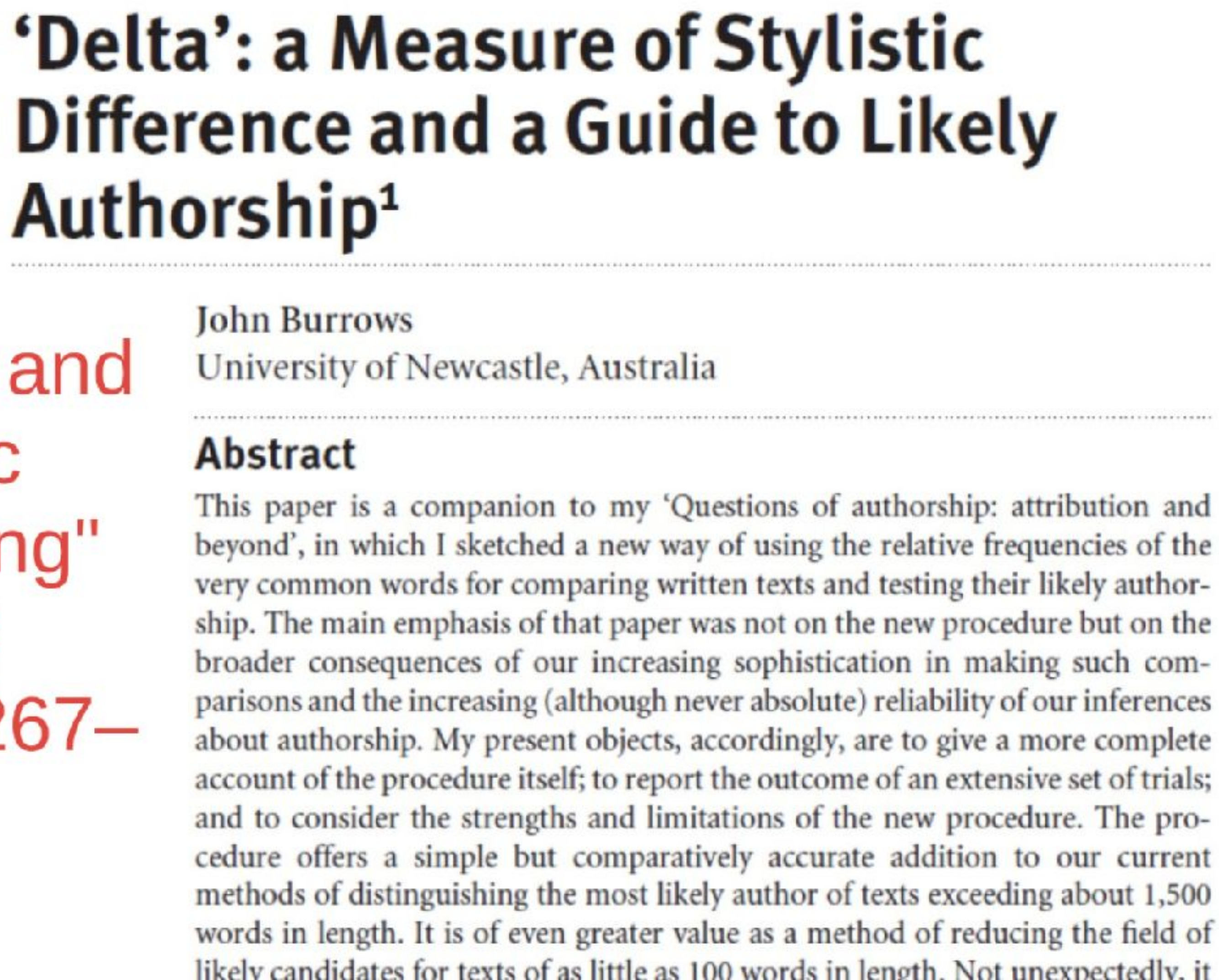
STYLOMETRY

WORD-FREQUENCY BASED STYLOMETRY



‘Delta’: a Measure of Stylistic Difference and a Guide to Likely Authorship[1]

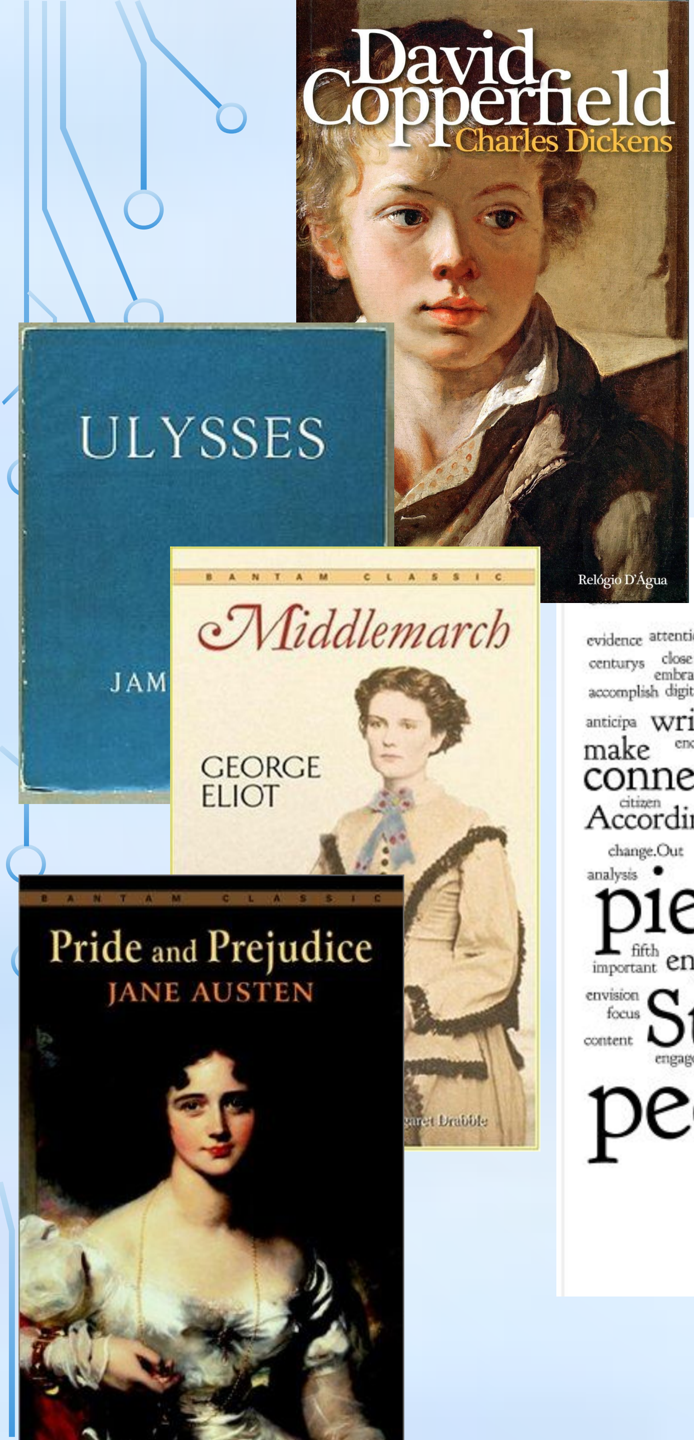
John Burrows
University of Newcastle, Australia

