Security provides three different files with thousands of first names classified as F (Female), M (Male) and the frequency of occurrences of the name. The SS Readme file states: “Each file is sorted first on sex and then on number of occurrences in descending order. When there is a tie on the number of occurrences, names are listed in alphabetical order. This sorting makes it easy to determine a name's rank. The first record for each sex has rank 1, the second record for each sex has rank 2, and so forth. To safeguard privacy, we restrict our list of names to those with at least 5 occurrences.”

<https://www.ssa.gov/oact/babynames/limits.html>

In the SS database, names are classified by year (year of birth, yob) giving users the ability to identify the popularity of certain names over time.

Name	Type	Compressed size
 NationalReadMe.pdf	Adobe Acrobat Document	225 KB
 yob1880.txt	TXT File	9 KB
 yob1881.txt	TXT File	8 KB
 yob1882.txt	TXT File	9 KB
 yob1883.txt	TXT File	9 KB
 yob1884.txt	TXT File	10 KB
 yob1885.txt	TXT File	10 KB
 yob1886.txt	TXT File	10 KB
 yob1887.txt	TXT File	10 KB
 yob1888.txt	TXT File	11 KB
 yob1889.txt	TXT File	11 KB
 yob1890.txt	TXT File	11 KB
 yob1891.txt	TXT File	11 KB
 yob1892.txt	TXT File	12 KB
 yob1893.txt	TXT File	12 KB
 yob1894.txt	TXT File	12 KB
 yob1895.txt	TXT File	13 KB
 yob1896.txt	TXT File	13 KB

The current (June 2020) list goes all the way up to yob2018.txt

## *Carnegie Mellon*

Mark Kantrowitz, at Carnegie Mellon, developed a list of first names by gender (<http://www.cgi.cs.cmu.edu/Groups/AI/util/areas/nlp/corpora/names/0.html>): 4987 female names and 2940 male names.

### ***Proper noun first names***

The algorithm uses the **Stanford CoreNLP NER annotator** to extract every **PERSON** tag. It then takes the lemma value of each word tagged (e.g., John) and matches the word lemma with an entry in a selected dictionary to find its gender (e.g., John, male).

### ***Personal pronouns***

The algorithm also marks every instance of the word “he,” “him,” “his,” and “she,” “her,” “hers” and annotates them, respectively, as male and female.

### **INPUT**

The algorithm can either process a single file or all the txt files in a directory.

### **OUTPUT**

### **References**

- Argamon, Shlomo, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. “Gender, Genre, and Writing Style in Formal Written Texts,” *Text*, Vol. 23, No. 3, pp. 321–346.
- Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, David McClosky. 2014. “The Stanford CoreNLP Natural Language Processing Toolkit.” *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.