Best practices in distance-based stylometry: Evidence using the Modern Greek corpus

GEORGE MIKROS

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS – UNIVERSITY OF MASSACHUSETTS, BOSTON

Research aims

Compile a Modern Greek corpus based on copyright-free novels with the following restrictions:

- Have at least 2 books available from each author
- Includes at least 10 authors
- All authors included are important to the development of the 19th century Greek literary production.

Explore best practices in authorship attribution using distance-based measures. Variables to consider:

- Distance metric
- Number of most frequent features
- Different features (words, characters, ngrams)
- Culling values
- Text sampling (whole texts, truncated texts etc)

The Modern Greek corpus

Training Corpus

Authors	Titles	Tokens	Types
Chatzopoulos	Fthinoporo	42,247	5,212
Christovasilis	Diigimata Ksenitias	30,486	6,129
Eftaliotis	Mazoxtra	50,071	9,489
Kondylakis	Patouxas	58,866	12,198
Mitsakis	Aftoxeir	5,597	2,308
Moraitidis	Diigimata A vol.	50,837	14,240
Nirvanas	Sinaksari	43,404	7,860
Papadiamantis	Fonissa.txt	35,229	8,381
Psycharis	Roses	89,073	12,245
Roidis	Pappisa Ioanna	76,459	16,809
Vikelas	Diigimata	51,869	11,380
	Total	**534,138**	**106,251**

Testing Corpus

Authors	Titles	Tokens	Types
Chatzopoulos	Yperanthropos	38,404	6,714
Christovasilis	Agapi	17,522	4,708
Eftaliotis	Fillades	50,942	9,800
Kondylakis	Proti agapi	30,907	7,061
Mitsakis	Oiwnos	1,523	844
Moraitidis	Diigimata B vol.	48,609	12,293
Nirvanas	Voskopoula	26,350	5,253
Papadiamantis	Emporoi	59,739	11,652
Psycharis	Taksidi	65,573	9,781
Roidis	Diigimata	53,088	15,710
Vikelas	Laras	36,597	9,012
	Total	**429,254**	**92,828**

Experiments

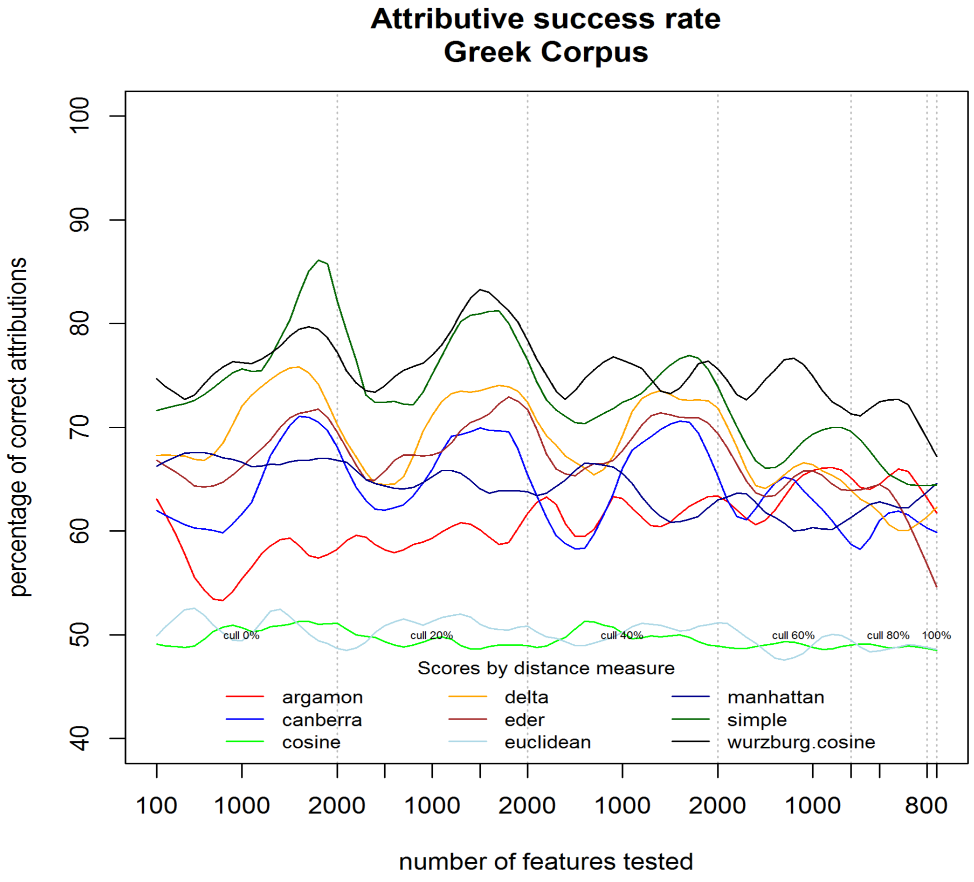
Perform multiple classification experiments using the above mentioned corpus and varying:

- Distance measures
  - Argamon, Canberra, Cosine, Delta, Eder, Euclidean, Manhattan, Simple, Wurzburg
- Culling values
  - 0%, 20%, 40%, 60%, 80%, 100%
- Number of features
  - 100 – 2000 (increment value: 100)

Analyze the classification accuracies using :

- Multi-way ANOVA
  - Main and Interaction effects
  - Post-hoc multiple comparisons

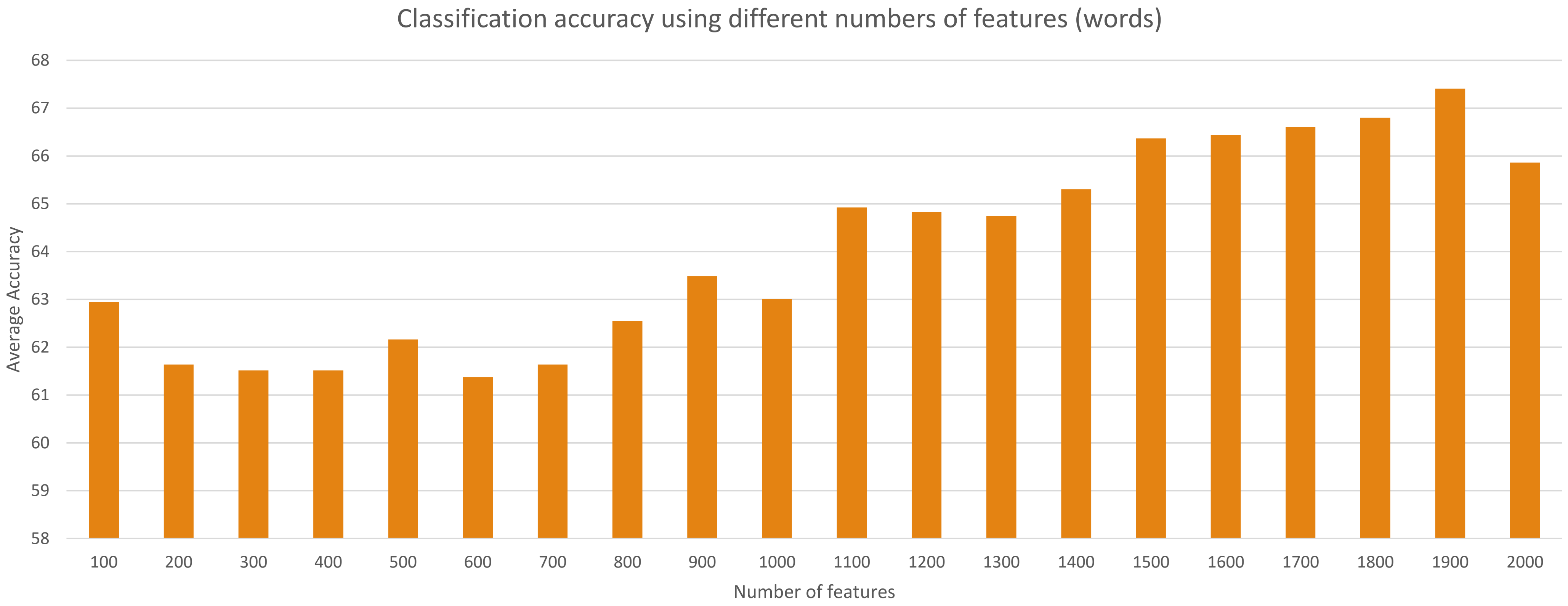
The big picture



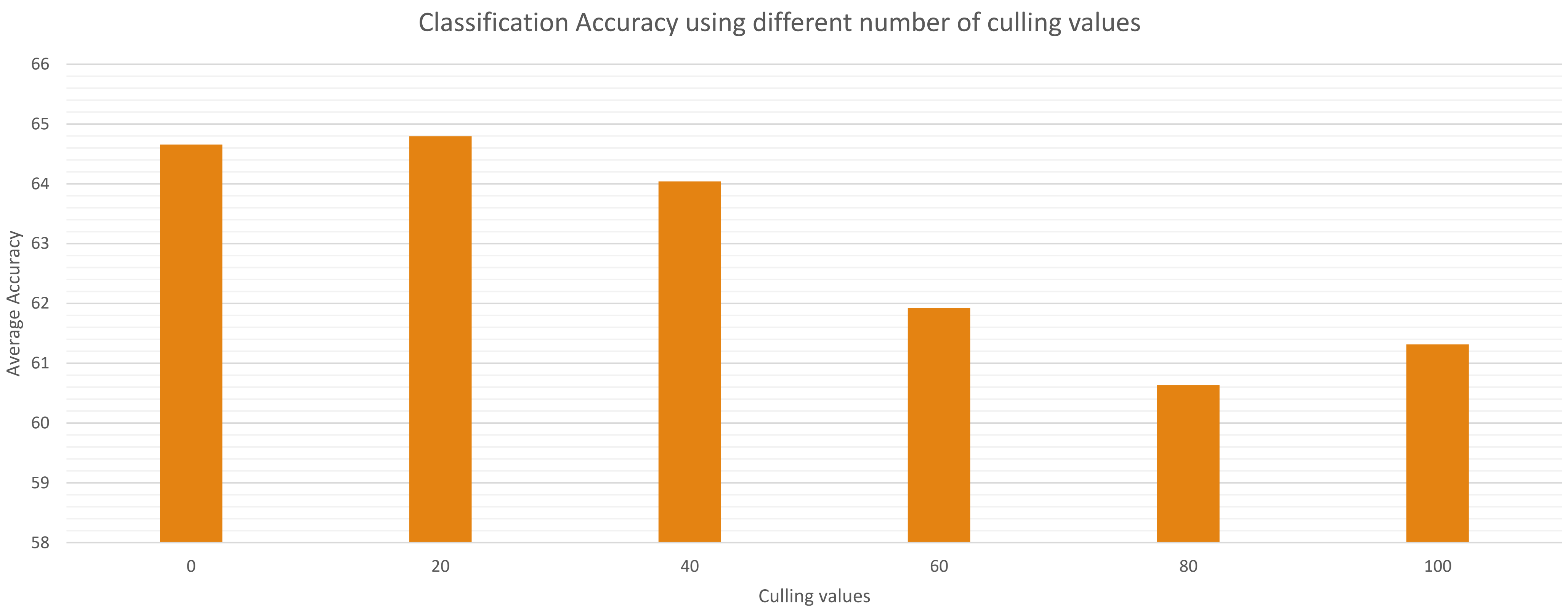
3-way ANOVA

```