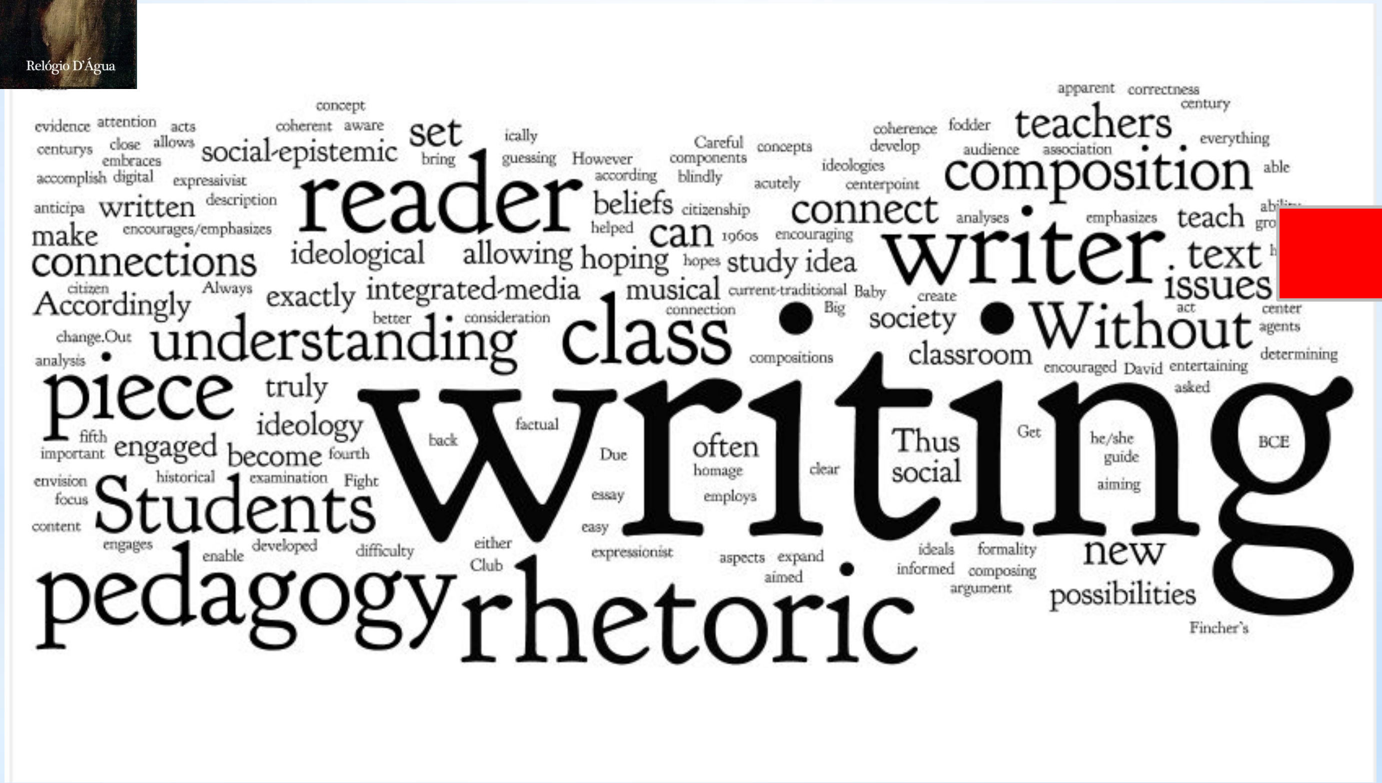
Abstract

This paper is a companion to my ‘Questions of authorship: attribution and beyond’, in which I sketched a new way of using the relative frequencies of the very common words for comparing written texts and testing their likely authorship. The main emphasis of that paper was not on the new procedure but on the broader consequences of our increasing sophistication in making such comparisons and the increasing (although never absolute) reliability of our inferences about authorship. My present objects, accordingly, are to give a more complete account of the procedure itself; to report the outcome of an extensive set of trials; and to consider the strengths and limitations of the new procedure. The procedure offers a simple but comparatively accurate addition to our current methods of distinguishing the most likely author of texts exceeding about 1,500 words in length. It is of even greater value as a method of reducing the field of likely candidates for texts of as little as 100 words in length. Not unexpectedly, it

"Literary and Linguistic Computing" 17, no. 3 (2002): 267–87

DELTA DISTANCE





1. the
2. and
3. of
4. to
5. a
6. i
7. in
8. he
9. was
10. it
11. that
12. you
13. his
14. her
15. with
16. as
17. had
18. she
19. for

