Extended Process Similarity Review Panel

(EPSRP)

Report for

Corresponding ISO3166 Entry: BG [BULGARIA]

A-Label: xn—90ae

U-Label: бг

Unicode Code Points: U+0431 U+0433

String in English: bg

String Language: Bulgarian Language

Scripts: Cyrillic

September 2014

Contents

Executive Summary

The Extended Process Similarity Review Panel (EPSRP) presents its recommendations on the following IDN ccTLD application:

Corresponding ISO3166 Entry: BG [BULGARIA]
A-Label: xn—90ae
U-Label: бг
Unicode Code Points: U+0431 U+0433
String in English: bg
String Language: BulgarianLanguage
Scripts: Cyrillic

The Extended Process Similarity Review Panel (EPSRP) was created under the Final Implementation Plan for IDN ccTLD Fast Track Process to provide ICANN with recommendations regarding IDN ccTLD applications being confusingly similar to ISO 3166-1 entries.

The EPSRP is composed of panel members which are internationally recognized researchers in the relevant field as well as a research team which was responsible for carrying out the experimentation.

The research team in collaboration with panel members developed an empirical evaluation methodology based on the latest scientific findings in the relevant field to determine if an applied for IDN ccTLD string should be considered confusingly similar to any ISO 3166-1 entries.

The methodology was used by the research team to establish threshold values for its tasks using ISO 3166-1 entries. All of the ISO 3166-1 are in use or potentially available as ccTLDs regardless of their potential for being confusingly similar within this group. The threshold values essentially allow for IDN ccTLD applications to be as similar as any ISO 3166-1 pair.

The methodology was then used on the applied for IDN ccTLD strings and the results compared to the threshold values to determine if they were confusingly similar or not. If the applied for IDN ccTLD in upper or lower case exceeds a threshold value for a given ISO 3166-1comparison for both tasks then it will be considered confusingly similar.

The panel provides separate recommendations for upper and lower case versions of the applied for IDN ccTLD strings given that from a visual similarity point of view upper and lower case characters of the same letter are distinct entities.

As such the Extended Process Similarity Review Panel presents the following recommendations for this application:

- The panel recommends that the IDN ccTLD application in upper case should not be considered confusingly similar to any ISO 3166-1 entries.

- The panel recommends that the IDN ccTLD application in lower case should not be considered confusingly similar to any ISO 3166-1 entries.

1 Background

The **Final Implementation Plan for IDN ccTLD Fast Track Process** (http://www.icann.org/en/resources/idn/fast-track/idn-cctld-implementation-plan-05nov13-en.pdf) instituted **the Extended Process Similarity Review Panel (EPSRP)**.

The guidelines for the EPSRP were published on 4 December 2013 and can be found at http://www.icann.org/en/resources/idn/fast-track/epsrp-guidelines-04dec13-en.pdf .

The objective of the EPSRP is described as follows in the guidelines:

> *In the event a requested string is found to be confusingly similar by the DNS Stability Panel, an external and independent Extended Process Similarity Review Panel ("EPSRP") conducts a review of the requested IDN ccTLD string, using a different framework from the DNS Stability Panel, and, only upon request of the applicant.*

2 Methodology

The methodology was developed by the research team and approved by the Panel after rigorous review.

Two tasks were selected to evaluate visual similarity:

- **Delayed match-to sample (two-alternative forced-choice) task (DMTS).** In this task, participants briefly see one candidate pairs on the screen, after which it is masked. Then, that pair plus a foil appears after a short delay, and they must identify which option was presented.
- **Go/No-go same-different task (GNG).** In this task, participants see two pairs on the screen, left and right of center, outside their central vision. They must respond only when the two differ.

For each task two evaluations of similarity were calculated from the observations, one for response time (RT) and another for response accuracy (error rate). These evaluations combined with the tasks produce four measurements:

- DMTS inv(RT)
- DMTS error rate
- GNG inv(RT)
- GNG error rate

The basic testing procedure involved presenting test subjects with a number of visual stimuli which consist of 2 characters in various versions to obtain data on both tasks. Versions include variations on fonts, font types as well as upper and lower case.

This testing was initially performed on a set of ISO 3166-1 two character codes, all of which are delegated or admissible as ccTLDs, and focused on visually confusable entries to establish the threshold for each of the 4 measurements. The threshold values essentially allow for IDN ccTLD applications to be as confusingly similar as any ISO 3166-1 pair of entries.

The threshold values derived from this experimentation were:

- DMTS inv(RT) - values less than 0.9 would indicate the entry is confusingly similar.
- DMTS error rate - values greater than 0.14 would indicate the entry is confusingly similar.
- GNG inv(RT) - values less than 0.77 would indicate the entry is confusingly similar.
- GNG error rate - values greater than 0.34 would indicate the entry is confusingly similar.

Further testing, which included the requested IDN ccTLD string against a number of ISO 3166-1 entries (selected for their potential for confusion with the requested string – see Section 6 of this report for details), was also carried out to generate measurements for this string for each version.

For an applied for string to be considered confusingly similar, there must be evidence that the candidate is highly similar to potentially-confusing ISO 3166-1 entries for both behavioral tasks. The DMTS task assesses memory confusion after brief delays, whereas the GNG task assesses the potential confusion of simultaneous glyphs.

For a given task, highly-similar refers to one or to both measures (Inv RT and error rate) exceeding the established threshold criterion (to exceed a given threshold both the mean and the 95% confidence interval must exceed the threshold). If only one of these two measures (invRT or error rate) exceeds threshold this is sufficient evidence for rejection for this task provided that the result cannot be due to a speed-accuracy trade-off. This pattern does not need to be in same font face for the given testing pair combination in both tasks.

Notes:

- This is simply a summary of the methodology that was developed by the research team in collaboration with the Panel to evaluate the candidate strings. A complete description of the methodology and the results can be found in the annexes of this document.
- Separate recommendations for upper and lower case versions of the candidate string. The Panel was requested to consider both upper and lower case versions of the candidate strings to evaluate if it is confusingly similar to any ISO 3166-1 entry in both upper and lower case. From a visual similarity point of view upper and lower case characters of the same letter are distinct entities – as such upper and lower case versions of the candidate strings needed to be tested separately. Given there is no scientific or policy basis as to how to combine these separate results of upper and lower case for IDN ccTLDs the Panel concluded it could only provide separate recommendations for each of these.

3 Panel Members and Research Team

Dr. Max Coltheart (chair), Emeritus Professor, Department of Cognitive Science, Macquarie University, Australia

Dr. Jonathan Grainger, Directeur de recherches au CNRS Aix-Marseille Université, France

Dr. Kevin Larson, United States

Research Institute: Department of Cognitive and Learning Sciences, Michigan Technological University, United States ; Leader of the research team: Professor Dr. Shane T. Mueller

4 Information on string to evaluate

Corresponding ISO3166 Entry: BG [BULGARIA]
A-Label: xn—90ae
U-Label: бг
Unicode Code Points: U+0431 U+0433
String in English: bg
String Language: BulgarianLanguage
Scripts: Cyrillic

5 Documents provided to the panel by ICANN

Submitted to the panel by ICANN:

- EPSRP Application form
- BG IDN Tables

Submitted by the applicant in the 30 day window following the application:

- None

Documents requested by the panel:

- None

Other documents:

- DNS Stability Evaluation results – original application

6 Research Report Summary

The following is a summary of the research report for the string being considered.

The complete research report, which was submitted to the EPSRP by Dr. Mueller can be found in Annex A of this document.

The following is a listing of the version information as well as the characters used in the experimentation for this application:

6.1 Stimuli for Candidate: бг/ БГ in Cyrillic

	Serif lowercase **Times New Roman**	Sans serif lowercase **Segoe UI**
Evaluation target	бг	бг
Similar Latin	br bt	br bt
Dissimilar Latin comparisons:	nk ja ld	nk ja ld
Other highly similar comparisons	бτ 6r 6г	бτ 6r 6г

	Garamond Cyrillic
Evaluation Target	бг
Similar Latin	bs
Dissimilar Latin comparisons:	gk ld
Other highly similar comparisons	бτ 6s

	Serif uppercase Times new roman	Sans serif uppercase Segoe UI Uppercase
Evaluation Target	БГ	БГ
Similar Latin	BT BF	BT BF
Dissimilar Latin comparisons:	KD OS AK	KD OS AK
Other Highly similar comparisons	БТ ЪТ ЪГ	БТ ЪТ ЪГ

Note: Some non-Latin character pairs were tested in early experimentation but these were not considered in the final analysis.

6.2 Results

The following is a summary of the results obtained.

6.2.1 DMTS

Summary of invRT below threshold (if both are below 0.9 then the result is a fail - bold)

Pair:	Fontface	Mean	Confidence interval
BT	***Sans Uppercase***	***0.87***	***0.897***
BF	***Sans Uppercase***	***0.84***	***0.879***
BT	***Serif Uppercase***	***0.855***	***0.887***
BF	*Serif Uppercase*	*0.898*	*0.929*

Italic indicates mean exceeds threshold. Bold indicates mean significantly exceeds threshold.

Summary of Error rate above threshold (if both are greater than 0.14 then the result is a fail - bold)

Pair:	Fontface	Mean	Confidence interval
None			

Italic indicates mean exceeds threshold. Bold indicates mean significantly exceeds threshold.

6.2.2 Same/different go/no-go task

Summary of invRT below threshold (if both are below 0.77 then the result is a fail - bold)

Pair:	Fontface	Mean:	Confidence interval
BT	Sans Uppercase	0.704	0.771
BF	Sans Uppercase	0.726	0.798

Summary of Error rate above threshold (if both are above 0.34 then the result is a fail - bold)

Pair:	Fontface	Mean:	Confidence interval
None			

Italic indicates mean exceeds threshold. Bold indicates mean significantly exceeds threshold.

7 Analysis by panel members

The panel reviewed the research report and was satisfied that it met the requirements it set out.

