Multidimensional Analysis Tagger (v. 1.1) – Manual

The Multidimensional Analysis Tagger (MAT) is a program for Windows that replicates Biber's (1988) tagger for the multidimensional functional analysis of English texts, generally applied for studies on text type or genre variation. The program generates a grammatically annotated version of the corpus or text selected as well as the statistics needed to perform a text-type or genre analysis. The program plots the input text or corpus on Biber's (1988) Dimensions and it determines its closest text type, as proposed by Biber (1989). Finally, the program offers a tool for visualising the Dimensions features of an input text.

This is an implementation of the tagger used in Biber (1988) and in many other works. This tagger tries to replicate the analysis in Biber (1988) as closely as possible by taking into account the algorithms that the author presented in the Appendix of the book. The basic analysis of the text is done through the Stanford Tagger. The present tagger includes a copy of the Stanford Tagger (2013) which is run automatically to produce a preliminary grammatical analysis. MAT then expands the Stanford Tagger tag set by identifying the linguistic features used in Biber (1988).

This document includes an extensive description of the tagger as well as some instructions for the user.

Referencing the tagger

To reference the tagger, please use the following:

Nini, A. 2014. *Multidimensional Analysis Tagger 1.1 - Manual*. Retrieved from: http://sites.google.com/site/multidimensionaltagger.

This program is based on the Stanford Tagger and it is therefore necessary to reference the Stanford Tagger any time the program is used. To reference the Stanford Tagger, please refer to the Stanford Tagger website: http://nlp.stanford.edu/software/tagger.shtml.

Architecture of the program

Requirements: the program requires Java to run. This can be downloaded from http://java.com/en/download/index.jsp

Tagger

This module of the program accepts as input only plain text files in the format '.txt'. The user can select either a folder of .txt files or a single .txt file.

MAT tagger uses the Stanford Tagger for an initial segmentation in parts of speech and then finds the patterns described in Biber (1988). Some basic Stanford Tagger tags are replaced by new tags that are more specific. For example, negations and prepositions are distinguished, respectively, from general adverbs and general subordinators. The word *to* used as an infinitive marker is disambiguated from the word *to* used as a preposition. Three tags are added in order to facilitate the identification of Biber's (1988) linguistic features, these are: (1) indefinite pronouns (INPR): *anybody, anyone, anything, everybody, everyone, everything, nobody, none, nothing, nowhere, somebody, someone, something*; (2) quantifiers (QUAN): *each, all, every, many, much, few, several, some, any*; (3) quantifier pronouns (QUPR): *everybody, somebody, anybody, everyone, someone, anyone, everything, something, anything*. A full list of tags and a description of the algorithms used to find them is given below.

The Stanford tagged texts will appear in a folder called 'ST_name_of_folder' or 'ST_name_of_file'. The MAT tagged texts will appear in a folder called 'MAT_name_of_folder' or 'MAT_name_of_text'. Both folders will be created in the folder selected for the analysis.

When the tagger is launched, a module of the tagger will check the encoding of the .txt files selected. The tagger will then flag any text in UNICODE and it is up to the user to change this to a compatible format, such as ANSI or UTF-8.

After this stage, the tagger will scan each of the .txt files in order to find instances of curly inverted commas. This step is necessary as otherwise some contractions are not tagged properly. If the tagger finds any instance of curly commas it will replace them with standard commas. This will overwrite the file, so the original .txt file with the curly commas will be lost. If it is necessary to keep the original with curly commas then it is recommended to create a backup copy before running MAT.

Analyser

This module of the program can be called either via the 'Analyse' button or via the 'Tag and Analyse' button. When this module starts, the user will be asked to input the number of tokens for which the type-token ratio should be calculated (for details see the entry on type-token ratio in the list of variables). By default, this number is 400, as set in Biber (1988). The user will then asked to choose which Dimensions to display graphically. The result of the analysis consists of a number of output files that will be created in a folder called 'Statistics' contained in the same folder that contains the MAT tagged texts. These files are:

1) 'Corpus_Statistics.txt': a tab delimited file that shows the frequency per 100 tokens for all the linguistic variables (see below) found in the input text or corpus. If the user selects the option 'all tags', then this file will display the counts for all the tags in the text, including the punctuation items. On the other hand, if the user selects the option 'only VASW tags', then only the tags used in Biber (1988) will be displayed.
2) 'Zscores.txt': a tab delimited file that includes the z-scores of the linguistic variables for the input file or corpus. If the user has selected a folder of text files as input, then the averages for the corpus are showed. The z-scores are calculated on the basis of the means and standard deviations presented in Biber (1988: 77). For each text and for the corpus as a whole, the program will flag all the z-scores with a magnitude higher than 2 as 'Interesting variables'. The z-scores displayed in this file are not affected by the user's selection of the z-score correction. The option 'z-score correction' affects only the calculation of the Dimension scores.
3) 'Dimensions.txt': a tab delimited file that contains the scores for the Dimensions as well as the averages for the corpus, if the user has selected a folder of text files. The Dimension scores are calculated using the z-scores of the variables that presented a mean higher than 1 in the chart presented in Biber (1988: 77). The reliability of the Dimension scores produced by MAT was checked against the LOB and the Brown corpus. The results of the tests are presented below. The program classifies each text according to its closer text type as proposed by Biber (1989) using Euclidean distance. If the user has selected as input a folder of texts, then the averages for the corpus are provided. If the user has chosen to use the z-score correction, then these Dimension score reflect the choice. When the user selects to use the z-score correction, all the z-scores used to calculate the

Dimension scores are first checked for their magnitude. If the absolute value of the magnitude is higher than 5, the program will change it to 5. This correction avoids the problem of few infrequent variables affecting the overall Dimension scores. This option should be used with caution and it is particularly advised only for very short texts.