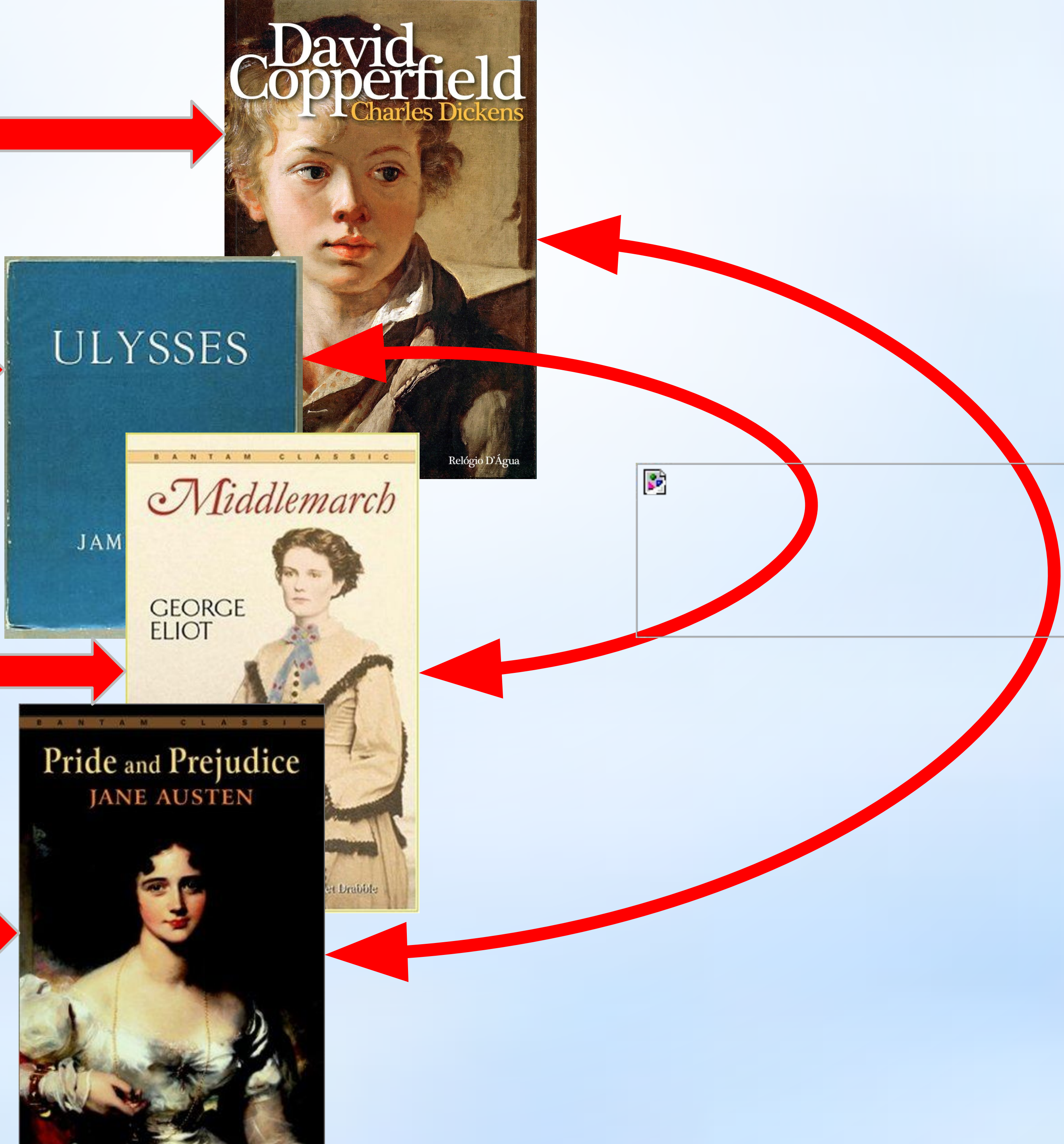
1. the
2. and
3. of
4. to
5. a
6. i
7. in
8. he
9. was
10. it
11. that
12. you
13. his
14. her
15. with
16. as
17. had
18. she
19. for