The panel was requested to consider both upper and lower case versions of the candidate string to evaluate if it is confusingly similar to any ISO 3166-1 entry in both upper and lower case. From a visual similarity point of view upper and lower case characters of the same letter are distinct entities or glyphs – as such upper and lower case versions of the candidate strings needed to be tested separately. Given there is no scientific or policy basis as to how to combine these separate results of upper and lower case for IDN ccTLDs the Panel concluded it could only provide separate recommendations for each of these.

For an applied for string to be considered confusingly similar, there must be evidence that the candidate is highly similar to potentially-confusing ISO 3166-1 entries for both behavioral tasks. The DMTS task assesses memory confusion after brief delays, whereas the GNG task assesses the potential confusion of simultaneous glyphs.

For a given task, highly-similar refers to one or to both measures (Inv RT and error rate) exceeding the established threshold criterion (to exceed a given threshold both the mean and the 95% confidence interval must exceed the threshold). If only one of these two measures (invRT or error rate) exceeds threshold this is sufficient evidence for rejection for this task provided that the result cannot be due to a speed-accuracy trade-off. This pattern does not need to be in same font face for the given testing pair combination in both tasks.

The established threshold criteria are:

- DMTS inv(RT) - values less than 0.9 would indicate the entry is confusingly similar.
- DMTS error rate - values greater than 0.14 would indicate the entry is confusingly similar.
- GNG inv(RT) - values less than 0.77 would indicate the entry is confusingly similar.
- GNG error rate - values greater than 0.34 would indicate the entry is confusingly similar.

The panel considered the research results for upper case and noted that the candidate string generated no results which exceeded the thresholds in both tasks for the same comparison.

The panel also considered the research results for lower case and noted that the candidate string generated no results which exceeded the thresholds for both the mean and a 95% confidence interval.

The panel therefore concludes that the IDN ccTLD application in upper case should not be considered confusingly similar to any ISO 3166-1 entries.

The panel also concludes that the IDN ccTLD application in lower case should not be considered confusingly similar to any ISO 3166-1 entries.

Note: The full report of the EPSRP can be found in Annex B

8 Recommendations of the EPSRP

For the candidate string:

Corresponding ISO3166 Entry: BG [BULGARIA]
A-Label: xn—90ae
U-Label: бг
Unicode Code Points: U+0431 U+0433
String in English: bg

String Language: BulgarianLanguage
Scripts: Cyrillic

The panel recommends that the IDN ccTLD application in upper case should not be considered confusingly similar to any ISO 3166-1 entries.

The panel recommends that the IDN ccTLD application in lower case should not be considered confusingly similar to any ISO 3166-1 entries.

Annex A - Results of the Research Team Experimentation

Results of the Research Team Experimentation

Behavioral Evaluation of candidate 2-letter similarity using Match-to-sample task (DMTS)

Candidate: бг/ БГ in Cyrillic

This document evaluates the candidate with respect to its overall discriminability from other pairs, using a delayed match-to sample (two-alternative forced-choice) task. In this task, participants briefly see one candidate pairs on the screen, after which it is masked. Then, that pair plus a foil appears after a short delay, and they must identify which option was presented.

Note: Some non-Latin character pairs were tested but these were not considered in the final analysis.

Presentation

•Sans serif stimuli were displayed as rendered in the location bar of a popular internet browser running on Microsoft Windows. Serif and italic stimuli were obtained via screenshots from a word processing application using Times New Roman font face to match the size of the sans serif font (Approximately 10-11pt size, non-italic, non-bold with normal spacing).

•Participants were instructed to view the screen from a comfortable distance, to best match their naturalistic screen viewing conditions.

Procedures

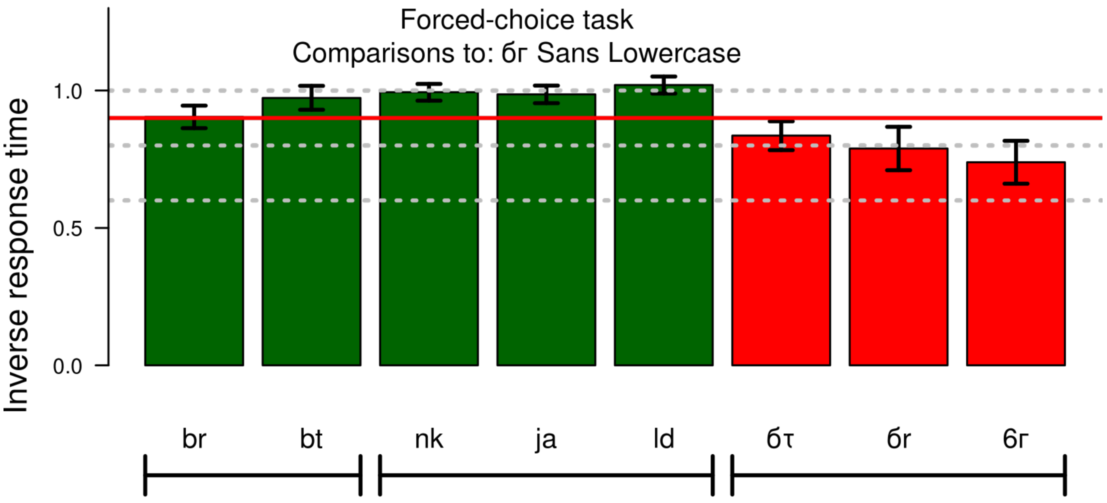
- Testing used two procedures: 1. A delayed match-to-sample forced-choice identification task, and 2. A go/no-go response same-different judgment task. The advantage of method 1 is that it tends to produce differences in response time based on confusability that are highly reliable with minimal observations, the advantage of method 2 is that it induces larger differences is accuracy, and requires a participant to detect a specific difference.
- Each test was performed in a blocked design in the same order across participants. Each set of stimuli will appear in a contiguous block. Testing was designed to assess the similarity between the target and (1) any of a set of highly-similar Latin character pairs in the same case (2) a set of 3-4 dissimilar Latin character pairs, and (3) any highly-similar comparisons, which may not directly bear on the decision, but may help to calibrate and validate the measures.

Participants

- In this study, we intend to test 20 undergraduate students, primarily students of U.S. origin. Because Cyrillic characters are relatively unfamiliar to them, and because they are experts in Latin orthography which is the orthography where the confusions are most likely to occur, they serve as a reasonable population for evaluating these characters sets to make inference about a general internet population

Inverse response time: Sans Lowercase

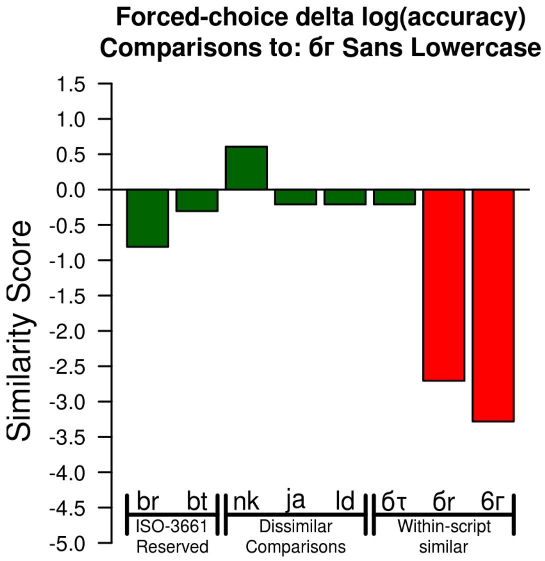
Critical value: 0.9

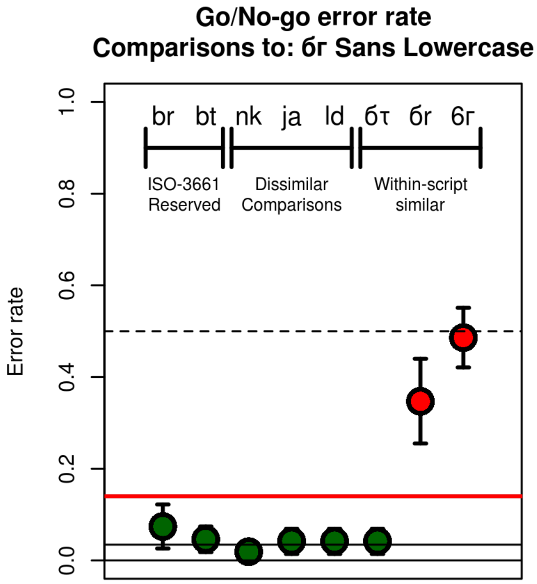


	mean:	sd:	N:	se:	5%	95%
br	0.904	0.102	27	0.02	0.863	0.945
bt	0.973	0.108	27	0.021	0.93	1.017
nk	0.994	0.076	27	0.015	0.963	1.024
ja	0.986	0.078	27	0.015	0.954	1.018
ld	1.02	0.077	27	0.015	0.989	1.051
бτ	0.836	0.13	27	0.025	0.783	0.888
бr	0.789	0.196	27	0.038	0.71	0.868
6г	0.739	0.194	27	0.037	0.661	0.817