4) 'Dimension#.png': a graph that displays the location of the input text's Dimension score compared to a number of genres as shown in Biber (1988: 172). The graph displays the mean and the range for each genre. If the user has selected as input only one text, then the Dimension score for that text is shown. On the other hand, if the user has selected a corpus as input, then the mean and the range for that corpus are displayed. The program will print the closest genre to the user's text or corpus next to the title of the graph. MAT produces as many Dimension graphs as the user has selected.
5) 'Text_types.png': a graph representing the location of the analysed text or corpus in relation to Biber's (1989) eight text types. The program will print the closest text type to the user's text or corpus next to the title of the graph. Text types are assigned using Euclidean distance.

Inspect tool

This tool allows the user to display the Dimension features of a single text. The user can choose which Dimensions to visualise. Once the tool is used, a new file named 'FILENAME_features.html' will be created in the folder where the selected text is located. This tool can be used only with MAT tagged texts.

Reliability tests for the program

The program was tested for reliability on the LOB and on the Brown corpus. These results are reproduced below.

Table 1 – MAT analysis of the LOB corpus compared to Biber's (1988) results

	D1	D2	D3	D4	D5	D6	
Press reportage - MAT	-14.02	0.97	2.81	-0.38	0.52	-0.72	59% General narrative exposition; 39% Learned exposition; 2% Involved persuasion; 2% Scientific exposition
Press reportage - Biber (1988)	-15.01	0.4	-0.3	-0.7	0.6	-0.9	73% General narrative exposition; 25% Learned exposition; 2% Scientific exposition
Difference	0.99	0.57	3.11	0.32	0.08	0.18	
Press editorials - MAT	-8.4	-0.28	4.38	3.3	1.5	0.33	81% General narrative exposition; 7% Involved persuasion; 7% Scientific exposition; 4% Learned exposition
Press editorials - Biber (1988)	-10	-0.8	1.9	3.1	0.3	1.5	86% General narrative exposition; 11% Involved persuasion; 4% Learned exposition
Difference	1.6	0.52	2.48	0.2	1.2	1.17	
Press reviews - MAT	-12.45	-0.74	5.38	-2.32	0.36	-1.01	53% General narrative exposition; 47% Learned exposition
Press reviews - Biber (1988)	-13.9	-1.6	4.3	-2.8	0.8	-1	47% Learned exposition; 47% General narrative exposition; 6% Scientific exposition
Difference	1.45	0.86	1.08	0.48	0.44	0.01	
Religion - MAT	-4.26	0.17	4.69	0.85	2.22	1.01	65% General narrative exposition; 29% Involved persuasion; 6% Scientific exposition
Religion - Biber (1988)	-7	-0.7	3.7	0.2	1.4	1	59% General narrative exposition; 18% Involved persuasion; 18% Learned exposition; 6% Imaginative narrative
Difference	2.74	0.87	0.99	0.65	0.82	0.01	
Hobbies - MAT	-9.42	-2.1	3.15	1.51	2.54	-0.35	34% General narrative exposition; 24% Learned exposition; 24% Involved persuasion; 18% Scientific exposition
Hobbies - Biber (1988)	-10.1	-2.9	0.3	1.7	1.2	-0.7	43% General narrative exposition; 21% Learned exposition; 21% Involved persuasion; 7% Scientific exposition; 7% Situated reportage
Difference	0.68	0.8	2.85	0.19	1.34	0.35	
Popular lore - MAT	-9.58	0.31	3.42	-0.61	1.4	-0.64	36% Learned exposition; 32% General narrative exposition; 20% Involved persuasion; 2% Imaginative narrative; 9% Scientific exposition
Popular lore - Biber (1988)	-9.3	-0.1	2.3	-0.3	0.1	-0.8	36% Learned exposition; 36% Involved persuasion; 21% General narrative exposition; 7% Imaginative narrative
Difference	0.28	0.41	1.12	0.31	1.3	0.16	

Academic prose - MAT	-12.16	-2.16	5.38	-0.02	5.14	0.23	56% Scientific exposition; 24% Learned exposition; 14% General narrative exposition; 6% Involved persuasion
Academic prose - Biber (1988)	-14.09	-2.6	4.2	-0.5	5.5	0.5	44% Scientific exposition; 31% Learned exposition; 17% General narrative exposition; 9% Involved persuasion
Difference	1.93	0.44	1.18	0.48	0.36	0.27	
General fiction - MAT	0.35	6.26	0.03	1.79	-0.45	-0.75	55% Imaginative narrative; 31% General narrative exposition; 10% Involved persuasion; 3% Learned exposition
General fiction - Biber (1988)	-0.8	5.9	-3.1	0.9	-2.5	-1.6	51% Imaginative narrative; 41% General narrative exposition; 3% Informational interaction; 3% Involved persuasion
Difference	1.15	0.36	3.13	0.89	2.05	0.85	
Mystery fiction - MAT	0.82	5.76	-0.7	1.55	-0.69	-1.13	67% Imaginative narrative; 29% General narrative exposition; 4% Involved persuasion
Mystery fiction - Biber (1988)	-0.2	6	-3.6	-0.7	-2.8	-1.9	70% Imaginative narrative; 23% General narrative exposition; 8% Situated reportage
Difference	1.02	0.24	2.9	2.25	2.11	0.77	
Science fiction - MAT	-5.01	6.1	1.08	0.21	-0.54	-0.54	83% General narrative exposition; 17% Imaginative narrative
Science fiction - Biber (1988)	-6.1	5.9	-1.4	-0.7	-2.5	-1.6	50% General narrative exposition; 33% Imaginative narrative; 17% Situated reportage
Difference	1.09	0.2	2.48	0.91	1.96	1.06	
Adventure fiction - MAT	-0.85	5.89	-1.29	0.19	-0.97	-1.29	69% Imaginative narrative; 24% General narrative exposition; 3% Involved persuasion; 3% Learned exposition
Adventure fiction - Biber (1988)	0	5.5	-3.8	-1.2	-2.5	-1.9	70% Imaginative narrative; 31% General narrative exposition
Difference	0.85	0.39	2.51	1.39	1.53	0.61	
Romantic fiction - MAT	3.55	6.71	-0.88	2.35	-1.26	-1	79% Imaginative narrative; 17% General narrative exposition; 3% Involved persuasion
Romantic fiction - Biber (1988)	4.3	7.2	-4.1	1.8	-3.1	-1.2	92% Imaginative narrative; 8% General narrative exposition
Difference	0.75	0.49	3.22	0.55	1.84	0.2	

Humour - MAT	-6.19	1.43	1.62	0.43	0.65	-0.56	78% General narrative exposition; 11% Imaginative narrative; 11% Involved persuasion
Humour - Biber (1988)	-7.8	0.9	-0.8	-0.3	-0.4	-1.5	89% General narrative exposition; 11% Involved persuasion
Difference	1.61	0.53	2.42	0.73	1.05	0.94	

The scores obtained by MAT for the Dimensions show that MAT is largely successful in replicating Biber's (1988) analysis.