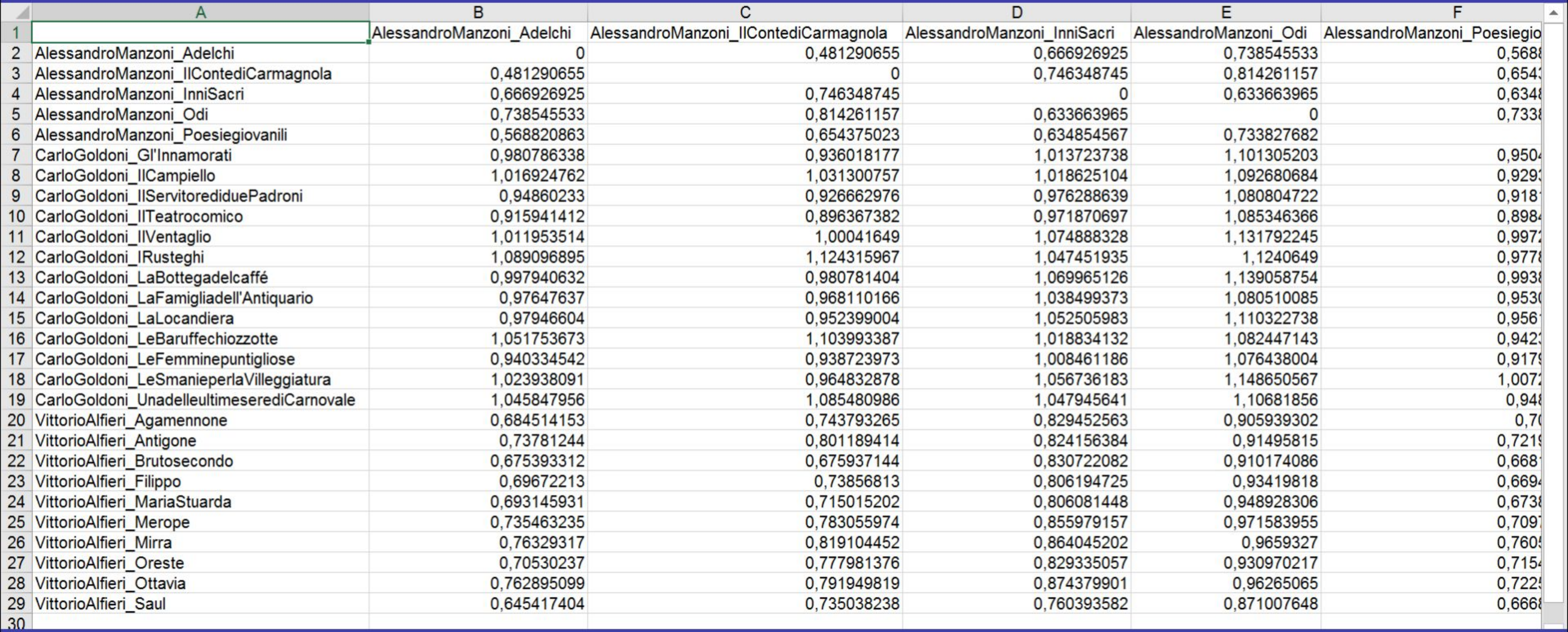
5.1%
3.2%
2.4%
2.5%

4.1%
3.3%
2.2%
2.7%

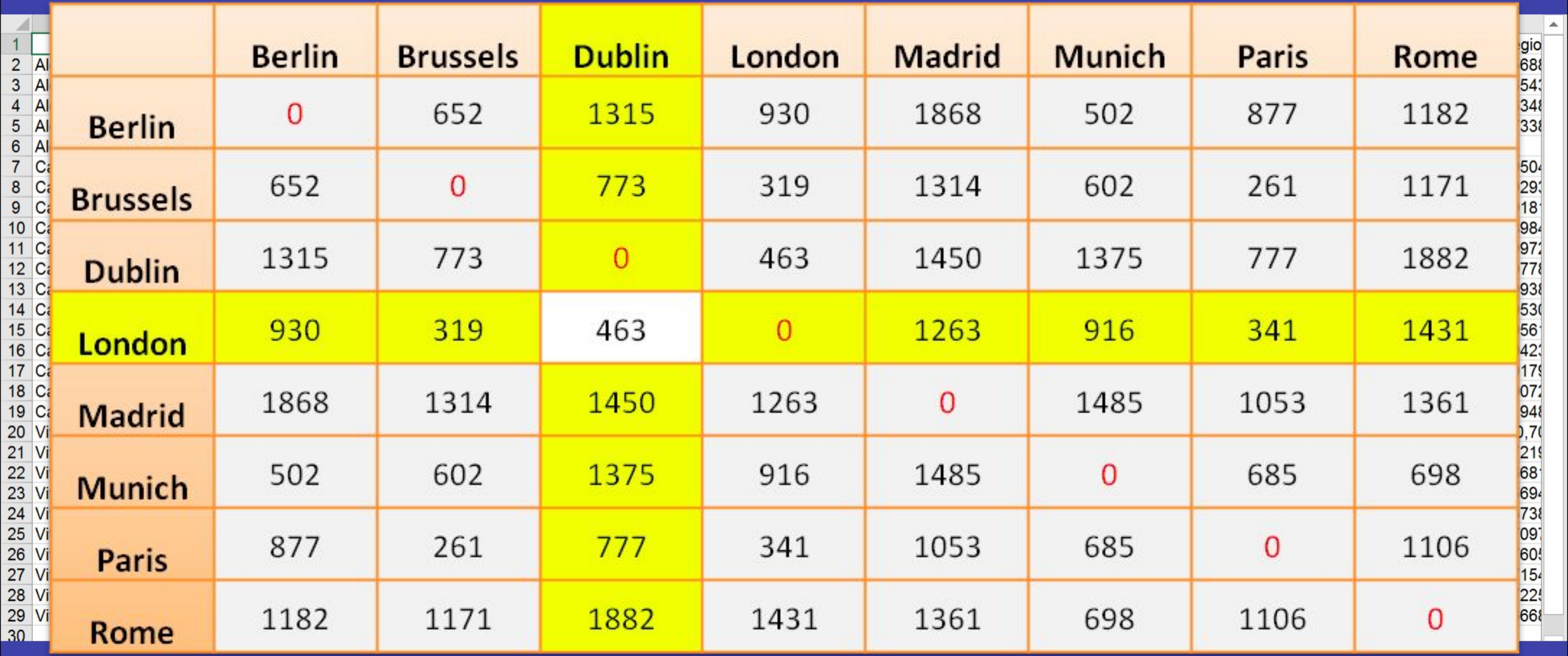
3.1%
4.2%
1.4%
1.2%

5.2%
3.2%
2.4%
2.5%





	A	B	C	D	E	F
1		AlessandroManzoni_Adelchi	AlessandroManzoni_IlContediCarmagnola	AlessandroManzoni_InniSacri	AlessandroManzoni_Odi	AlessandroManzoni_Poesiegio
2	AlessandroManzoni_Adelchi	0	0,481290655	0,666926925	0,738545533	0,5688
3	AlessandroManzoni_IlContediCarmagnola	0,481290655	0	0,746348745	0,814261157	0,6543
4	AlessandroManzoni_InniSacri	0,666926925	0,746348745	0	0,633663965	0,6348
5	AlessandroManzoni_Odi	0,738545533	0,814261157	0,633663965	0	0,7338
6	AlessandroManzoni_Poesiegiovanili	0,568820863	0,654375023	0,634854567	0,733827682	
7	CarloGoldoni_Gl'Innamorati	0,980786338	0,936018177	1,013723738	1,101305203	0,9504
8	CarloGoldoni_IlCampiello	1,016924762	1,031300757	1,018625104	1,092680684	0,9293
9	CarloGoldoni_IlServitoredidue Padroni	0,94860233	0,926662976	0,976288639	1,080804722	0,9181
10	CarloGoldoni_IlTeatrocomico	0,915941412	0,896367382	0,971870697	1,085346366	0,8984
11	CarloGoldoni_IlVentaglio	1,011953514	1,00041649	1,074888328	1,131792245	0,9972
12	CarloGoldoni_IRusteghi	1,089096895	1,124315967	1,047451935	1,1240649	0,9778
13	CarloGoldoni_LaBottegadelcaffé	0,997940632	0,980781404	1,069965126	1,139058754	0,9938
14	CarloGoldoni_LaFamigliadell'Antiquario	0,97647637	0,968110166	1,038499373	1,080510085	0,9530
15	CarloGoldoni_LaLocandiera	0,97946604	0,952399004	1,052505983	1,110322738	0,9561
16	CarloGoldoni_LeBaruffechiozzotte	1,051753673	1,103993387	1,018834132	1,082447143	0,9423
17	CarloGoldoni_LeFemminepuntigliose	0,940334542	0,938723973	1,008461186	1,076438004	0,9179
18	CarloGoldoni_LeSmanieperlaVilleggiatura	1,023938091	0,964832878	1,056736183	1,148650567	1,0072
19	CarloGoldoni_UnadelleultimeserediCarnovale	1,045847956	1,085480986	1,047945641	1,10681856	0,948
20	VittorioAlfieri_Agamennone	0,684514153	0,743793265	0,829452563	0,905939302	0,70
21	VittorioAlfieri_Antigone	0,73781244	0,801189414	0,824156384	0,91495815	0,7219
22	VittorioAlfieri_Brutosecondo	0,675393312	0,675937144	0,830722082	0,910174086	0,6681
23	VittorioAlfieri_Filippo	0,69672213	0,73856813	0,806194725	0,93419818	0,6694
24	VittorioAlfieri_MariaStuarda	0,693145931	0,715015202	0,806081448	0,948928306	0,6738
25	VittorioAlfieri_Merope	0,735463235	0,783055974	0,855979157	0,971583955	0,7097
26	VittorioAlfieri_Mirra	0,76329317	0,819104452	0,864045202	0,9659327	0,7605
27	VittorioAlfieri_Oreste	0,70530237	0,777981376	0,829335057	0,930970217	0,7154
28	VittorioAlfieri_Ottavia	0,762895099	0,791949819	0,874379901	0,96265065	0,7225
29	VittorioAlfieri_Saul	0,645417404	0,735038238	0,760393582	0,871007648	0,6668
30						