Error rate: Sans Lowercase

Critical value: 0.14

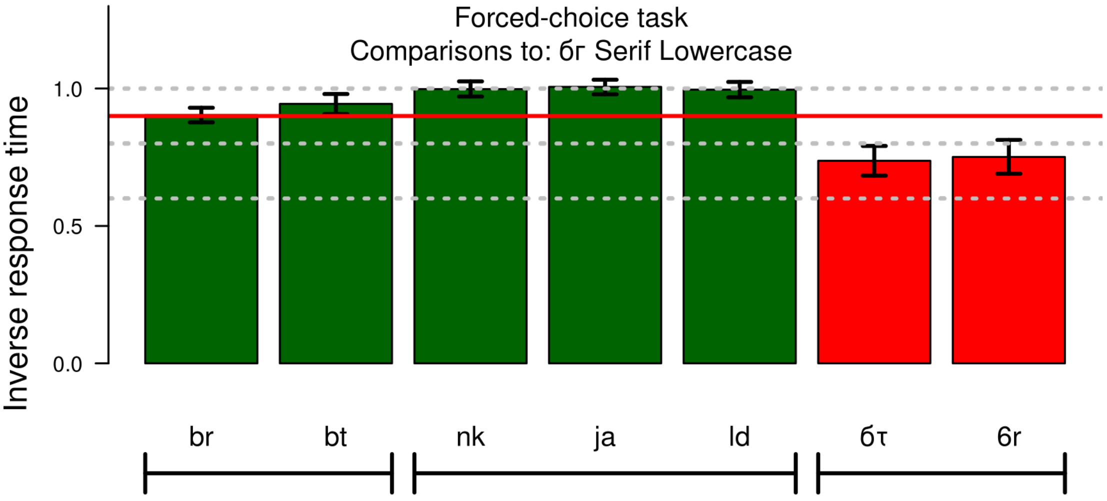




	mean:	sd:	N:	se:	5%	95%
br	0.074	0.121	27	0.023	0.026	0.122
bt	0.046	0.071	27	0.014	0.018	0.074
nk	0.019	0.045	27	0.009	0.001	0.036
ja	0.042	0.069	27	0.013	0.014	0.069
ld	0.042	0.069	27	0.013	0.014	0.069
бτ	0.042	0.069	27	0.013	0.014	0.069
бr	0.347	0.233	27	0.045	0.255	0.44
6г	0.486	0.164	27	0.032	0.421	0.551

Inverse response time: Serif Lowercase

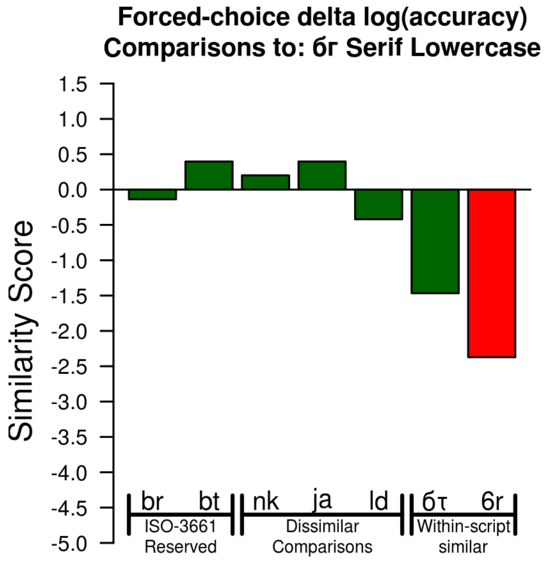
Critical value: 0.9

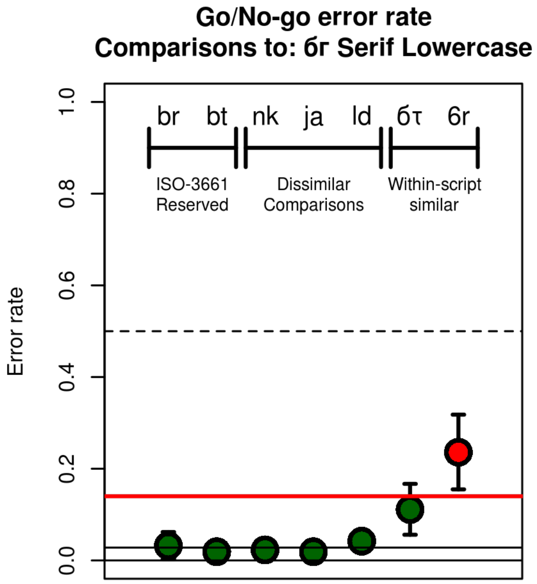


	mean:	sd:	N:	se:	5%	95%
br	0.903	0.067	27	0.013	0.877	0.93
bt	0.944	0.093	27	0.018	0.907	0.98
nk	0.998	0.07	27	0.013	0.971	1.026
ja	1.006	0.067	27	0.013	0.979	1.032
ld	0.996	0.071	27	0.014	0.968	1.024
бτ	0.737	0.136	27	0.026	0.683	0.791
6r	0.751	0.155	27	0.03	0.69	0.813

Error rate: Serif Lowercase

Critical value: 0.14

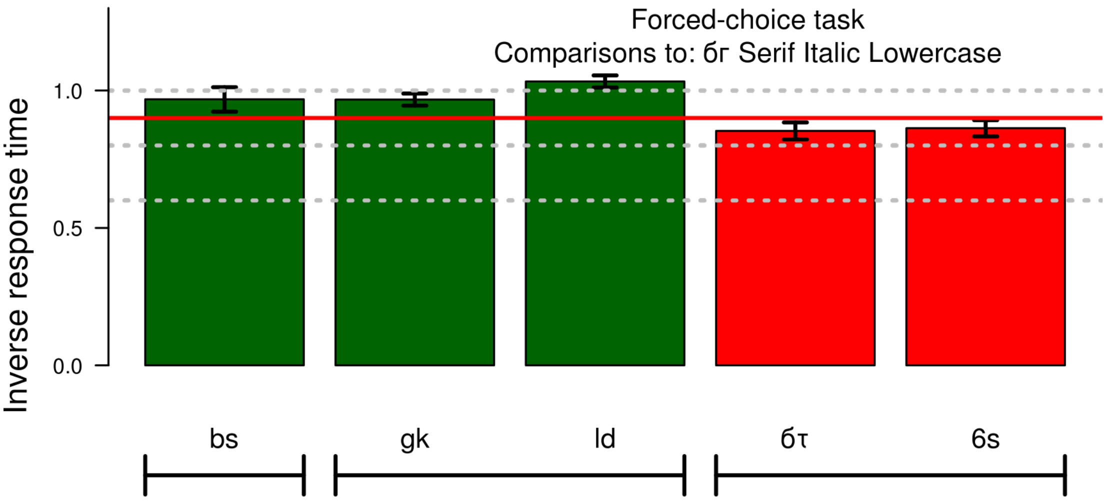




	mean:	sd:	N:	se:	5%	95%
br	0.032	0.074	27	0.014	0.003	0.062
bt	0.019	0.057	27	0.011	-0.004	0.041
nk	0.023	0.06	27	0.012	-0.001	0.047
ja	0.019	0.045	27	0.009	0.001	0.036
ld	0.042	0.06	27	0.012	0.018	0.065
бт	0.111	0.14	27	0.027	0.056	0.167
бr	0.236	0.206	27	0.04	0.155	0.318

Inverse response time: Serif Italic Lowercase

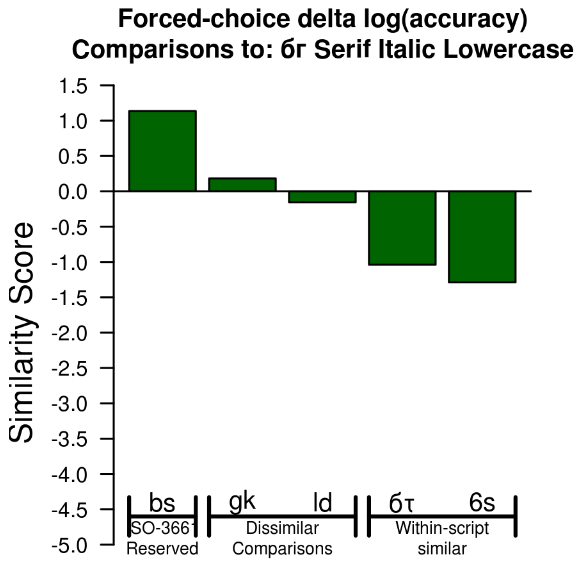
Critical value: 0.9

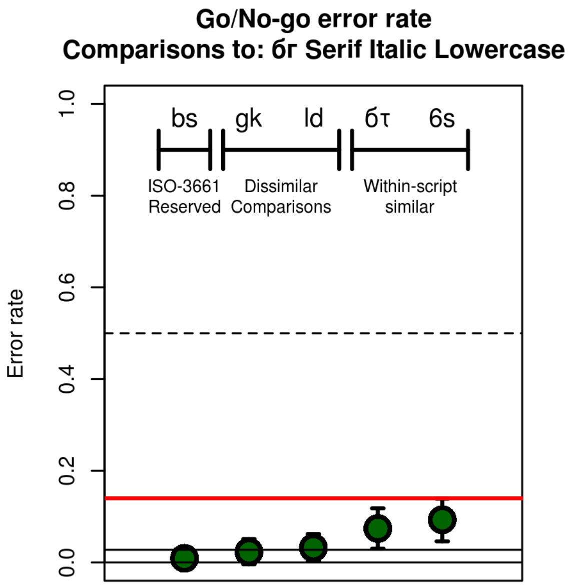


	mean:	sd:	N:	se:	5%	95%
bs	0.968	0.113	27	0.022	0.923	1.012
gk	0.967	0.056	27	0.011	0.945	0.989
ld	1.033	0.056	27	0.011	1.011	1.055
бτ	0.853	0.078	27	0.015	0.822	0.884
6s	0.863	0.074	27	0.014	0.833	0.892