For Dimension 1, the difference ranges from a minimum of 0.28 for Popular Lore to a maximum of 2.74 for Religion. However, given the wide span of Dimension 1 scores, even a difference of 3 does still correctly locate the text in the right area of Dimension 1.

For Dimension 2, the difference ranges from a minimum of 0.2 for Science Fiction to a maximum of 0.87 for Religion. This difference of less than a point is not enough to cause any significant difference in terms of text type assignation and/or location of the analysed text(s) along Dimension 2.

For Dimension 3, the difference ranges from a minimum of 0.99 for Religion to a maximum of 3.22 for Romantic Fiction. Given the limited range of Dimension 3, differences of magnitude 2 or more can create some problems in the reliability of MAT Dimension 3 scores.

For Dimension 4, the differences range from a minimum of 0.19 for Hobbies to a maximum of 2.25 for Mystery Fiction. Apart from this value, all other values show that there are no large differences between Biber's (1988) scores and MAT's.

For Dimension 5, the differences range from a minimum of 0.08 for Press Reportage to a maximum of 2.11 for Mystery Fiction. Apart from this value, all other values show that there are no large differences between Biber's (1988) scores and MAT's.

Finally, for Dimension 6, the differences range from a minimum of 0.01 for Press Reviews and Religion to a maximum of 1.06 for Science Fiction, confirming that there are no large differences between Biber's (1988) scores and MAT's analysis.

In general, therefore, it is possible to conclude that MAT performs well in replicating Biber's (1988) study. The only anomalous scores are the ones obtained for Dimension 3. An exploration of the z-scores pointed out that the scores produced by MAT for Dimension 3 are inflated because of high z-scores of general adverbs. However, to this stage no cause was individuated as being responsible for this variation. Until the problem is resolved, Dimension 3 scores produced by MAT should be treated with caution. Although the differences for Dimension 3 are moderate, these do not influence the assignation of the text type in many cases, since most of the genres are unmarked for Dimension 3.

The assignation of text types given by MAT are generally accurate with some small inaccuracies probably caused by the small differences between the dictionaries or rules employed by Stanford Tagger and the tagger used in Biber (1988).

Another test was run for the Brown corpus and the results are presented below.

Table 2 - MAT analysis of the Brown corpus compared to Biber's (1988) results

	D1	D2	D3	D4	D5	D6	
Press reportage - MAT	-17.61	0.09	4.51	-1.55	0.85	-1.11	75% Learned exposition; 20% General narrative exposition; 4% Scientific exposition
Press reportage - Biber (1988)	-15.01	0.4	-0.3	-0.7	0.6	-0.9	73% General narrative exposition; 25% Learned exposition; 2% Scientific exposition
Difference	2.6	0.31	4.81	0.85	0.25	0.21	
Press editorials - MAT	-10.71	-0.59	4.5	1.39	0.63	-0.28	63% General narrative exposition; 7% Involved persuasion; 26% Learned exposition; 4% Scientific exposition
Press editorials - Biber (1988)	-10	-0.8	1.9	3.1	0.3	1.5	86% General narrative exposition; 11% Involved persuasion; 4% Learned exposition
Difference	0.71	0.21	2.6	1.71	0.33	1.78	
Press reviews - MAT	-13.83	-1.32	5.27	-3.31	0.41	-1.08	59% Learned exposition; 41% General narrative exposition
Press reviews - Biber (1988)	-13.9	-1.6	4.3	-2.8	0.8	-1	47% Learned exposition; 47% General narrative exposition; 6% Scientific exposition
Difference	0.07	0.28	0.97	0.51	0.39	0.08	
Religion - MAT	-7.17	-0.11	5.1	0.39	2.11	0.49	35% General narrative exposition; 29% Involved persuasion; 24% Learned exposition; 12% Scientific exposition
Religion - Biber (1988)	-7	-0.7	3.7	0.2	1.4	1	59% General narrative exposition; 18% Involved persuasion; 18% Learned exposition; 6% Imaginative narrative
Difference	0.17	0.59	1.4	0.19	0.71	0.51	
Hobbies - MAT	-12.44	-2.66	4.47	-0.86	1.34	-1.15	50% Learned exposition; 36% General narrative exposition; 6% Involved persuasion; 8% Scientific exposition
Hobbies - Biber (1988)	-10.1	-2.9	0.3	1.7	1.2	-0.7	43% General narrative exposition; 21% Learned exposition; 21% Involved persuasion; 7% Scientific exposition; 7% Situated reportage
Difference	2.34	0.24	4.17	2.56	0.14	0.45	
Popular lore - MAT	-13.3	-0.1	3.9	-1.03	1.38	-0.67	44% Learned exposition; 42% General narrative exposition; 8% Involved persuasion; 6% Scientific exposition
Popular lore - Biber (1988)	-9.3	-0.1	2.3	-0.3	0.1	-0.8	36% Learned exposition; 36% Involved persuasion; 21% General narrative exposition; 7% Imaginative narrative
Difference	4	0	1.6	0.73	1.28	0.13	