Analysis of Variance Table

Response: Accuracy

              Df      Sum Sq   Mean Sq     F value     Pr(>F)

Features      1       23527    23527       278.213     < 2.2e-16 ***

Culling       5       5671     1134        13.413      4.794e-13 ***

Distance      8       540610   67576       799.098     < 2.2e-16 ***

Residuals     7455    630437   85

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

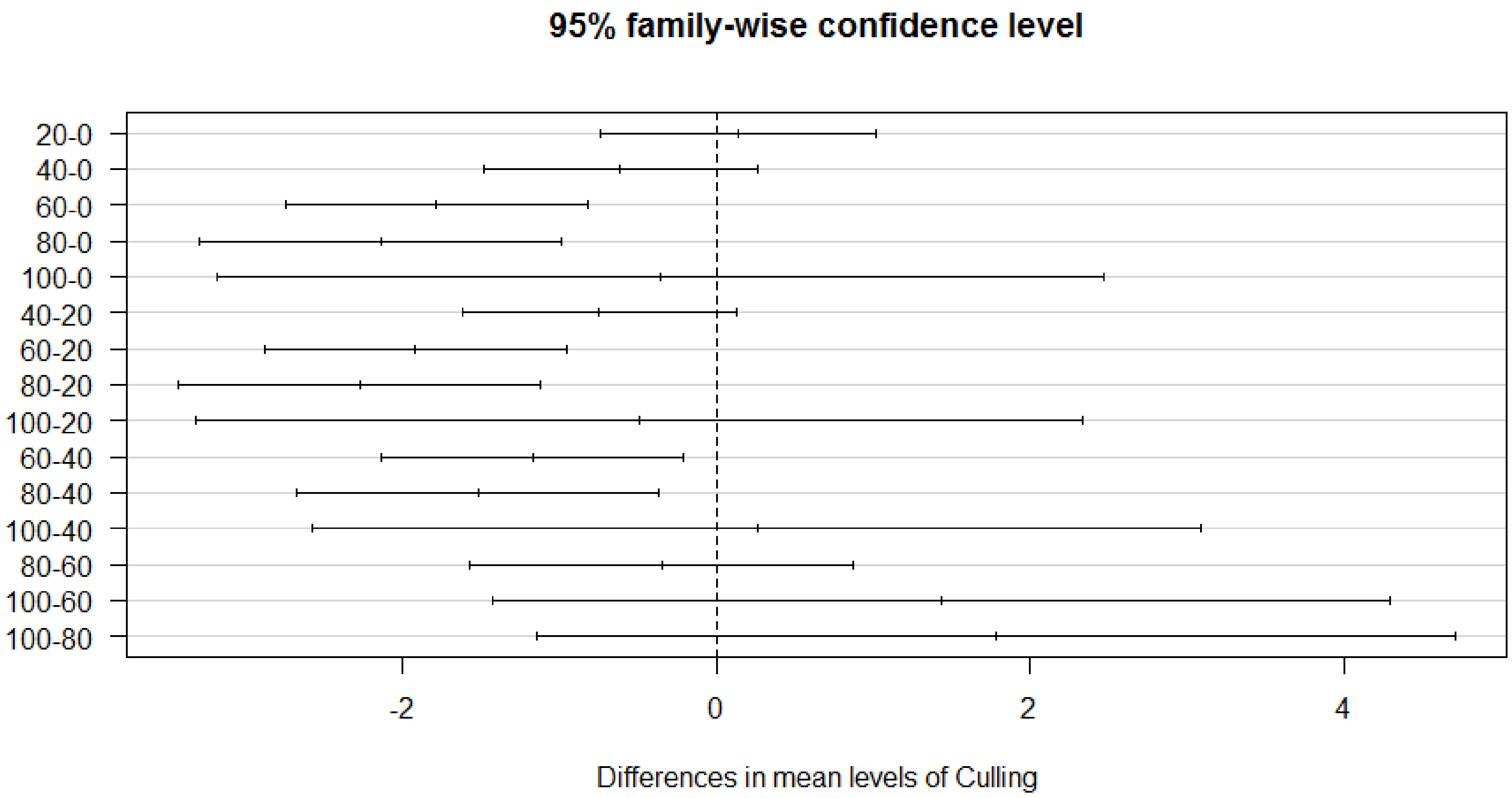
Main effects: Number of features