Error rate: Serif Italic Lowercase

Critical value: 0.14

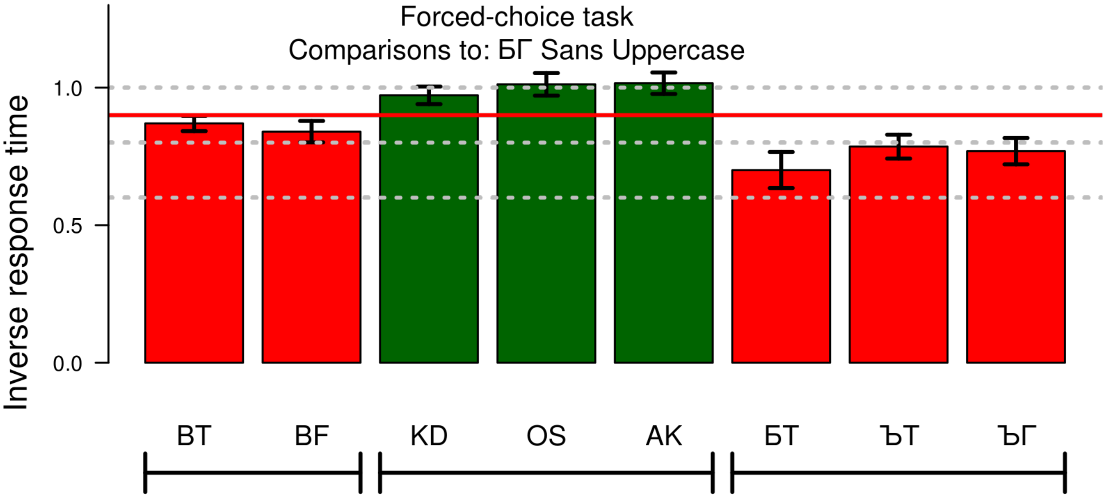




	mean:	sd:	N:	se:	5%	95%
bs	0.009	0.048	27	0.009	-0.01	0.028
gk	0.023	0.07	27	0.013	-0.004	0.051
ld	0.032	0.074	27	0.014	0.003	0.062
бτ	0.074	0.111	27	0.021	0.03	0.118
6s	0.093	0.118	27	0.023	0.046	0.139

Inverse response time: Sans Uppercase

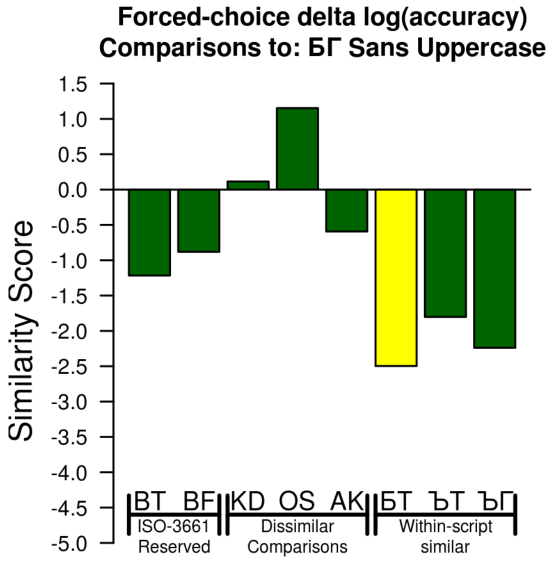
Critical value: 0.9

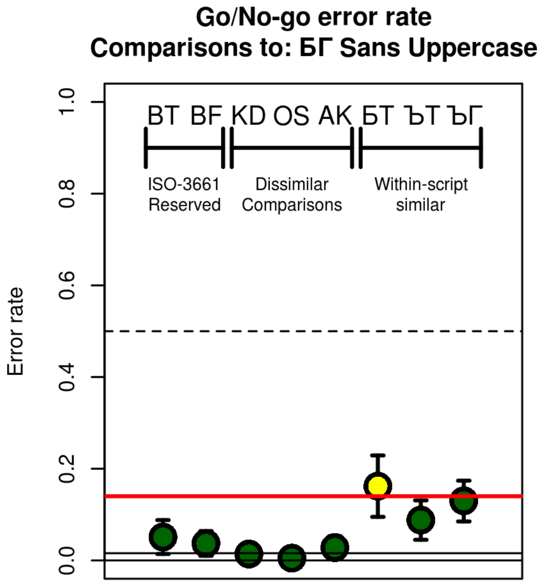


	mean:	sd:	N:	se:	5%	95%
BT	0.87	0.07	27	0.014	0.842	0.897
BF	0.84	0.098	27	0.019	0.801	0.879
KD	0.972	0.081	27	0.016	0.94	1.004
OS	1.012	0.104	27	0.02	0.971	1.053
AK	1.016	0.099	27	0.019	0.977	1.055
БТ	0.7	0.165	27	0.032	0.635	0.766
ЪТ	0.786	0.111	27	0.021	0.742	0.829
ЪГ	0.769	0.121	27	0.023	0.721	0.817

Error rate: Sans Uppercase

Critical value: 0.14

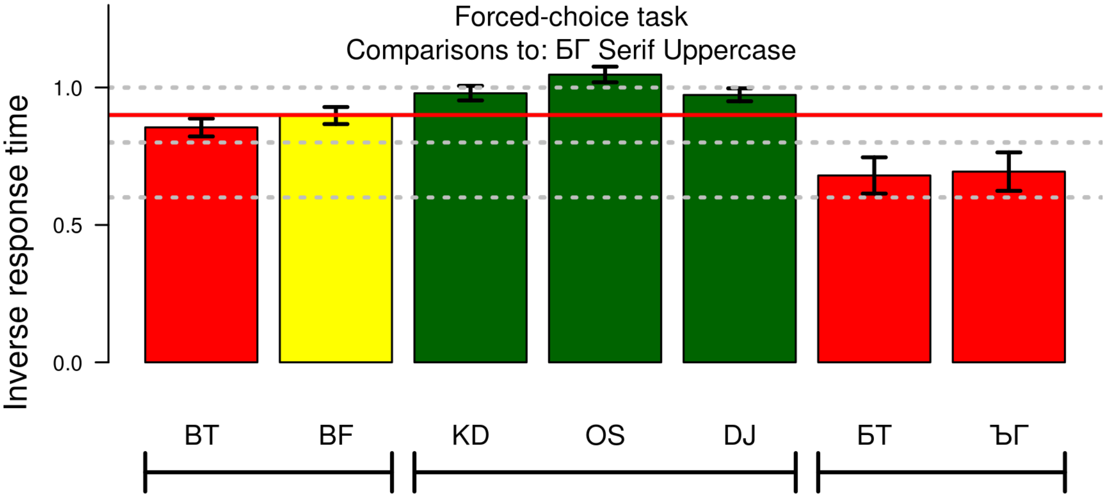




	mean:	sd:	N:	se:	5%	95%
BT	0.051	0.093	27	0.018	0.014	0.088
BF	0.037	0.068	27	0.013	0.01	0.064
KD	0.014	0.04	27	0.008	-0.002	0.03
OS	0.005	0.024	27	0.005	-0.005	0.014
AK	0.028	0.063	27	0.012	0.003	0.053
БТ	0.162	0.169	27	0.033	0.095	0.229
ЪТ	0.088	0.109	27	0.021	0.045	0.131
ЪГ	0.13	0.112	27	0.022	0.085	0.174

Inverse response time: Serif Uppercase

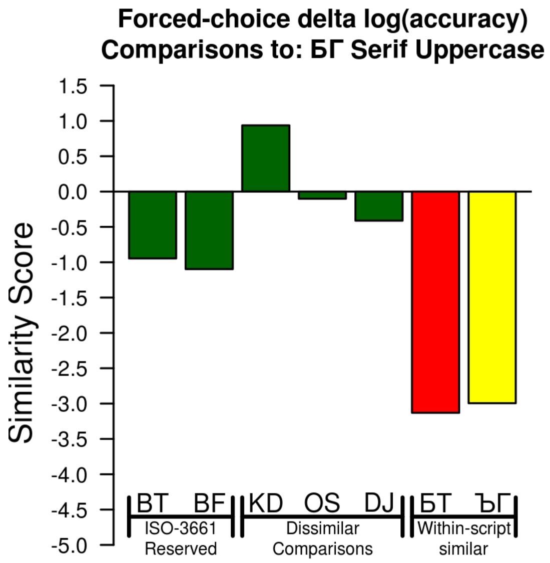
Critical value: 0.9

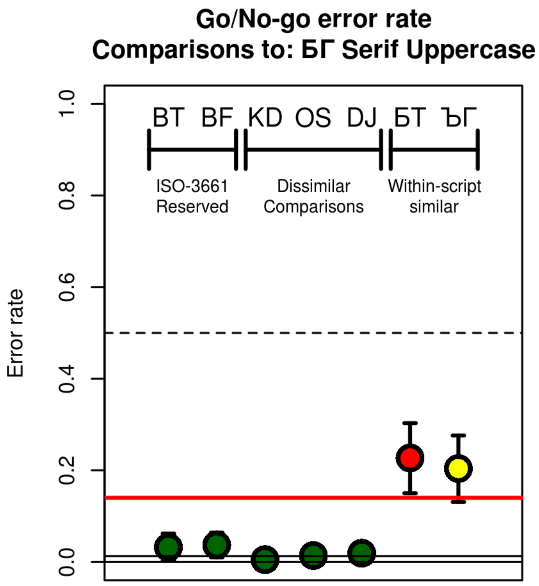


	mean:	sd:	N:	se:	5%	95%
BT	0.855	0.08	27	0.015	0.822	0.887
BF	0.898	0.076	27	0.015	0.867	0.929
KD	0.979	0.066	27	0.013	0.953	1.006
OS	1.047	0.071	27	0.014	1.019	1.076
DJ	0.973	0.059	27	0.011	0.95	0.997
БТ	0.68	0.164	27	0.032	0.614	0.746
ЪГ	0.694	0.173	27	0.033	0.624	0.764

Error rate: Serif Uppercase

Critical value: 0.14





	mean:	sd:	N:	se:	5%	95%
BT	0.032	0.074	27	0.014	0.003	0.062
BF	0.037	0.068	27	0.013	0.01	0.064
KD	0.005	0.024	27	0.005	-0.005	0.014
OS	0.014	0.04	27	0.008	-0.002	0.03
DJ	0.019	0.045	27	0.009	0.001	0.036
БТ	0.227	0.193	27	0.037	0.15	0.303
ЪГ	0.204	0.184	27	0.035	0.131	0.276

Summary of RT below threshold

Pair:	Fontface	Mean:	Confidence interval	< 0.9
BT	***Sans Uppercase***	***0.87***	***0.897***	
BF	***Sans Uppercase***	***0.84***	***0.879***	
BT	***Serif Uppercase***	***0.855***	***0.887***	
BF	*Serif Uppercase*	*0.898*	*0.929*	

Italic indicates mean surpasses threshold. Bold indicates mean significantly surpasses threshold.