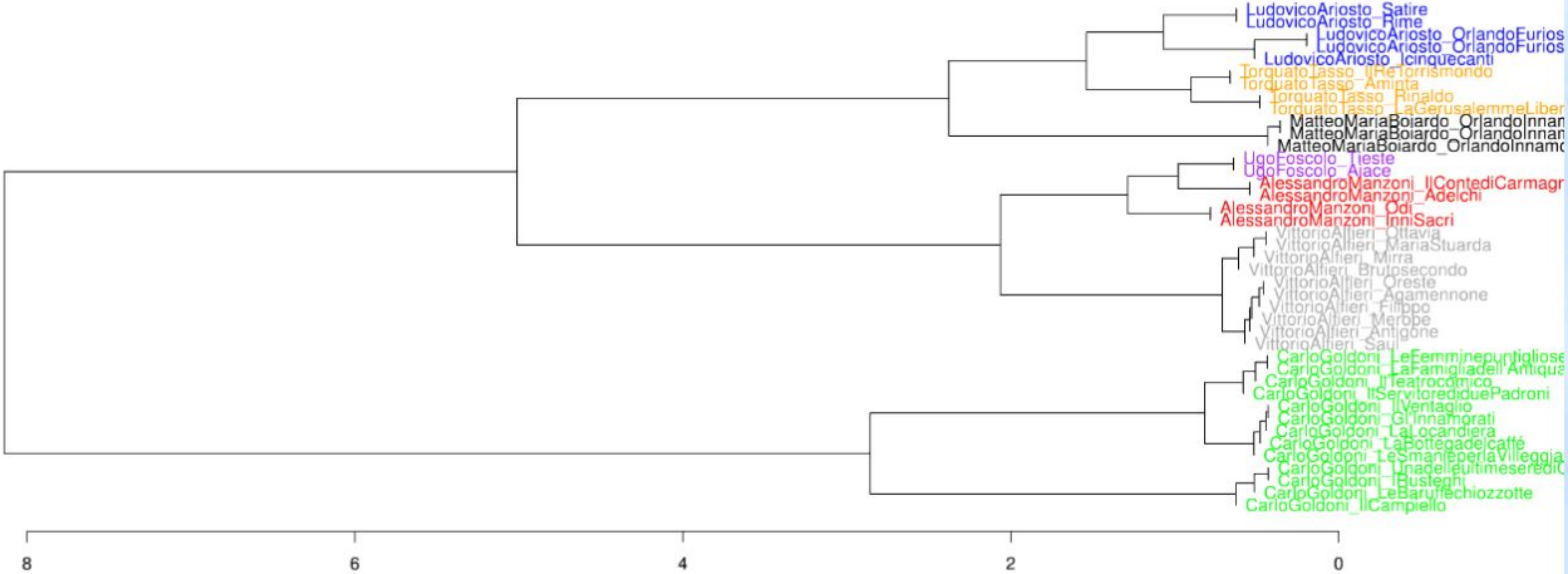
	Berlin	Brussels	Dublin	London	Madrid	Munich	Paris	Rome
Berlin	0	652	1315	930	1868	502	877	1182
Brussels	652	0	773	319	1314	602	261	1171
Dublin	1315	773	0	463	1450	1375	777	1882
London	930	319	463	0	1263	916	341	1431
Madrid	1868	1314	1450	1263	0	1485	1053	1361
Munich	502	602	1375	916	1485	0	685	698
Paris	877	261	777	341	1053	685	0	1106
Rome	1182	1171	1882	1431	1361	698	1106	0