Academic prose - MAT	-13.58	-2.33	5.93	-0.88	4.48	0.01	38% Scientific exposition; 38% Learned exposition; 23% General narrative exposition; 3% Involved persuasion
Academic prose - Biber (1988)	-14.09	-2.6	4.2	-0.5	5.5	0.5	44% Scientific exposition; 31% Learned exposition; 17% General narrative exposition; 9% Involved persuasion
Difference	0.51	0.27	1.73	0.38	1.02	0.49	
General fiction - MAT	-5.83	5.86	0.19	-0.33	-0.44	-1.22	66% General narrative exposition; 24% Imaginative narrative; 10% Involved persuasion
General fiction - Biber (1988)	-0.8	5.9	-3.1	0.9	-2.5	-1.6	51% Imaginative narrative; 41% General narrative exposition; 3% Informational interaction; 3% Involved persuasion
Difference	5.03	0.04	3.29	1.23	2.06	0.38	
Mystery fiction - MAT	-2.21	5.57	-1.22	0.13	-1.03	-1	46% General narrative exposition; 42% Imaginative narrative; 13% Involved persuasion
Mystery fiction - Biber (1988)	-0.2	6	-3.6	-0.7	-2.8	-1.9	70% Imaginative narrative; 23% General narrative exposition; 8% Situated reportage
Difference	2.01	0.43	2.38	0.83	1.77	0.9	
Science fiction - MAT	-4.1	4.79	1.3	0.12	0.79	-0.78	50% General narrative exposition; 17% Imaginative narrative; 17% Involved persuasion; 17% Learned exposition
Science fiction - Biber (1988)	-6.1	5.9	-1.4	-0.7	-2.5	-1.6	50% General narrative exposition; 33% Imaginative narrative; 17% Situated reportage
Difference	2	1.11	2.7	0.82	3.29	0.82	
Adventure fiction - MAT	-6.05	5.88	-0.81	-1.78	-1.05	-1.39	66% General narrative exposition; 31% Imaginative narrative; 3% Learned exposition
Adventure fiction - Biber (1988)	0	5.5	-3.8	-1.2	-2.5	-1.9	70% Imaginative narrative; 31% General narrative exposition
Difference	6.05	0.38	2.99	-0.58	1.45	0.51	
Romantic fiction - MAT	0.83	6.02	0.41	-0.08	-1.15	-1.08	59% Imaginative narrative; 31% General narrative exposition; 10% Involved persuasion

Romantic fiction - Biber (1988)	4.3	7.2	-4.1	1.8	-3.1	-1.2	92% Imaginative narrative; 8% General narrative exposition
Difference	3.47	1.18	4.51	1.88	1.95	0.12	
Humour - MAT	-6.76	2.96	2.56	-1.16	0.42	-0.46	67% General narrative exposition; 22% Imaginative narrative; 11% Learned exposition
Humour - Biber (1988)	-7.8	0.9	-0.8	-0.3	-0.4	-1.5	89% General narrative exposition; 11% Involved persuasion
Difference	1.04	2.06	3.36	0.86	0.82	1.04	

Greater differences can be observed between MAT scores and Biber's (1988) scores. However, given that the Brown corpus contains identical genres but different texts from the LOB corpus, the results obtained from the analysis of the Brown corpus suggest that the Dimensions found by Biber (1988) are still valid for those genres even when considering different texts.

The results obtained with the latter experiment are encouraging and suggest that MAT can be used to assign Biber's (1988) Dimension scores to texts. Furthermore, MAT can be used to categorise a text for its text type, as proposed by Biber (1989).

List of the variables

Each variable is described in a short paragraph. Next to the name of the variable is the tag used by the present tagger to identify it. An asterisk appears next to the name of the variables for which Biber (1988) manually checked the results. The present version of the tagger does not allow any manual intervention in the tagging process. However, the texts can be manually checked before the analysis takes place.

Past tense (VBD)

The Stanford Tagger tag VBD is used for this variable (for further reference: http://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf).

Perfect aspect (PEAS)

This is calculated by counting how many times a form of HAVE is followed by: a VBD or VBN tag (a past or participle form of any verb). These are also counted when an adverb (RB) or negation (XX0) occurs between the two. The interrogative version is counted too. This is achieved by counting how many times a form of HAVE is followed by a nominal form (noun, NN, proper noun, NP or personal pronoun, PRP) and then followed by a VBD or VBN tag. As for the affirmative version, the latter algorithm also accounts for intervening adverbs or negations.

Present tense (VPRT)

Any verb that received by the Stanford Tagger a VBP or VBZ tag (present tense or third person present verb) is tagged as VPRT (for further reference: http://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf).

Place adverbials (PLACE)

Any item in this list: *aboard, above, abroad, across, ahead, alongside, around, ashore, astern, away, behind, below, beneath, beside, downhill, downstairs, downstream, east, far, hereabouts, indoors, inland, inshore, inside, locally, near, nearby, north, nowhere, outdoors, outside, overboard, overland, overseas, south, underfoot, underground, underneath, uphill, upstairs, upstream, west*. If an item is tagged by the Stanford Tagger as a proper noun (NNP), this is not tagged as place adverbial.

Time adverbials (TIME)

Any item in this list: *afterwards, again, earlier, early, eventually, formerly, immediately, initially, instantly, late, lately, later, momentarily, now, nowadays, once, originally, presently, previously, recently, shortly, simultaneously, subsequently, today, to-day, tomorrow, to-morrow, tonight, to-night, yesterday*. The list used in Biber (1988) was improved by adding that the word *soon* is not a time adverbial if it is followed by the word *as*. Furthermore, old spellings of the time adverbials starting with *to-* were added (e.g. *to-morrow*).

First person pronouns (FPP1)

Any item of this list: *I, me, us, my, we, our, myself, ourselves*.

Second person pronouns (SPP2)

Any item of this list: *you, your, yourself, yourselves, thy, thee, thyself, thou*.

Third person pronouns (TPP3)

Any item of this list: *she, he, they, her, him, them, his, their, himself, herself, themselves*.

Pronoun it *(PIT)*

Any pronoun *it*. Although not specified in Biber (1988), the present program also tags *its* and *itself* as "Pronoun *it*".