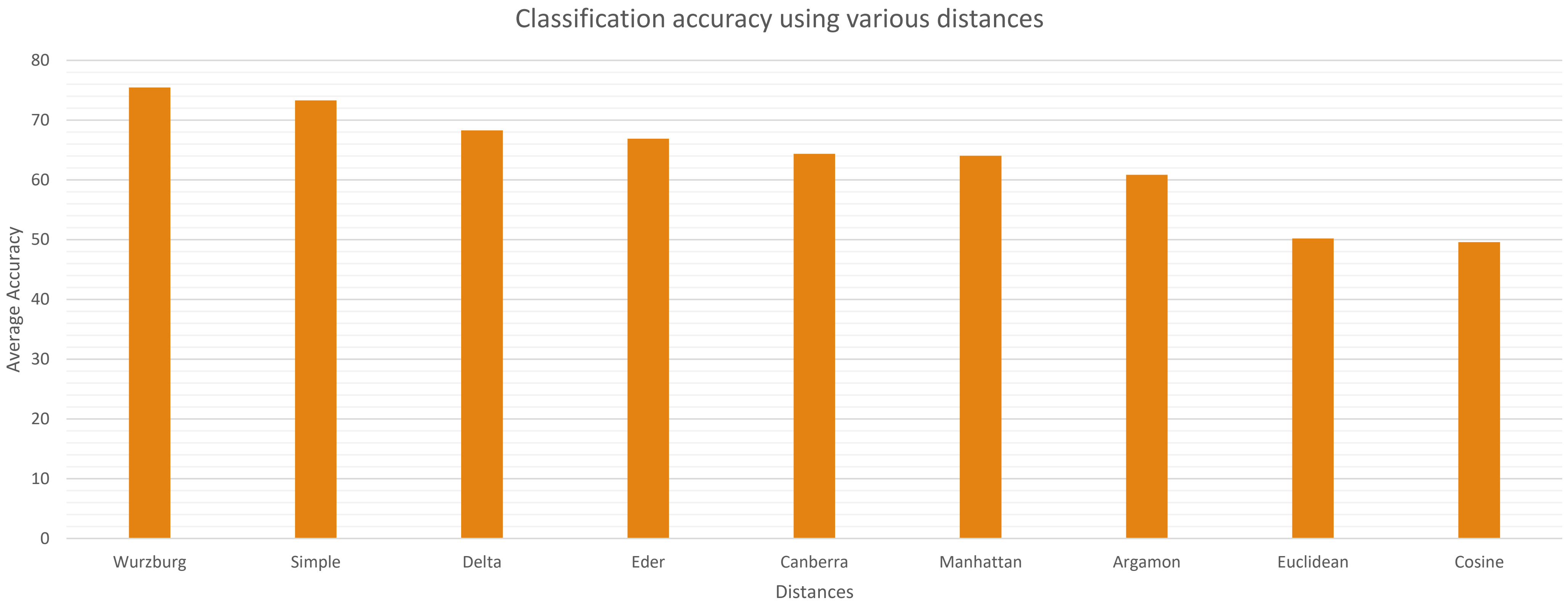
Main effects: Culling values

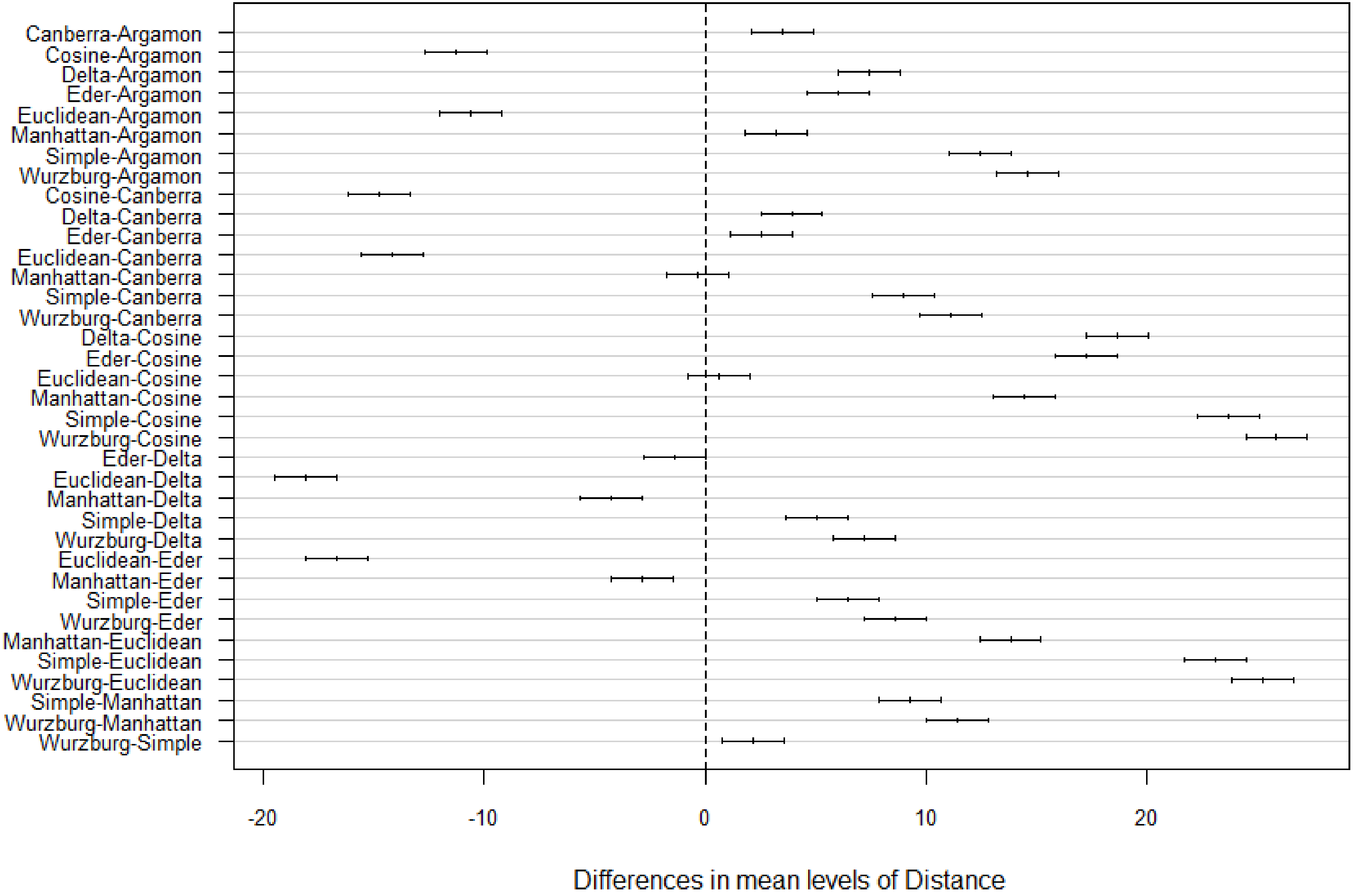


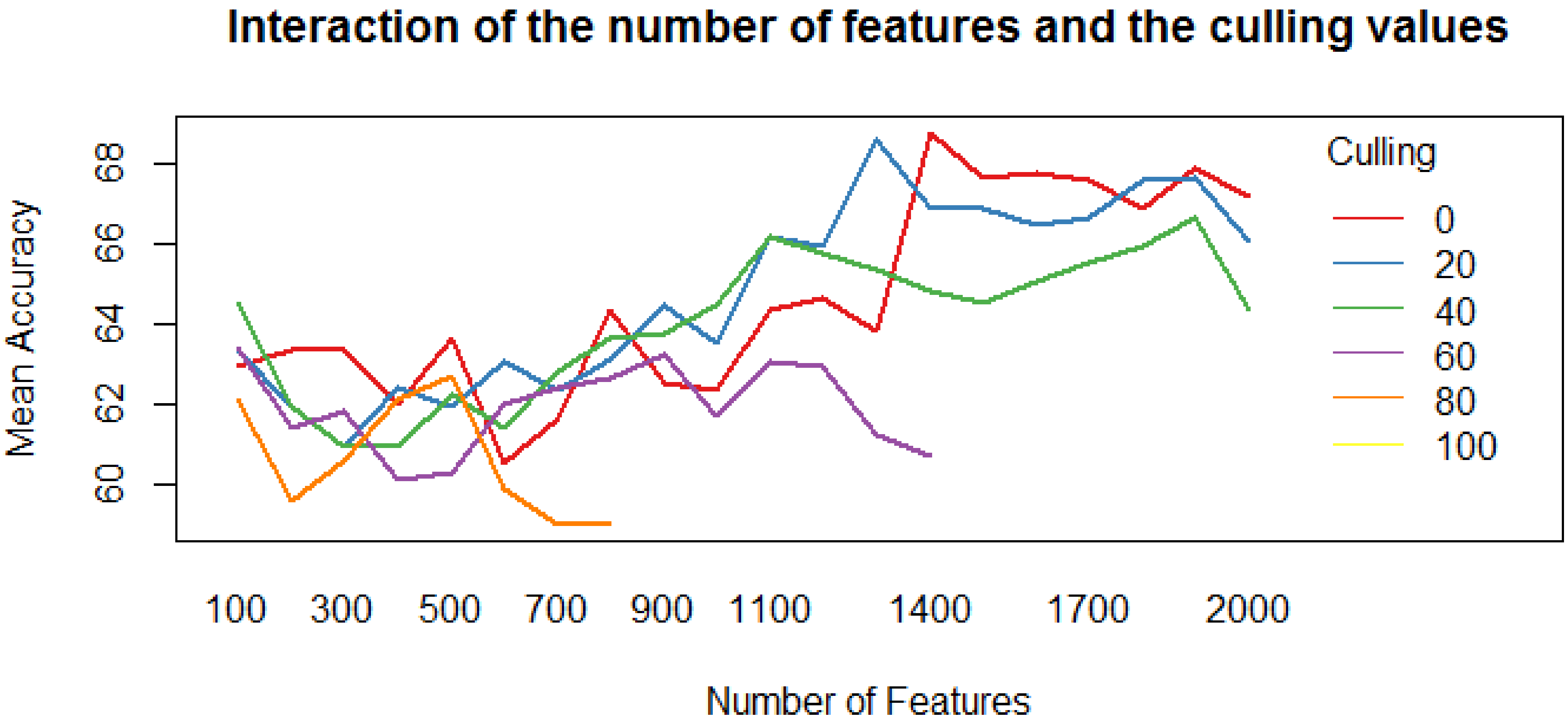
Post-hoc comparisons: Culling



Main effects: Distances



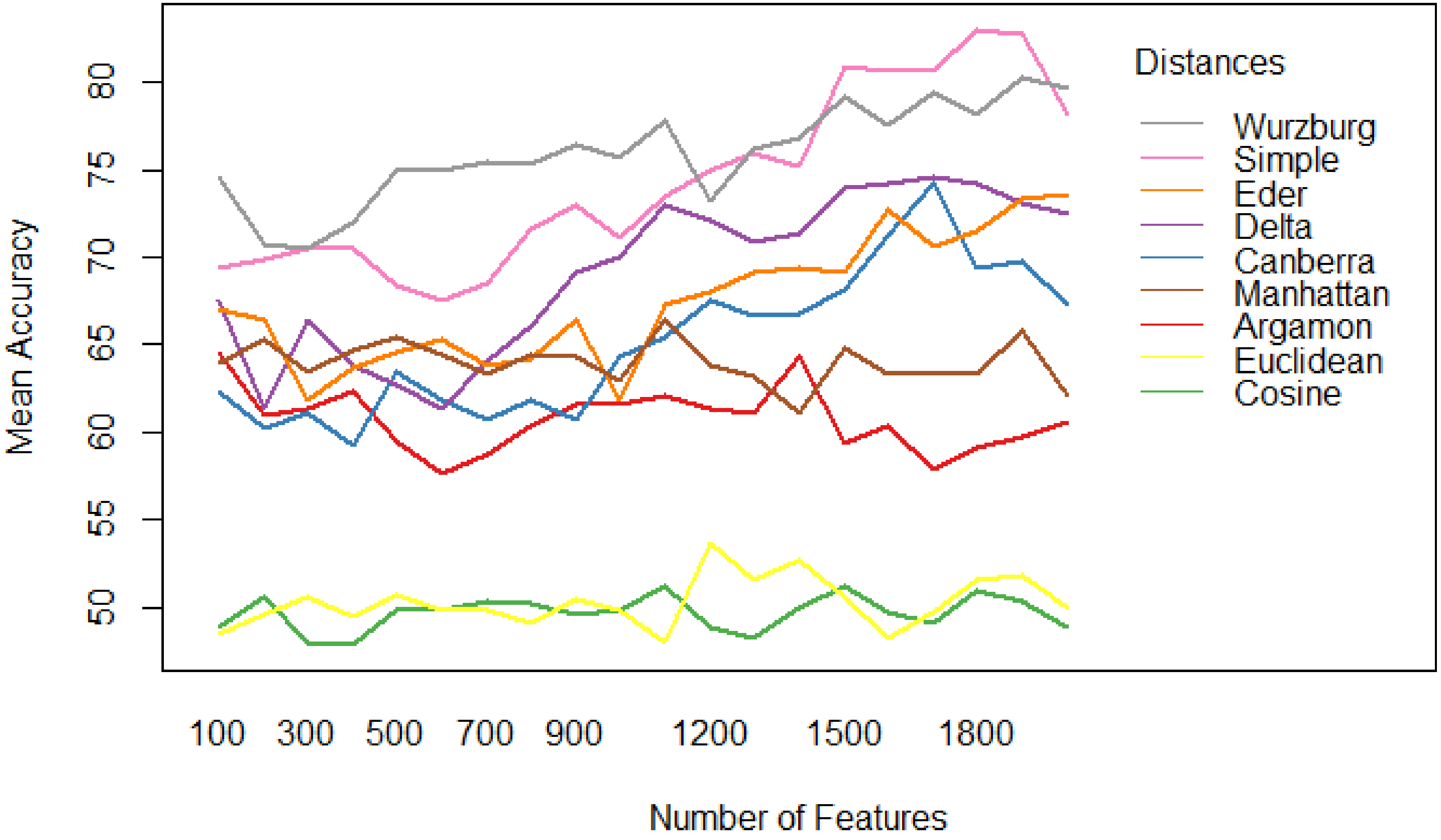

Canberra-Argamon
Cosine-Argamon
Delta-Argamon
Eder-Argamon
Euclidean-Argamon
Manhattan-Argamon
Simple-Argamon
Wurzburg-Argamon
Cosine-Canberra
Delta-Canberra
Eder-Canberra
Euclidean-Canberra