Summary of Error rate above threshold

Pair:	Fontface	Mean:	Confidence interval	> 0.14

None

Italic indicates mean surpasses threshold. Bold indicates mean significantly surpasses threshold.

Behavioral Evaluation of candidate 2-letter similarity using Same/different go/no-go task

Candidate: бг/ БГ in Cyrillic)

This document evaluates the candidate with respect to its overall discriminability from other pairs, using a Go/No-go same-different task. In this task, participants see two pairs on the screen, left and right of center, outside their central vision. They must respond only when the two differ.

Note: Some non-Latin character pairs were tested but not considered in the final analysis.

Presentation

•Sans serif stimuli were displayed as rendered in the location bar of a popular internet browser running on Microsoft Windows. Serif and italic stimuli were obtained via screenshots from a word processing application using Times New Roman font face to match the size of the sans serif font (Approximately 10-11pt size, non-italic, non-bold with normal spacing).

•Participants were instructed to view the screen from a comfortable distance, to best match their naturalistic screen viewing conditions.

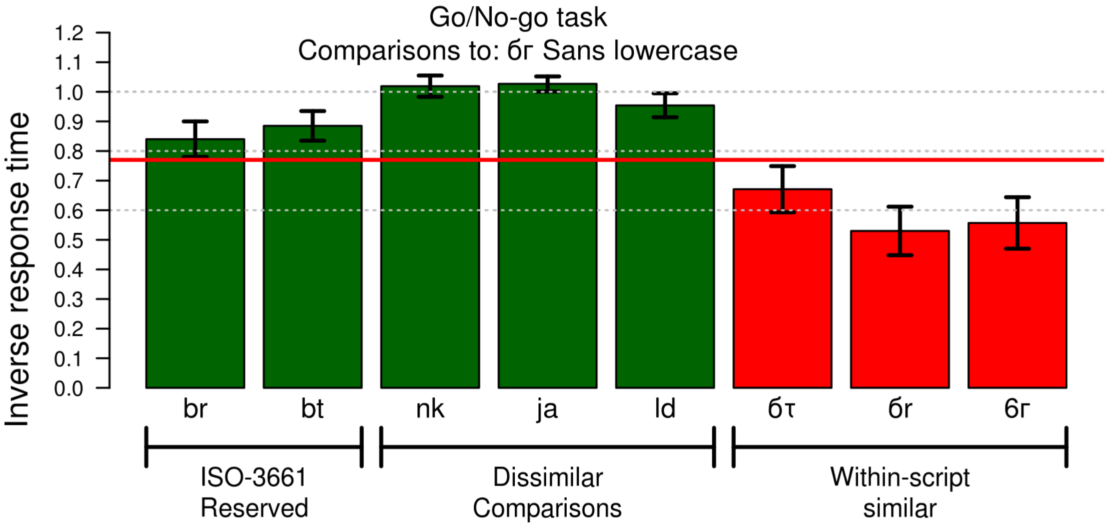
Procedures

- Testing used two procedures: 1. A delayed match-to-sample forced-choice identification task, and 2. A go/no-go response same-different judgment task. The advantage of method 1 is that it tends to produce differences in response time based on confusability that are highly reliable with minimal observations, the advantage of method 2 is that it induces larger differences is accuracy, and requires a participant to detect a specific difference.
- Each test was performed in a blocked design in the same order across participants. Each set of stimuli will appear in a contiguous block. Testing was designed to assess the similarity between the target and (1) any of a set of highly-similar Latin character pairs in the same case (2) a set of 3-4 dissimilar Latin character pairs, and (3) any highly-similar comparisons, which may not directly bear on the decision, but may help to calibrate and validate the measures.

Participants

- In this study, we intend to test 20 undergraduate students, primarily students of U.S. origin. Because Cyrillic characters are relatively unfamiliar to them, and because they are experts in Latin orthography which is the orthography where the confusions are most likely to occur, they serve as a reasonable population for evaluating these characters sets to make inference about a general internet population

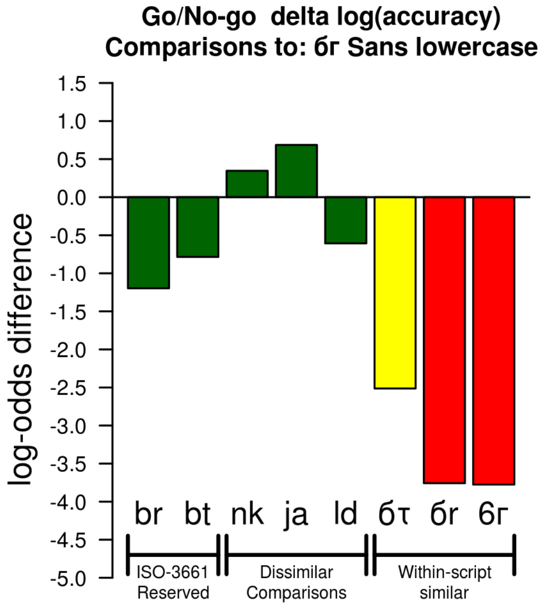
- **Inverse response time: Sans lowercase**
- Critical value: 0.77

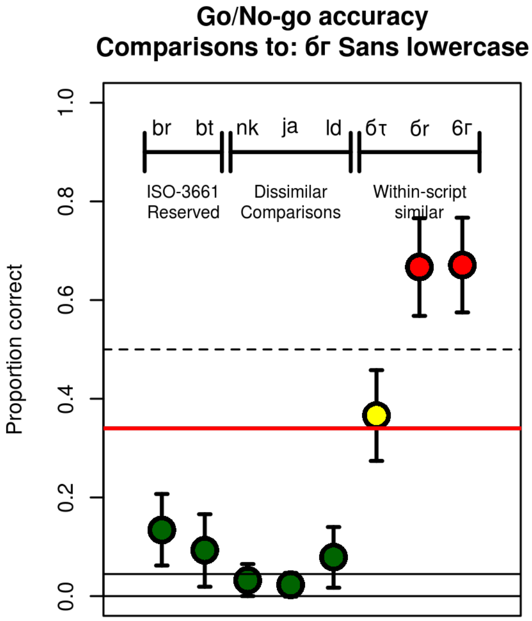


-

	mean	sd	N	se	5%	95%
br	0.84	0.153	27	0.029	0.78	0.9
bt	0.885	0.126	27	0.024	0.835	0.935
nk	1.019	0.091	27	0.017	0.983	1.055
ja	1.027	0.064	27	0.012	1.001	1.052
ld	0.954	0.102	27	0.02	0.914	0.995
бτ	0.671	0.197	27	0.038	0.593	0.749
бr	0.53	0.208	27	0.04	0.448	0.612
6г	0.557	0.22	27	0.042	0.47	0.644

- **Error rate: Sans lowercase**
- Critical value: 0.34

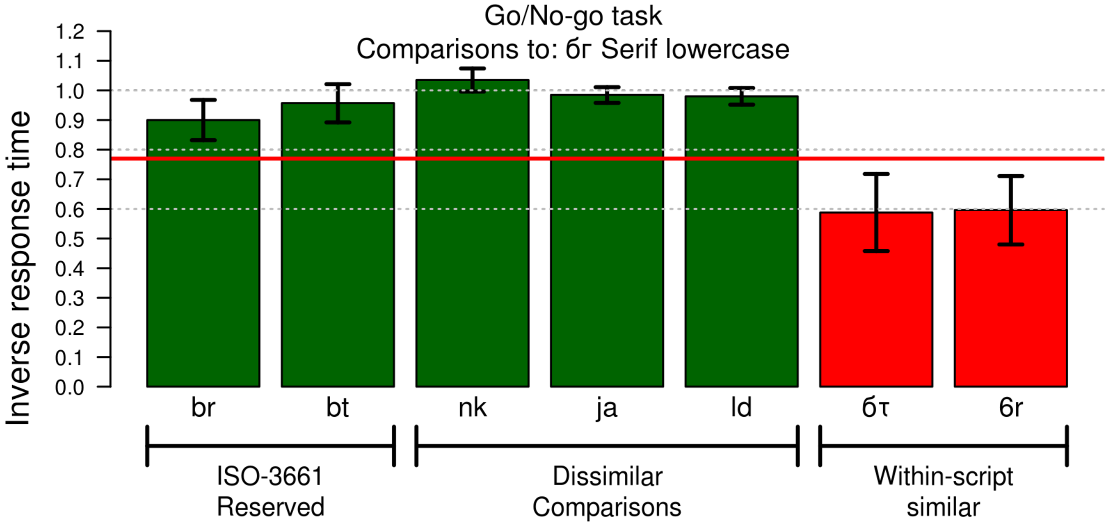

Go/No-go delta log(accuracy)
Comparisons to: бг Sans lowercase


Go/No-go accuracy
Comparisons to: бг Sans lowercase

-

	mean:	sd:	N:	se:	5%	95%
br	0.134	0.183	27	0.035	0.062	0.207
bt	0.093	0.185	27	0.036	0.019	0.166
nk	0.032	0.082	27	0.016	0	0.065
ja	0.023	0.06	27	0.012	-0.001	0.047
ld	0.079	0.156	27	0.03	0.017	0.14
бτ	0.366	0.232	27	0.045	0.274	0.458
бr	0.667	0.25	27	0.048	0.568	0.766
6г	0.671	0.243	27	0.047	0.575	0.767