*Demonstrative pronouns (DEMP) **

The program tags as demonstrative pronouns the words *those, this, these* when they are followed by a verb (any tag starting with V) or auxiliary verb (modal verbs in the form of

MD tags or forms of DO or forms of HAVE or forms of BE) or a punctuation mark or a WH pronoun or the word *and*. The word *that* is tagged as a demonstrative pronoun when it follows the said pattern or when it is followed by *'s* or *is* and, at the same time, it has not been already tagged as a TOBJ, TSUB, THAC or THVC.

Indefinite pronouns (INPR)

Any item of this list: *anybody, anyone, anything, everybody, everyone, everything, nobody, none, nothing, nowhere, somebody, someone, something*.

Pro-verb do *(PROD)*

Any form of DO that is used as main verb and, therefore, excluding DO when used as auxiliary verb. The tagger tags as PROD any DO that is NOT in neither of the following patterns: (a) DO followed by a verb (any tag starting with V) or followed by adverbs (RB), negations and then a verb (V); (b) DO preceded by a punctuation mark or a WH pronoun (the list of WH pronouns is in Biber (1988)).

Direct WH-questions (WHQU)

Any punctuation followed by a WH word (*what, where, when, how, whether, why, whoever, whomever, whichever, wherever, whenever, whatever, however*) and followed by any auxiliary verb (modal verbs in the form of MD tags or forms of DO or forms of HAVE or forms of BE). This algorithm was slightly changed by allowing an intervening word between the punctuation mark and the WH word. This allows WH-questions containing discourse markers such as 'so' or 'anyways' to be recognised. Furthermore, Biber's algorithm was improved by excluding WH words such as *however* or *whatever* that do not introduce WH-questions.

Nominalizations (NOMZ)

Any noun ending in *-tion, -ment, -ness,* or *-ity*, plus the plural forms. Although Biber (1988) does not mention that this variables was checked manually, it is likely that a stop list was used to avoid obviously erroneous tagging (e.g. *city*). However, this was not indicated in the appendix of Biber (1988).

*Gerunds (GER)**

The program tags as gerunds any nominal form (N) that ends in *–ing* or *–ings*. To improve the accuracy, only words longer than 10 characters are considered as gerunds.

Total other nouns (NN)

Any noun that has been tagged by the Stanford Tagger as NN and that has not been identified a nominalisation or a gerund is left as such. Plural nouns (NNS) and proper nouns (NNP and NNPS) tags are changed to NN and included in this count.

Agentless passives (PASS)

This tag is assigned when one of the two following patterns is found: (a) any form of BE followed by a participle (VBN or VBD) plus one or two optional intervening adverbs (RB) or negations; (b) any form of BE followed by a nominal form (a noun, NN, NNP or personal pronoun, PRP) and a participle (VBN or VBD). This algorithm was slightly changed from Biber's version in the present tagger. It was felt necessary to implement the possibility of an intervening negation in the pattern (b). This tag is therefore assigned also in the cases in which a negation precedes the nominal form of pattern (b).

By-*passives (BYPA)*

The tagger assigns this tag every time the patterns for PASS are found and the preposition *by* follows it.

Be *as main verb (BEMA)*

BE is tagged as being a main verb in the following pattern: BE followed by a determiner (DT), or a possessive pronoun (PRP$) or a preposition (PIN) or an adjective (JJ). This algorithm was improved in the present tagger by taking into account that adverbs or negations can appear between the verb BE and the rest of the pattern. Furthermore, the algorithm was slightly modified and improved: (a) the problem of a double-coding of any Existential *there* followed by a form of BE as a BEMA was solved by imposing the condition that *there* should not appear before the pattern; (b) the cardinal numbers (CD) tag and the personal pronoun (PRP) tag were added to the list of items that can follow the form of BE.

Existential there *(EX)*

Existential *there* is tagged by the Stanford Tagger as EX (for further reference: http://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf).

That *verb complements (THVC)**

This tag is assigned when the word *that* is: (1) preceded by *and, nor, but, or, also* or any punctuation mark and followed by a determiner (DT, QUAN, CD), a pronoun (PRP), *there*, a plural noun (NNS) or a proper noun (NNP); (2) preceded by a public, private or suasive verb or a form of *seem* or *appear* and followed by any word that is NOT a verb (V), auxiliary verb (MD, form of DO, form of HAVE, form of BE), a punctuation or the word *and*; (3) preceded by a public, private or suasive verb or a form of *seem* or *appear* and a preposition and up to four words that are not nouns (N).

That *adjective complements (THAC)**

The program tags as THAC any word *that* preceded by an adjective (JJ or a predicative adjective, PRED).

WH-clauses (WHCL)

*(e.g. I believed **what he told me**)*

This tag is assigned when the following pattern is found: any public, private or suasive verb followed by any WH word, followed by a word that is NOT an auxiliary (tag MD for modal verbs, or a form of DO, or a form of HAVE, or a form of BE).

Infinitives (TO)

The tag for infinitives is the Stanford Tagger Treebank tag TO. The Stanford Tagger does not distinguish when the word *to* is used as an infinitive marker or a preposition. Therefore, an algorithm was implemented to identify instances of *to* as preposition. This algorithm finds any occurrence of *to* followed by a subordinator (IN), a cardinal number (CD), a determiner (DT), an adjective (JJ), a possessive pronoun (PRP$), WH words (WP$, WDT, WP, WRB), a pre-determiner (PDT), a noun (N, NNS, NP, NPs), or a pronoun (PRP) and tags it as a preposition. The remaining instances of *to* are considered as being infinitive markers and are therefore identifying occurrences of infinitive clauses.

*Present participial clauses (PRESP)**

(e.g. ***Stuffing his mouth with cookies****, Joe ran out the door)*

This tag is assigned when the following pattern is found: a punctuation mark is followed by a present participial form of a verb (VBG) followed by a preposition (PIN), a determiner (DT, QUAN, CD), a WH pronoun, a WH possessive pronoun (WP$), any WH word, any pronoun (PRP) or any adverb (RB).

*Past participial clauses (PASTP)**

(e.g. ***Built in a single week****, the house would stand for fifty years)*

This tag is assigned when the following pattern is found: a punctuation mark followed by a past participial form of a verb (VBN) followed by a preposition (PIN) or an adverb (RB).