VISUALIZATIONS

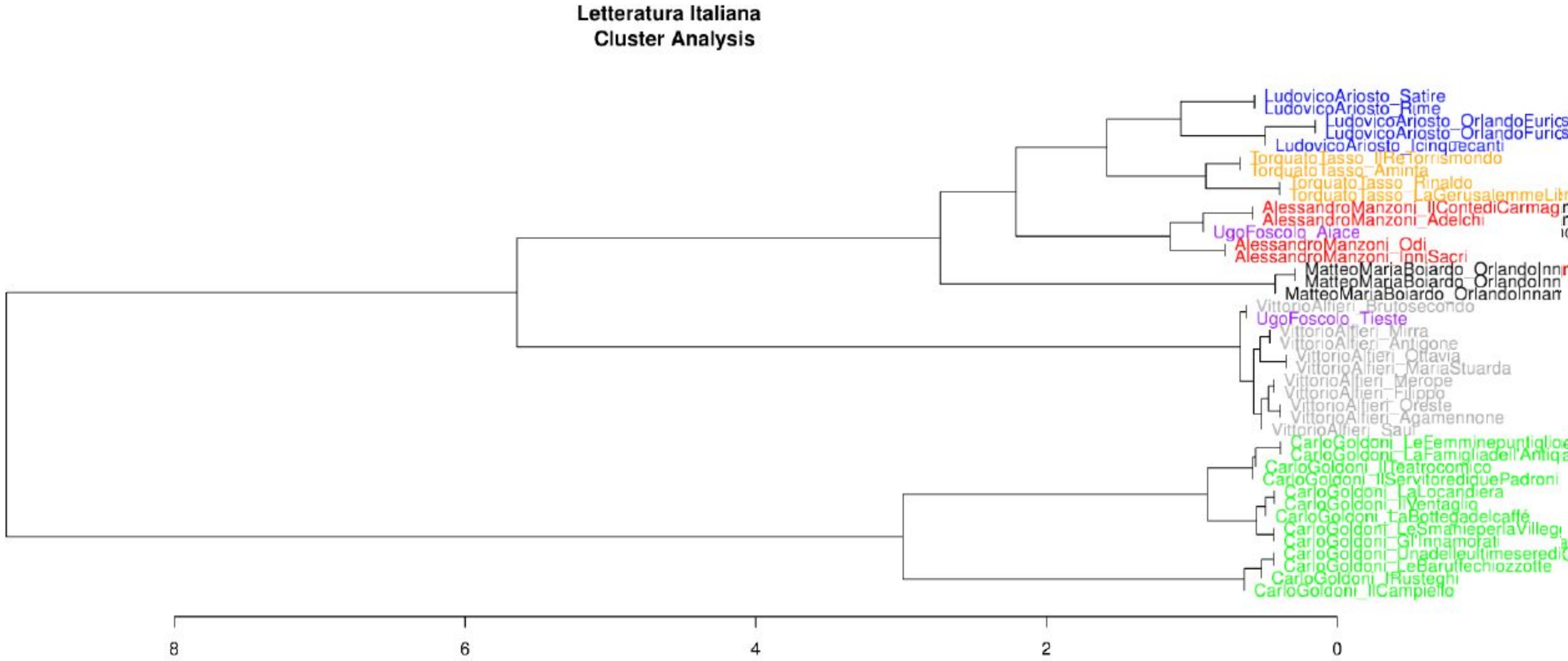
1. Dendrograms

Ward’s clustering algorithm (Ward, 1963)

Letteratura Italiana
Cluster Analysis

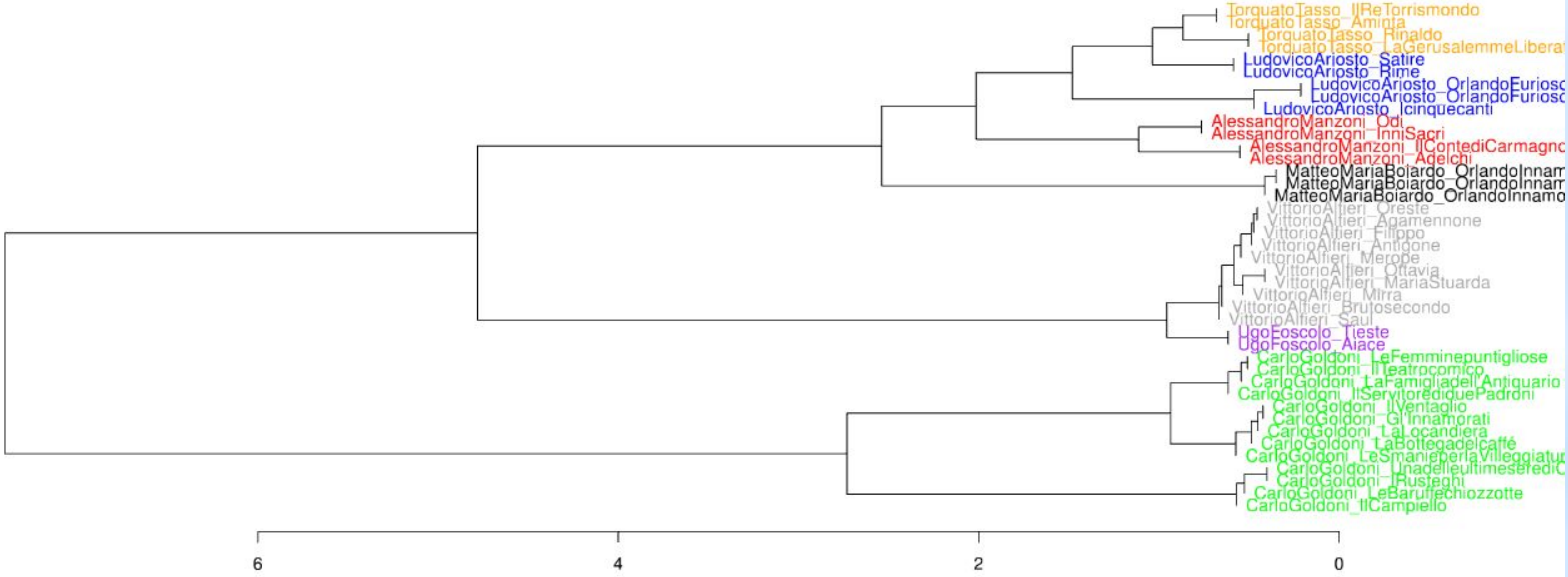


Burrows Delta
with 100 most frequent words (MFW)