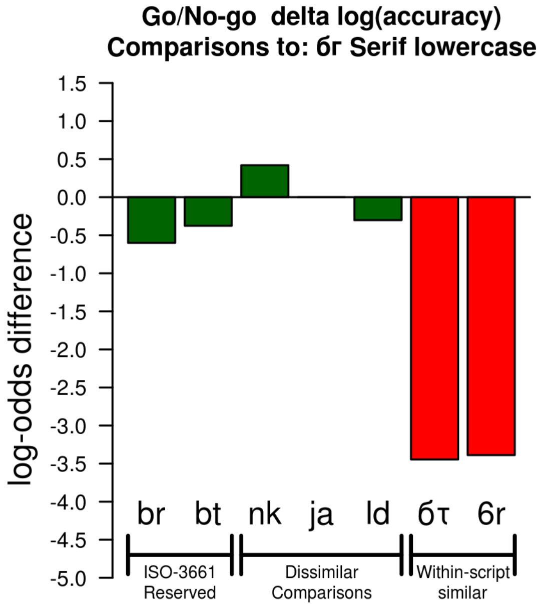
- **Inverse response time: Serif lowercase**
- Critical value: 0.77

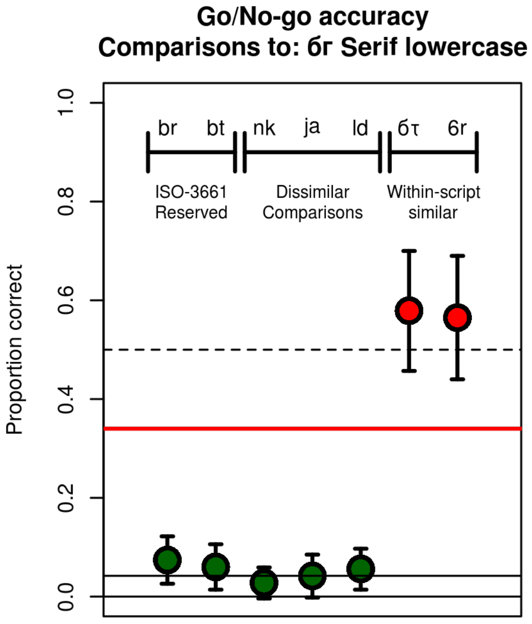


-

	mean	sd	N	se	5%	95%
br	0.9	0.173	27	0.033	0.832	0.968
bt	0.957	0.163	27	0.031	0.892	1.021
nk	1.035	0.099	27	0.019	0.996	1.074
ja	0.985	0.068	27	0.013	0.958	1.011
ld	0.98	0.07	27	0.014	0.952	1.008
бτ	0.588	0.328	27	0.063	0.458	0.718
6г	0.596	0.293	27	0.056	0.48	0.711

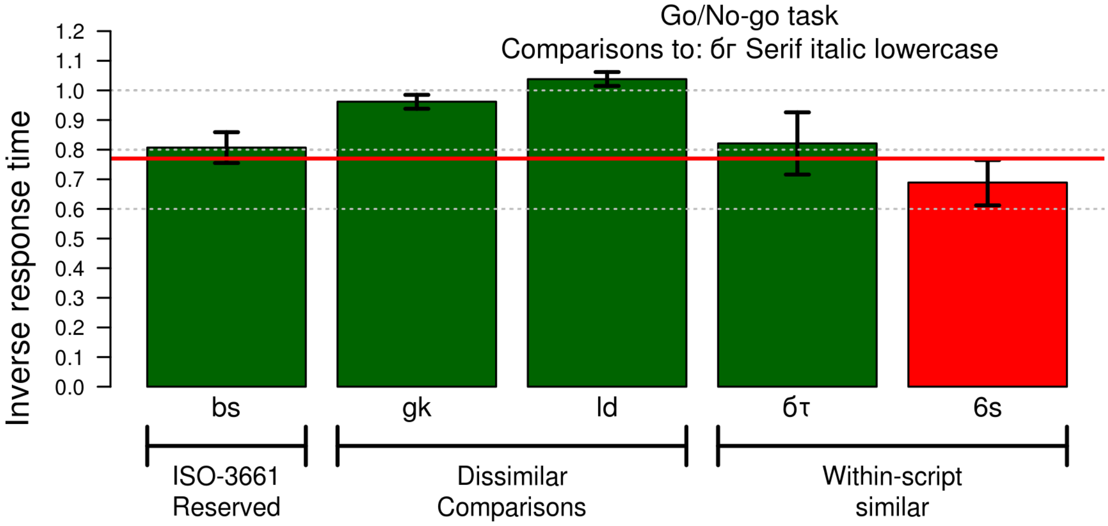
- **Error rate: Serif lowercase**
- Critical value: 0.34





	mean:	**sd:**	**N:**	**se:**	**5%**	**95%**
br	0.074	0.121	27	0.023	0.026	0.122
bt	0.06	0.117	27	0.022	0.014	0.106
nk	0.028	0.08	27	0.015	-0.004	0.059
ja	0.042	0.11	27	0.021	-0.002	0.085
ld	0.056	0.106	27	0.02	0.014	0.097
бτ	0.579	0.306	27	0.059	0.457	0.7
6r	0.565	0.317	27	0.061	0.44	0.69

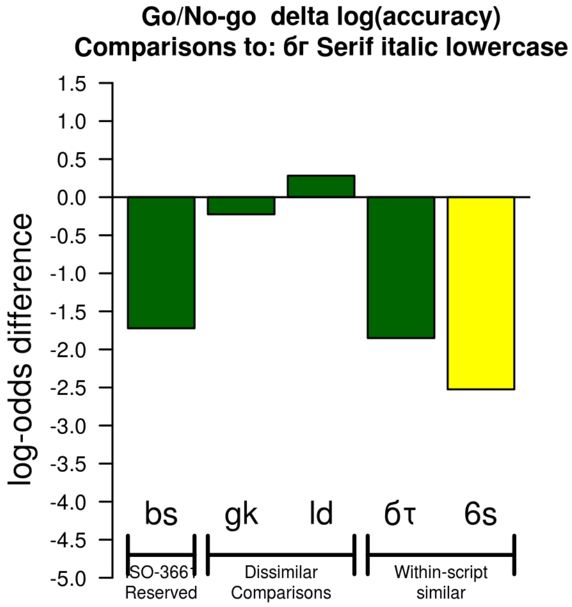
- **Inverse response time: Serif italic lowercase**
- Critical value: 0.77

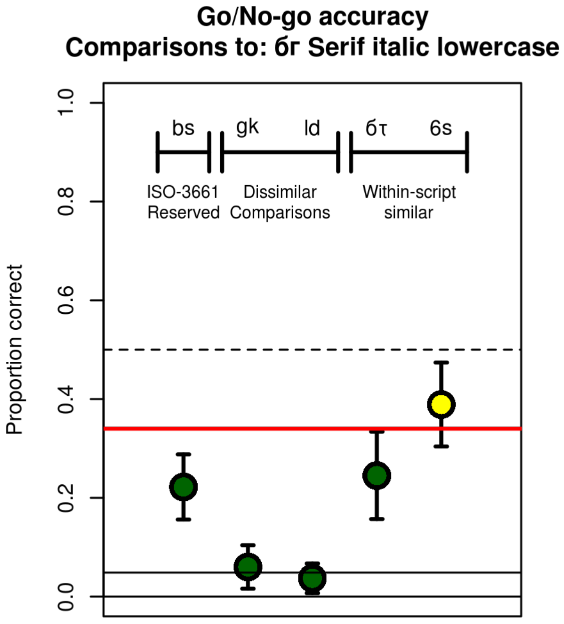


-

	mean	sd	N	se	5%	95%
bs	0.807	0.132	27	0.025	0.755	0.859
gk	0.962	0.06	27	0.012	0.938	0.985
ld	1.038	0.06	27	0.012	1.015	1.062
бτ	0.821	0.265	27	0.051	0.716	0.926
6s	0.689	0.194	27	0.037	0.612	0.765

- **Error rate: Serif italic lowercase**
- Critical value: 0.34

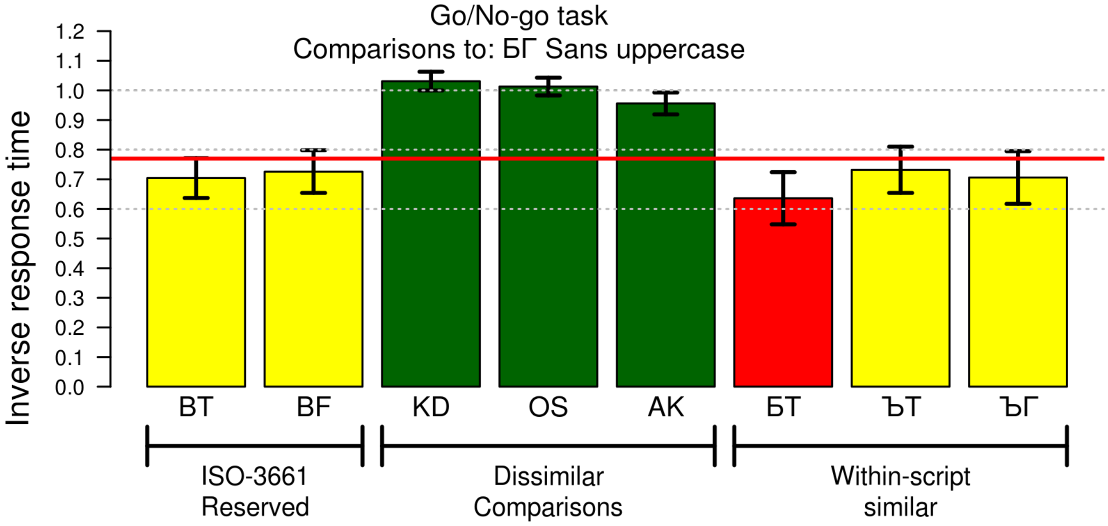




-

	mean:	sd:	N:	se:	5%	95%
bs	0.222	0.167	27	0.032	0.156	0.288
gk	0.06	0.112	27	0.021	0.016	0.104
ld	0.037	0.076	27	0.015	0.007	0.067
бτ	0.245	0.223	27	0.043	0.157	0.334
6s	0.389	0.215	27	0.041	0.304	0.474