*Past participial WHIZ deletion relatives (WZPAST)**

(e.g. The solution ***produced by this process****)*

This tag is assigned when the following pattern is found: a noun (N) or quantifier pronoun (QUPR) followed by a past participial form of a verb (VBN) followed by a preposition (PIN) or an adverb (RB) or a form of BE.

*Present participial WHIZ deletion relatives (WZPRES)**

(e.g. the event ***causing this decline*** *is….)*

This tag is assigned a present participial form of a verb (VBG) is preceded by a noun (NN).

That *relative clauses on subject position (TSUB)**

(e.g. the dog ***that bit me****)*

These are occurrences of *that* preceded by a noun (N) and followed by an auxiliary verb or a verb (V), with the possibility of an intervening adverb (RB) or negation (XX0).

That *relative clauses on object position (TOBJ)**

(e.g. the dog ***that I saw****)*

These are occurrences of *that* preceded by a noun and followed by a determiner (DT, QUAN, CD), a subject form of a personal pronoun, a possessive pronoun (PRP$), the pronoun *it*, an adjective (JJ), a plural noun (NNS), a proper noun (NNP) or a possessive noun (a noun (N) followed by a genitive marker (POS)). As Biber specifies, however, this algorithm does not distinguish between simple complements to nouns and true relative clauses.

WH relative clauses on subject position (WHSUB)

*(e.g. the man **who likes popcorn**)*

This tag is assigned when the following pattern is found: any word that is NOT a form of the words ASK or TELL followed by a noun (N), then a WH pronoun, then by any verb or auxiliary verb (V), with the possibility of an intervening adverb (RB) or negation (XX0) between the WH pronoun and the verb.

WH relative clauses on object position (WHOBJ)

*(e.g. the man **who Sally likes**)*

This tag is assigned when the following pattern is found: any word that is NOT a form of the words ASK or TELL followed by any word, followed by a noun (N), followed by any word that is NOT an adverb (RB), a negation (XX0) , a verb or an auxiliary verb (MD, forms of HAVE, BE or DO).

Pied-piping relative clauses (PIRE)

*(e.g. the manner **in which he was told**)*

This tag is assigned when the following pattern is found: any preposition (PIN) followed by *who, who, whose* or *which*.

*Sentence relatives (SERE)**

*(e.g. Bob likes fried mangoes, **which is disgusting**)*

A sentence relative is counted and tagged every time a punctuation mark is followed by the word *which*.

Causative adverbial subordinators (CAUS)

This tag identifies any occurrence of the word *because*.

Concessive adverbial subordinators (CONC)

This tag identifies any occurrence of the words *although* and *though*. Biber's algorithm was improved by including the abbreviation *tho*.

Conditional adverbial subordinators (COND)

This tag identifies any occurrence of the words *if* and *unless*.

Other adverbial subordinators (OSUB)

This tag identifies any occurrence of the words: *since, while, whilst, whereupon, whereas, whereby, such that, so that* (followed by a word that is neither a noun nor an adjective), *such that* (followed by a word that is neither a noun nor an adjective), *inasmuch as, forasmuch as, insofar as, insomuch as, as long as, as soon as*. In cases of multi-word units such as *as long as*, only the first word is tagged as OSUB and the other words are tagged with the tag NULL.

Total prepositional phrases (PIN)

This tag identifies any occurrence of the prepositions listed by Biber (1988) under this category. As described in the section on infinitives, the preposition *to* is disambiguated by the infinitive marker *to*. Biber (1988) does not specifies whether he included any instance of the word *to* or he distinguished the two grammatical functions of this word. However, it was felt the distinction needed to be applied to the present tagger for improved accuracy.

Attributive adjectives (JJ)

*(e.g. the **big** horse)*

Biber (1988) specifies that attributive adjectives were counted when an adjective was followed by another adjective or a noun. However, Biber states that also all the adjectives that were not identified as predicative were counted as attributive adjectives. Therefore, the present tagger does not have an algorithm to identify attributive adjectives. All the adjectives that the Stanford Tagger has already tagged as JJ, JJS, or JJR are considered attributive adjectives and are all re-assigned to the tag JJ. The predicative adjectives are tagged by another algorithm and therefore distinguished from the rest.

Predicative adjectives (PRED)

*(e.g. the horse is **big**)*

The tagger tags as PRED the adjectives that are found in the following pattern: any form of BE followed by an adjective (JJ) followed by a word that is NOT another adjective, an adverb (RB) or a noun (N). If any adverb or negation is intervening between the adjective and the word after it, the tag is still assigned. A modification to Biber's algorithm was implemented in the present tagger to improve its accuracy. An adjective is tagged as predicative if it is preceded by another predicative adjective followed by a phrasal coordinator (see below). This pattern accounts for cases such as: *the horse is big and fast*.

Total adverbs (RB)

All the adverbs that the Stanford Tagger has already tagged as RB, RBS, RBR or WRB are all re-assigned to the tag RB in order to have a final count of total adverbs (for further reference: http://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf).

Type-token ratio (TTR)

In Biber (1988), the tagger considered only the first 400 tokens of the text and counted how many types were present in these 400 tokens. The resulting number was therefore the number of types in the first 400 words of the text. If a text was shorter than 400 tokens, it was excluded from this analysis.

The number 400 was chosen by Biber supposedly as it provided a compromise between accuracy and number of texts that could be measured. Since the present tagger can be applied to corpora of different sizes, it was felt that this number should be left to the user to decide. The tagger will therefore ask to input the number before the tagging starts. It will then count how many types there are in the first X number of tokens given by the user. For texts shorter than X, the program will count the types for the whole text. The user can decide which number to use based on either the shortest text in the corpus or perhaps on the statistical mode of the population of the number of tokens for the whole corpus.

By default, this number is 400. The variable type-token ratio will be included in the calculation of Dimension 1 only if the user has not changed the default number. This is done in order to maintain compatibility with Biber's (1988) calculations.

Word length (AWL)

Mean length of the words in the text in orthographic letters. A word is any string separated by space in the text tokenised by the Stanford Tagger.