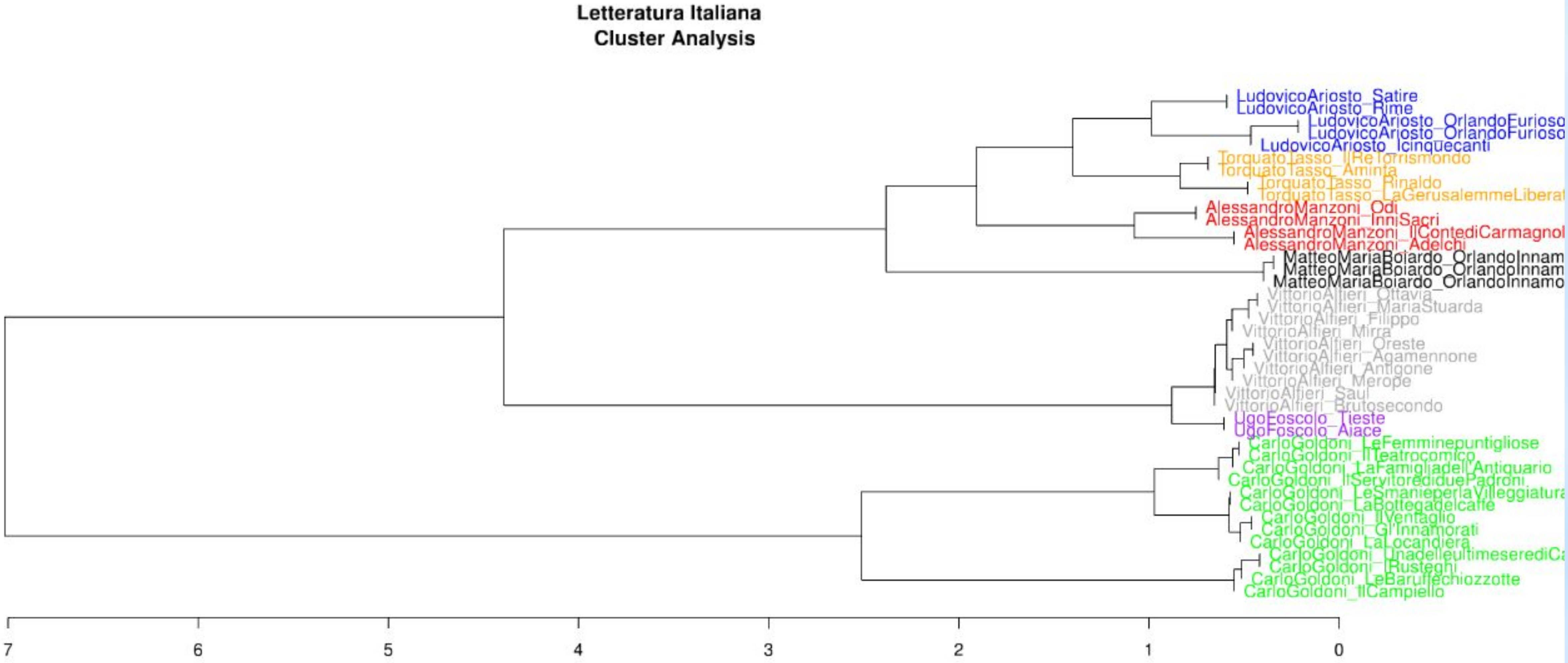


Burrows Delta
with 200 most frequent words (MFW)

**Letteratura Italiana
Cluster Analysis**



Cosine Delta
with 100 most frequent words (MFW)



My Weird Distance Measure
with 1,000,000 most frequent words (MFW)

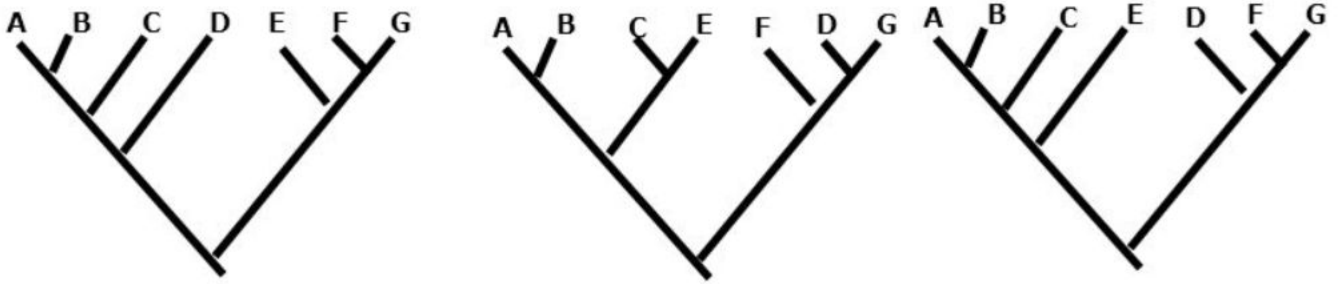
VISUALIZATIONS

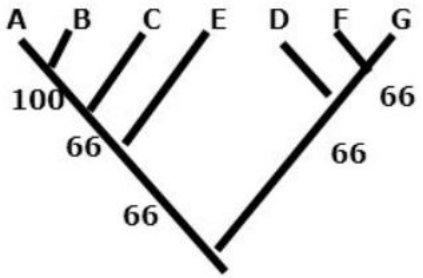
2. Consensus Trees

Method developed in philogenetics
(see Paradis et al. 2004)