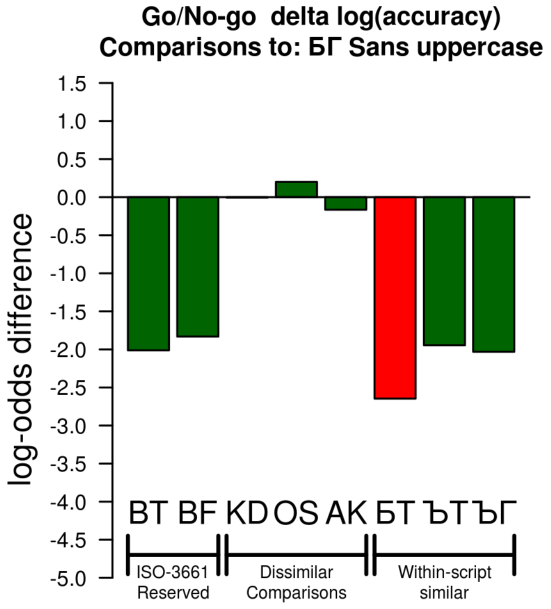
- **Inverse response time: Sans uppercase**
- Critical value: 0.77

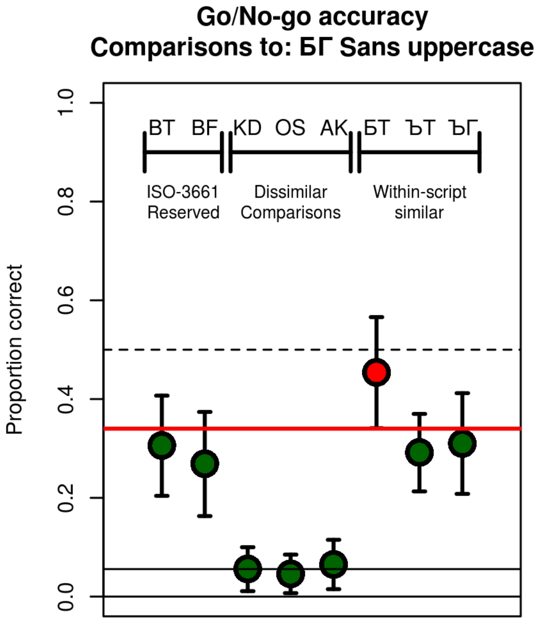


-

	mean	sd	N	se	5%	95%
BT	0.704	0.169	27	0.033	0.637	0.771
BF	0.726	0.182	27	0.035	0.654	0.798
KD	1.031	0.08	27	0.015	1	1.063
OS	1.013	0.076	27	0.015	0.983	1.043
AK	0.956	0.094	27	0.018	0.919	0.993
БТ	0.636	0.223	27	0.043	0.548	0.724
ЪТ	0.732	0.198	27	0.038	0.654	0.81
ЪГ	0.706	0.225	27	0.043	0.617	0.795

- **Error rate: Sans uppercase**
- Critical value: 0.34

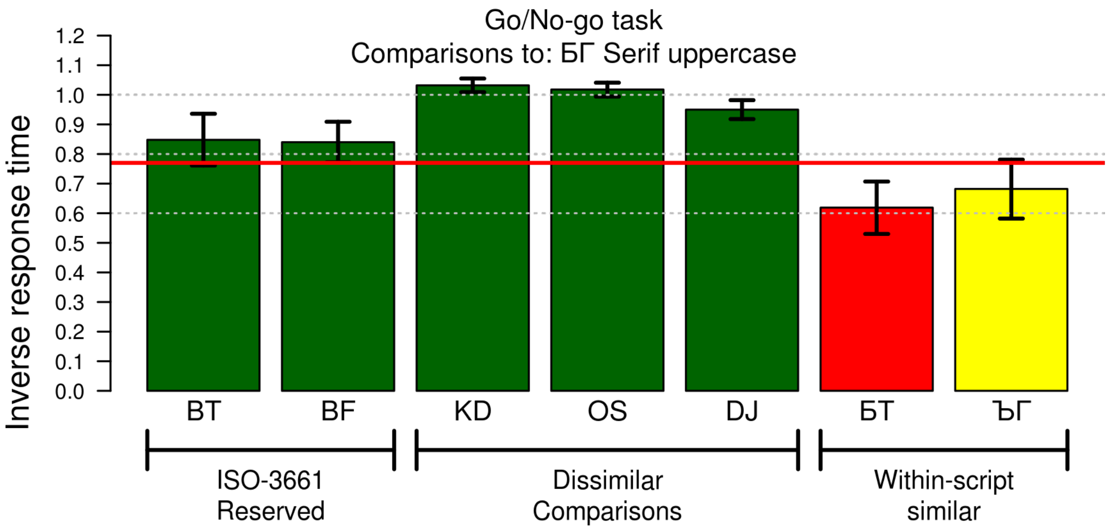

Go/No-go delta log(accuracy)
Comparisons to: БГ Sans uppercase


Go/No-go accuracy
Comparisons to: БГ Sans uppercase

-

	mean:	sd:	N:	se:	5%	95%
BT	0.306	0.256	27	0.049	0.204	0.407
BF	0.269	0.266	27	0.051	0.163	0.374
KD	0.056	0.111	27	0.021	0.011	0.1
OS	0.046	0.099	27	0.019	0.007	0.085
AK	0.065	0.127	27	0.024	0.015	0.115
БТ	0.454	0.284	27	0.055	0.341	0.566
ЪТ	0.292	0.199	27	0.038	0.213	0.37
ЪГ	0.31	0.258	27	0.05	0.208	0.412

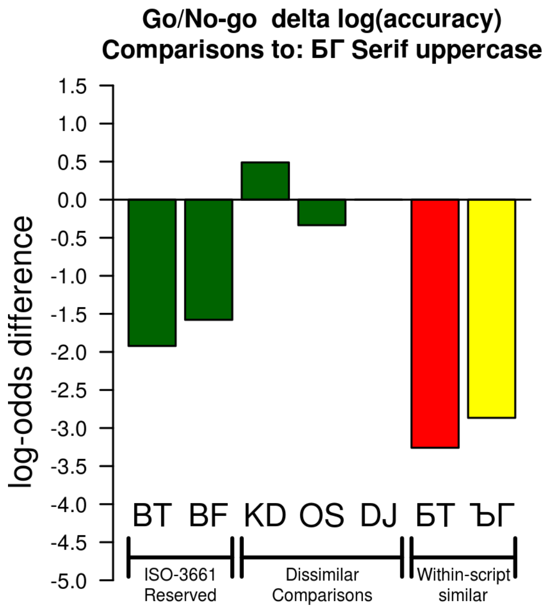
- **Inverse response time: Serif uppercase**
- Critical value: 0.77



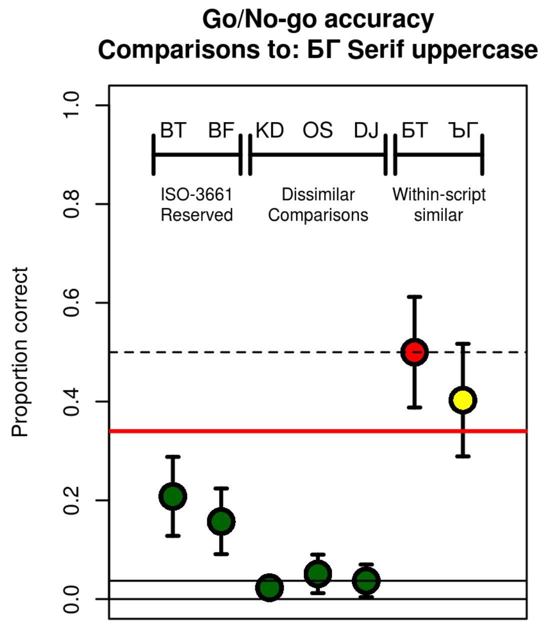
-

	mean	sd	N	se	5%	95%
BT	0.848	0.22	27	0.042	0.761	0.936
BF	0.84	0.173	27	0.033	0.772	0.909
KD	1.032	0.058	27	0.011	1.009	1.055
OS	1.018	0.059	27	0.011	0.994	1.041
DJ	0.95	0.082	27	0.016	0.918	0.982
БТ	0.619	0.224	27	0.043	0.53	0.707
ЪГ	0.682	0.251	27	0.048	0.582	0.781

- **Error rate: Serif uppercase**
- Critical value: 0.34



Go/No-go delta log(accuracy)
Comparisons to: БГ Serif uppercase



Go/No-go accuracy
Comparisons to: БГ Serif uppercase

	mean:	sd:	N:	se:	5%	95%
BT	0.208	0.202	27	0.039	0.128	0.288
BF	0.157	0.168	27	0.032	0.091	0.224
KD	0.023	0.049	27	0.01	0.004	0.043
OS	0.051	0.1	27	0.019	0.012	0.09
DJ	0.037	0.084	27	0.016	0.004	0.07
БТ	0.5	0.284	27	0.055	0.388	0.612
ЪГ	0.403	0.289	27	0.056	0.289	0.517

- **Summary of RT below threshold**

Pair:	Fontface	Mean:	Confidence interval	< 0.77
BT	*Sans Uppercase*	*0.704*	*0.771*	
BF	*Sans Uppercase*	*0.726*	*0.798*	

Italic indicates mean surpasses threshold. Bold indicates mean significantly surpasses threshold.

- **Summary of Error rate above threshold**

Pair:	Fontface	Mean:	Confidence interval	> 0.34
None				

Italic indicates mean surpasses threshold. Bold indicates mean significantly surpasses threshold.

Annex B - Final Report of the EPSRP for the application for BG in Bulgarian-Cyrillic

Final Report of the EPSRP for the application for BG in Bulgarian-Cyrillic

1. We are using two tasks: Delayed Matching to Sample (DMTS) and Go/NoGo (GNG).

2. From each task we want to derive two measures of similarity, making sure that one of these measures pays attention to response speed and the other pays attention to response accuracy. Jonathan suggested a simple solution: 1/RT (taking the inverse makes RT distributions much closer to normal; raw RT distributions typically have considerable positive skew) and percent correct. The advantages of these two measures is that they are simple to explain and that they do, taken together, capture both speed and accuracy. We agreed on 5 June that we would use 1/RT i.e. inv(RT) and percent correct as our two measures.