Conjuncts (CONJ)

This tag finds any of the items in this list: punctuation+*else*, punctuation+*altogether*, punctuation+*rather*, *alternatively, consequently, conversely, e.g., furthermore, hence, however, i.e., instead, likewise, moreover, namely, nevertheless, nonetheless, notwithstanding, otherwise, similarly, therefore, thus, viz., in comparison, in contrast, in particular, in addition, in conclusion, in consequence, in sum, in summary, for example, for*

instance, instead of, by contrast, by comparison, in any event, in any case, in other words, as a result, as a consequence, on the contrary, on the other hand.

Some minor inconsistencies in the said list were fixed. For example, Biber lists the word *rather* two times in this list, making the second mentions redundant. *Rather* was counted only when it appeared after a punctuation mark. The same applies for *altogether*. In cases of multi-word units such as *on the other hand*, only the first word is tagged as OSUB and the other words are tagged with the tag NULL.

Downtoners (DWNT)

This tag finds any of the items in this list: *almost, barely, hardly, merely, mildly, nearly, only, partially, partly, practically, scarcely, slightly, somewhat*. The word *almost* was classified by Biber as being both a hedge and a downtoner. In the present tagger *almost* is considered a downtoner only.

Hedges (HDG)

This tag finds any of the items in this list: *maybe, at about, something like, more or less, sort of, kind of* (these two items must be preceded by a determiner (DT), a quantifier (QUAN), a cardinal number (CD), an adjective (JJ or PRED), a possessive pronouns (PRP$) or WH word (see entry on WH-questions)). In cases of multi-word units such as *more or less*, only the first word is tagged as HDG and the other words are tagged with the tag NULL.

Amplifiers (AMP)

This tag finds any of the items in this list: *absolutely, altogether, completely, enormously, entirely, extremely, fully, greatly, highly, intensely, perfectly, strongly, thoroughly, totally, utterly, very*.

Emphatics (EMPH)

This tag finds any of the items in this list: *just, really, most, more, real*+adjective, *so*+adjective, any form of DO followed by a verb, *for sure, a lot, such a*. In cases of multi-word units such as *a lot*, only the first word is tagged as OSUB and the other words are tagged with the tag NULL.

Discourse particles (DPAR)

The program tags as discourse particles the words *well, now, anyhow, anyways* preceded by a punctuation mark.

Demonstratives (DEMO)

A demonstrative is found when the words *that, this, these, those* have not been tagged as either DEMP, TOBJ, TSUB, THAC, or THVC.

Possibility modals (POMD)

The possibility modals listed by Biber (1988): *can, may, might, could*.

Necessity modals (NEMD)

The necessity modals listed by Biber (1988): *ought, should, must*.

Predictive modals (PRMD)

The predictive modals listed by Biber (1988): *will, would, shall* and their contractions: *'d_MD, ll_MD, wo_MD, sha_MD*.

Public verbs (PUBV)

This tag finds any of the items listed by Quirk *et al.* (1985: 1180–1): *acknowledge, acknowledged, acknowledges, acknowledging, add, adds, adding, added, admit, admits, admitting, admitted, affirm, affirms, affirming, affirmed, agree, agrees, agreeing, agreed, allege, alleges, alleging, alleged, announce, announces, announcing, announced, argue, argues, arguing, argued, assert, asserts, asserting, asserted, bet, bets, betting, boast, boasts, boasting, boasted, certify, certifies, certifying, certified, claim, claims, claiming, claimed, comment, comments, commenting, commented, complain, complains, complaining, complained, concede, concedes, conceding, conceded, confess, confesses, confessing, confessed, confide, confides, confiding, confided, confirm, confirms, confirming, confirmed, contend, contends, contending, contended, convey, conveys, conveying, conveyed, declare, declares, declaring, declared, deny, denies, denying, denied, disclose, discloses, disclosing, disclosed, exclaim, exclaims, exclaiming, exclaimed, explain, explains, explaining, explained, forecast, forecasts, forecasting, forecasted, foretell, foretells, foretelling, foretold, guarantee, guarantees, guaranteeing, guaranteed, hint, hints, hinting, hinted, insist, insists, insisting,*

insisted, maintain, maintains, maintaining, maintained, mention, mentions, mentioning, mentioned, object, objects, objecting, objected, predict, predicts, predicting, predicted, proclaim, proclaims, proclaiming, proclaimed, promise, promises, promising, promised, pronounce, pronounces, pronouncing, pronounced, prophesy, prophesies, prophesying, prophesied, protest, protests, protesting, protested, remark, remarks, remarking, remarked, repeat, repeats, repeating, repeated, reply, replies, replying, replied, report, reports, reporting, reported, say, says, saying, said, state, states, stating, stated, submit, submits, submitting, submitted, suggest, suggests, suggesting, suggested, swear, swears, swearing, swore, sworn, testify, testifies, testifying, testified, vow, vows, vowing, vowed, warn, warns, warning, warned, write, writes, writing, wrote, written.

Private verbs (PRIV)

This tag finds any of the items listed by Quirk *et al.* (1985: 1181–2): *accept, accepts, accepting, accepted, anticipate, anticipates, anticipating, anticipated, ascertain, ascertains, ascertaining, ascertained, assume, assumes, assuming, assumed, believe, believes, believing, believed, calculate, calculates, calculating, calculated, check, checks, checking, checked, conclude, concludes, concluding, concluded, conjecture, conjectures, conjecturing, conjectured, consider, considers, considering, considered, decide, decides, deciding, decided, deduce, deduces, deducing, deduced, deem, deems, deeming, deemed, demonstrate, demonstrates, demonstrating, demonstrated, determine, determines, determining, determined, discern, discerns, discerning, discerned, discover, discovers, discovering, discovered, doubt, doubts, doubting, doubted, dream, dreams, dreaming, dreamt, dreamed, ensure, ensures, ensuring, ensured, establish, establishes, establishing, established, estimate, estimates, estimating, estimated, expect, expects, expecting, expected, fancy, fancies, fancying, fancied, fear, fears, fearing, feared, feel, feels, feeling, felt, find, finds, finding, found, foresee, foresees, foreseeing, foresaw, forget, forgets, forgetting, forgot, forgotten, gather, gathers, gathering, gathered, guess, guesses, guessing, guessed, hear, hears, hearing, heard, hold, holds, holding, held, hope, hopes, hoping, hoped, imagine, imagines, imagining, imagined, imply, implies, implying, implied, indicate, indicates, indicating, indicated, infer, infers, inferring, inferred, insure, insures, insuring, insured, judge, judges, judging, judged, know, knows, knowing, knew, known, learn, learns, learning, learnt, learned, mean, means, meaning, meant, note, notes, noting, noted, notice, notices, noticing, noticed, observe, observes, observing, observed, perceive, perceives, perceiving, perceived, presume, presumes, presuming, presumed, presuppose, presupposes, presupposing, presupposed,*