Consensus Trees

Majority rule consensus

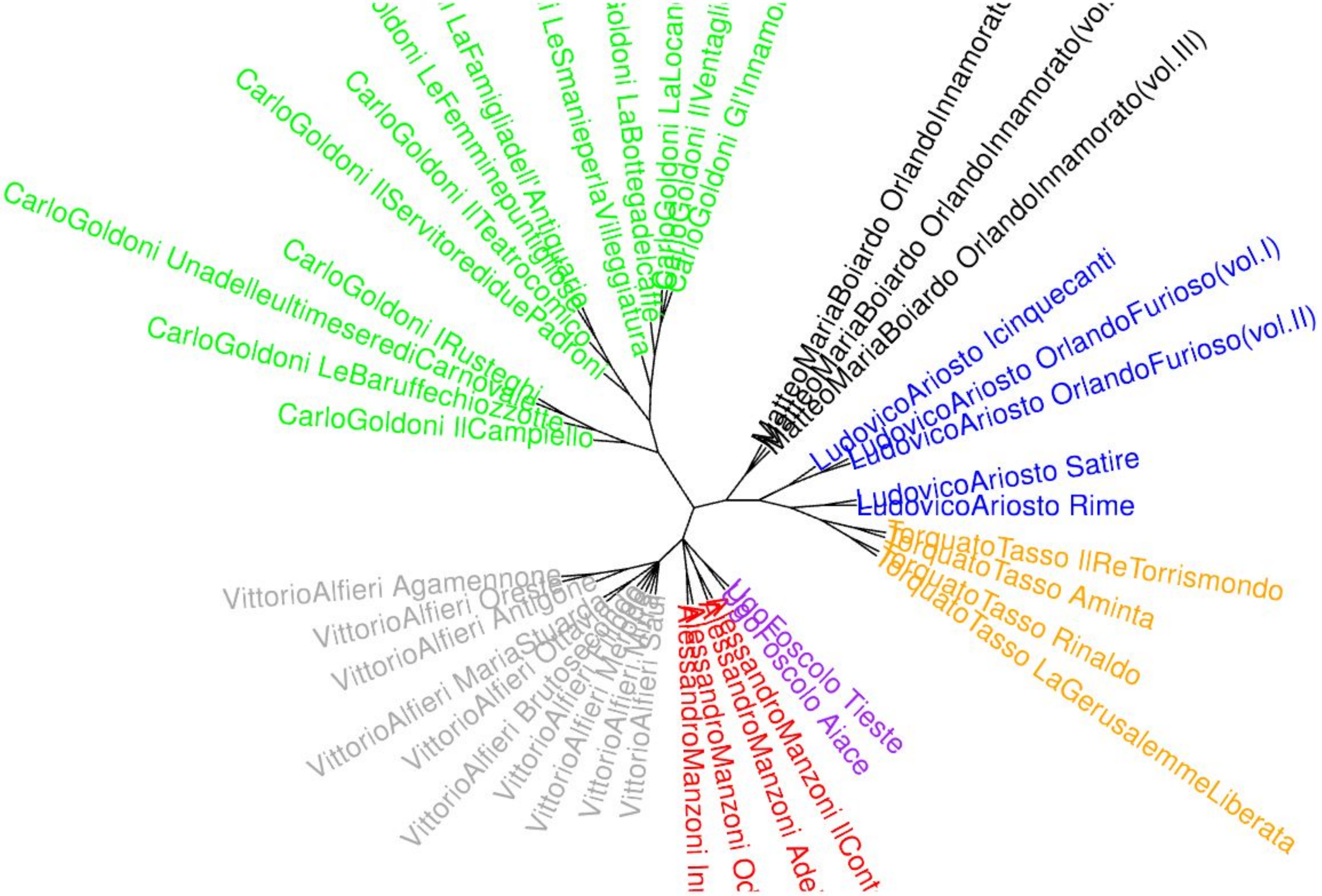




Numbers indicate frequency of clades in the fundamental trees

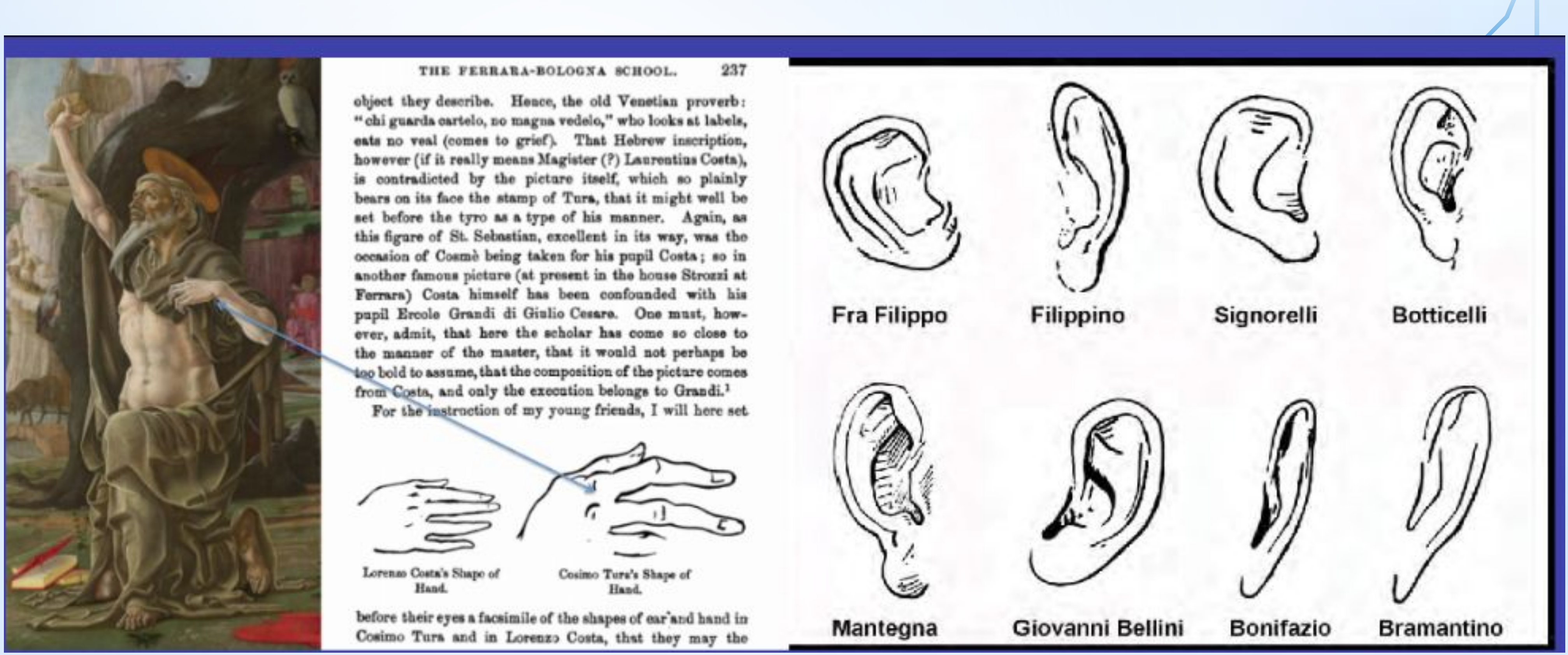
MAJORITY-RULE CONSENSUS TREE

Letteratura Italiana
Bootstrap Consensus Tree



100–1000 MFW Culled @ 0%
Classic Delta distance Consensus 0.5

WHY DOES IT WORK?



THE FERRARA-BOLOGNA SCHOOL. 237

object they describe. Hence, the old Venetian proverb: "chi guarda cartelo, no magna vedelo," who looks at labels, eats no veal (comes to grief). That Hebrew inscription, however (if it really means Magister (?) Laurentius Costa), is contradicted by the picture itself, which so plainly bears on its face the stamp of Tura, that it might well be set before the tyro as a type of his manner. Again, as this figure of St. Sebastian, excellent in its way, was the occasion of Cosmè being taken for his pupil Costa; so in another famous picture (at present in the house Strozzi at Ferrara) Costa himself has been confounded with his pupil Ercole Grandi di Giulio Cesare. One must, however, admit, that here the scholar has come so close to the manner of the master, that it would not perhaps be too bold to assume, that the composition of the picture comes from Costa, and only the execution belongs to Grandi.[1]

For the instruction of my young friends, I will here set

Lorenzo Costa's Shape of Hand.

Cosimo Tura's Shape of Hand.

before their eyes a facsimile of the shapes of ear and hand in Cosimo Tura and in Lorenzo Costa, that they may the

"It has been noted that the switch from content words to function words in authorship attribution studies has an interesting historic parallel in art-historic research. […] Giovanni Morelli (1816-1891) was among the first to suggest that the attribution of, for instance, a Quattrocento painting to some Italian master, could not happen based on 'content' […] Morelli thought it better to restrict an authorship analysis to discrete details such as ears, hands and feet" (Kestemont 2014)