Manhattan-Canberra
Simple-Canberra
Wurzburg-Canberra
Delta-Cosine
Eder-Cosine
Euclidean-Cosine
Manhattan-Cosine
Simple-Cosine
Wurzburg-Cosine
Eder-Delta
Euclidean-Delta
Manhattan-Delta
Simple-Delta
Wurzburg-Delta
Euclidean-Eder
Manhattan-Eder
Simple-Eder
Wurzburg-Eder
Manhattan-Euclidean
Simple-Euclidean
Wurzburg-Euclidean
Simple-Manhattan
Wurzburg-Manhattan
Wurzburg-Simple
-20
-10
0
10
20
Differences in mean levels of Distance


Interaction of the number of features and the culling values
Mean Accuracy
60
62
64
66
68
100
300
500
700
900
1100
1400
1700
2000
Number of Features
Culling
0
20
40
60
80
100

Interaction effects: Culling * Features

Variables	**p**
Features:Culling20	0.86243
Features:Culling40	0.820735
Features:Culling60	**0.013127**
Features:Culling80	0.701812

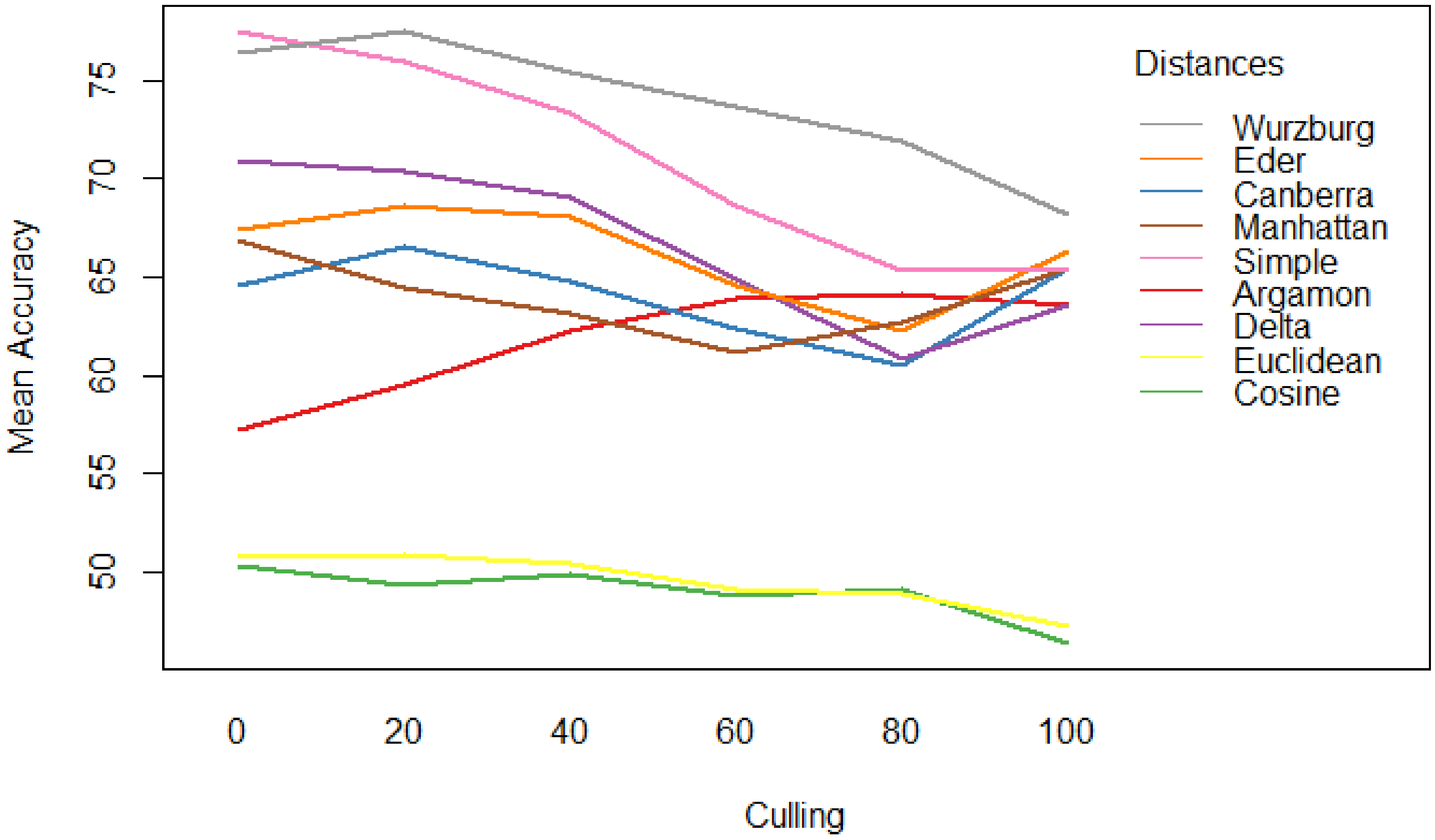
Interaction of the number of features and the distances used



Interaction effects: Distance*Features

Variables	**p**
Features:DistanceCanberra	**8.32E-07**
Features:DistanceCosine	0.220107
Features:DistanceDelta	**0.00016**
Features:DistanceEder	**0.000825**
Features:DistanceEuclidean	0.676201
Features:DistanceManhattan	0.796637
Features:DistanceSimple	**1.21E-09**
Features:DistanceWurzburg	**0.003139**

Interaction of the distances used and the culling values



Interaction effects: Distances*Culling

Variables	p
Culling80:DistanceEuclidean	0.004132
Culling40:DistanceEuclidean	0.004679
Culling40:DistanceDelta	0.006139
Culling80:DistanceManhattan	0.011323
Culling100:DistanceEuclidean	0.014905
Culling100:DistanceWurzburg	0.016457
Culling100:DistanceSimple	0.026945
Culling100:DistanceCosine	0.044966
Culling80:DistanceDelta	0.046296

More to experiment…

Word and character n-grams

Text sampling methods

N-order interactions

Develop the variation envelop of each distance metric using the above mentioned variables and in many different languages.