pretend, pretend, pretending, pretended, prove, proves, proving, proved, realize, realise, realising, realizing, realises, realizes, realised, realized, reason, reasons, reasoning, reasoned, recall, recalls, recalling, recalled, reckon, reckons, reckoning, reckoned, recognize, recognise, recognizes, recognises, recognizing, recognising, recognized, recognised, reflect, reflects, reflecting, reflected, remember, remembers, remembering, remembered, reveal, reveals, revealing, revealed, see, sees, seeing, saw, seen, sense, senses, sensing, sensed, show, shows, showing, showed, shown, signify, signifies, signifying, signified, suppose, supposes, supposing, supposed, suspect, suspects, suspecting, suspected, think, thinks, thinking, thought, understand, understands, understanding, understood.

Suasive verbs (SUAV)

This tag finds any of the items listed by Quirk *et al.* (1985: 1182–3): *agree, agrees, agreeing, agreed, allow, allows, allowing, allowed, arrange, arranges, arranging, arranged, ask, asks, asking, asked, beg, begs, begging, begged, command, commands, commanding, commanded, concede, concedes, conceding, conceded, decide, decides, deciding, decided, decree, decrees, decreeing, decreed, demand, demands, demanding, demanded, desire, desires, desiring, desired, determine, determines, determining, determined, enjoin, enjoins, enjoining, enjoined, ensure, ensures, ensuring, ensured, entreat, entreats, entreating, entreated, grant, grants, granting, granted, insist, insists, insisting, insisted, instruct, instructs, instructing, instructed, intend, intends, intending, intended, move, moves, moving, moved, ordain, ordains, ordaining, ordained, order, orders, ordering, ordered, pledge, pledges, pledging, pledged, pray, prays, praying, prayed, prefer, prefers, preferring, preferred, pronounce, pronounces, pronouncing, pronounced, propose, proposes, proposing, proposed, recommend, recommends, recommending, recommended, request, requests, requesting, requested, require, requires, requiring, required, resolve, resolves, resolving, resolved, rule, rules, ruling, ruled, stipulate, stipulates, stipulating, stipulated, suggest, suggests, suggesting, suggested, urge, urges, urging, urged, vote, votes, voting, voted,*

Seem|appear *(SMP)*

Any occurrence of any of the forms of the two verbs *seem* and *appear*.

Contractions (CONT)

The contractions were tagged by identifying any instance of apostrophe followed by a tagged word OR any instance of the item *n't*.

Subordinator that *deletion (THATD)*

The tag THATD is added when one of the following patterns is found: (1) a public, private or suasive verb followed by a demonstrative pronoun (DEMP) or a subject form of a personal pronoun; (2) a public, private or suasive verb is followed by a pronoun (PRP) or a noun (N) and then by a verb (V) or auxiliary verb; (3) a public, private or suasive verb is followed by an adjective (JJ or PRED), an adverb (RB), a determiner (DT, QUAN, CD) or a possessive pronoun (PRP$) and then a noun (N) and then a verb or auxiliary verb, with the possibility of an intervening adjective (JJ or PRED) between the noun and its preceding word.

Stranded preposition (STPR)

*(e.g. the candidate that I was thinking **of**)*

A stranded preposition is identified every time a preposition is followed by a punctuation mark. However, this algorithm was improved by adding that the preposition cannot be *besides*, since this word can also be a conjunct and, therefore, usually followed by a punctuation mark.

Split infinitives (SPIN)

*(e.g. he wants **to convincingly prove that…**)*

Split infinitives are identified every time an infinitive marker *to* is followed by one or two adverbs and a verb base form.

Split auxiliaries (SPAU)

*(e.g. they **are objectively shown** that…)*

Split auxiliaries are identified every time an auxiliary (any modal verb MD, or any form of DO, or any form of BE, or any form of HAVE) is followed by one or two adverbs and a verb base form.

Phrasal coordination (PHC)

This tag was assigned for any *and* that is preceded and followed by the same tag and when this tag is either an adverb tag, or an adjective tag, or a verb tag or a noun tag.

Independent clause coordination (ANDC)

This tag is assigned to the word *and* when it is found in one of the following patterns: (1) preceded by a comma and followed by *it, so, then, you, there* + BE, or a demonstrative

pronoun (DEMP) or the subject forms of a personal pronouns; (2) preceded by any punctuation; (3) followed by a WH pronoun or any WH word, an adverbial subordinator (CAUS, CONC, COND, OSUB) or a discourse particle (DPAR) or a conjunct (CONJ).

Synthetic negation (SYNE)

The following pattern was identified as synthetic negation: *no* followed by any adjective (both JJ and PRED) and any noun or proper noun. The words *neither* and *nor* were also tagged as instances of synthetic negation.

Analytic negation (XX0)

This tag was assigned to the word *not* and to the item *n't_RB*.

Other Stanford Tagger tags

If the user selects "all tags" from the main window then all the tags assigned by the Stanford Tagger are counted as well. A list of the Stanford Tagger tags and the description of how they are identified can be found here:

http://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf

References

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1989). A typology of English texts. *Linguistics*, *27*(1), 3–43.

Stanford Tagger v. 3.1.5. Retrieved from: http://nlp.stanford.edu/software/tagger.shtml.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.