3. The proposed new DNs to evaluate (in several fonts, in both uppercase and lowercase) are бг/БГ in Cyrillic.

4. The data against which we will evaluate any proposed new DN combination are similarity measures from a set of DNs that are already being used or reserved for future use. Let's call these sets *reference sets.* A specific reference set was chosen for each candidate DN; these sets are listed in Appendix A. Our basic approach is this: if in an experiment involving the reference set plus the new proposed DN, the average similarity of the new DN to any member of its reference set is higher than the set of average similarities of the reference set to all the other members of the reverence set, that is a negative result for the new proposed DN. This is done in three steps:

Step (a): We measure the similarity of the candidate DN to all members of its reference set (Appendix A). This provides us with a mean and one-sided 95% confidence interval for every comparison of the DN with each member of the reference set.

Step (b): We measure the similarity of pairs of existing DNs (the anchor set - Appendix B) and use the highest observed similarity as the criterion against which the similarities measured in Step (a) will be evaluated. These criteria are selected to be levels consistent across several different studies.

Step (c): To be rejected, there must be evidence that the candidate is highly similar to potentially-confusing IDNs for both behavioral tasks. The DMTS task assesses memory confusion after brief delays, whereas the GNG task assesses the potential confusion of simultaneous glyphs, and so our proposal is that confusability should be demonstrated in both tasks.

For a given task, highly-similar could refer to one or to both measures (Inv RT and error rate) passing the established threshold criterion. If only one of these two measures passes threshold, we treat this as sufficient evidence for rejection provided that the result cannot be due to a speed-accuracy tradeoff. We recommend that this pattern does not need to hold for any single fontface/IDN combination, but for at least one IDN/fontface in each task.

5. To compare the similarity of the new proposed DN to the set of similarities of the reference set we calculated the average similarity value for each subject across all the items in the reference set and construct a one-sided 95% confidence interval from that set of subject means. This produced a critical value for each of our four measures i.e. a value at the end of the one-sided 95% confidence interval. The resulting cutoff critical values were:

DMTS inv(RT): <0.9

DMTS error rate: >0.14

GNG inv(RT): <.77

GNG error rate: >.34

If the similarity of any new proposed DN to the members of the reference set is outside this 95% confidence interval for both tasks, that is a negative result for the new proposed DN.

The procedures by which we arrived at these values is summarized in Appendix B and described in detail in the documents dmts-anchors.pdf and gonogo-anchors.pdf.

6. Results

DMTS

Summary of invRT below threshold (if both are below 0.9 then the result is a fail - bold)

Pair:	Fontface	Mean	Confidence interval
BT	***Sans Uppercase***	***0.87***	***0.897***
BF	***Sans Uppercase***	***0.84***	***0.879***
BT	***Serif Uppercase***	***0.855***	***0.887***
BF	*Serif Uppercase*	*0.898*	*0.929*

Italic indicates mean exceeds threshold. Bold indicates mean significantly exceeds threshold.

Summary of Error rate above threshold (if both are greater than 0.14 then the result is a fail - bold)

Pair:	Fontface	Mean	Confidence interval
None			

Italic indicates mean exceeds threshold. Bold indicates mean significantly exceeds threshold.

Same/different go/no-go task

Summary of invRT below threshold (if both are below 0.77 then the result is a fail - bold)

Pair:	**Fontface**	**Mean:**	**Confidence interval**
BT	Sans Uppercase	0.704	0.771
BF	Sans Uppercase	0.726	0.798

Summary of Error rate above threshold (if both are above 0.34 then the result is a fail - bold)

Pair:	**Fontface**	**Mean:**	**Confidence interval**
None			

Italic indicates mean exceeds threshold. Bold indicates mean significantly exceeds threshold.

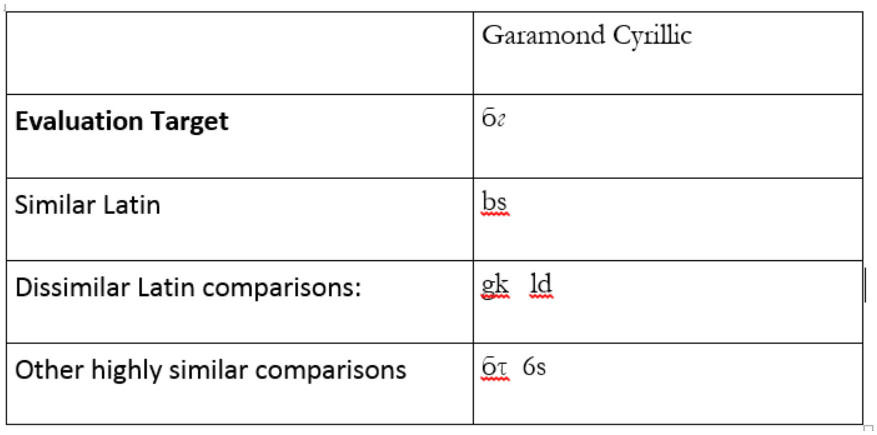
7. Conclusion

No testing pair failed both tasks in either upper or lower case. The candidate string is not confusingly similar to any ISO 3166-1 entries.

APPENDIX A: Reference sets and testing plans for each candidate DN.

Stimuli for Candidate: бг/ БГ in Cyrillic

	Serif lowercase **Times New Roman**	Sans serif lowercase **Segoe UI**
Evaluation target	бг	бг
Similar Latin	br bt	br bt
Dissimilar Latin comparisons:	nk ja ld	nk ja ld
Other highly similar comparisons	бτ 6r 6г	бτ 6r 6г



	Garamond Cyrillic
Evaluation Target	бг
Similar Latin	bs
Dissimilar Latin comparisons:	gk ld
Other highly similar comparisons	бτ 6s

	Serif uppercase Times new roman	Sans serif uppercase Segoe UI Uppercase
Evaluation Target	БГ	БГ
Similar Latin	BT BF	BT BF
Dissimilar Latin comparisons:	KD OS AK	KD OS AK
Other Highly similar comparisons	БТ ЪТ ЪГ	БТ ЪТ ЪГ

APPENDIX B:

General procedures for using the anchor sets to establish the critical values for the DMTS and GNG 1/RT and error measures. For full details of these procedures please consult the research results.

Candidate: Latin Comparison anchor sets

The purpose of these is to establish a set of high-similarity pairs that have an acceptable level of confusability/similarity. Nine pairs were selected from the highly-confusable pairings of the following letter sets, and measures compared to those same candidates with respect to dissimilar letter combinations. Each study and task contained two blocks of these trials. A single set of criteria was chosen based on all three studies.

Stimuli:

- it and lt
- fi and fj
- ai, al, at
- cx and ex

Presentation

- Sans serif stimuli were displayed as rendered in the location bar of a popular internet browser running on Microsoft Windows. Serif and italic stimuli were obtained via screenshots from a word processing application using Times New Roman font face to match the size of the sans serif font (Approximately 10-11pt size, non-italic, non-bold with normal spacing).
- Participants were instructed to view the screen from a comfortable distance, to best match their naturalistic screen viewing conditions.

Procedures

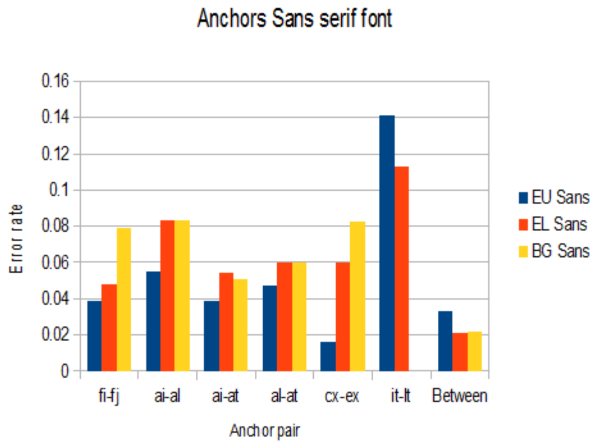
Testing used two procedures: 1. A delayed match-to-sample forced-choice identification task, and 2. A go/no-go response same-different judgment task. The advantage of method 1 is that it tends to produce differences in response time based on confusability that are highly reliable with minimal observations, the advantage of method 2 is that it induces larger differences is accuracy, and requires a participant to detect a specific difference.

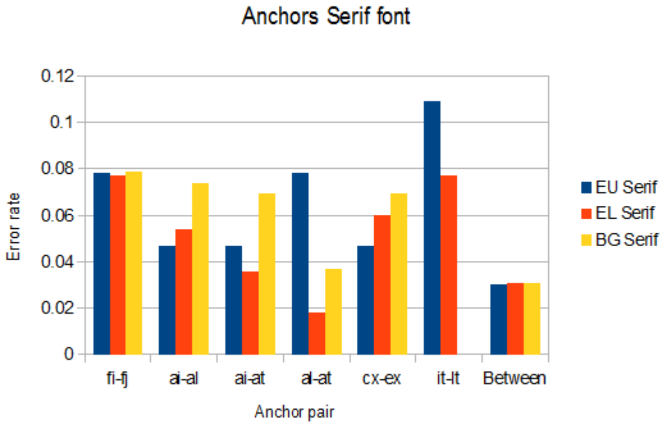
Each test was performed in a blocked design in the same order across participants. Each set of stimuli will appear in a contiguous block. Testing was designed to assess the similarity between the target and (1) any of a set of highly-similar Latin character pairs in the same case (2) a set of 3-4 dissimilar Latin character pairs, and (3) any highly-similar comparisons, which may not directly bear on the decision, but may help to calibrate and validate the measures.

Participants

In this study, we intend to test 20 undergraduate students, primarily students of U.S. origin. Because they are experts in Latin orthography, which is the orthography where the confusions are most likely to occur, they serve as a reasonable population for evaluating these characters sets to make inference about a general internet population

DMTS Anchor Summary





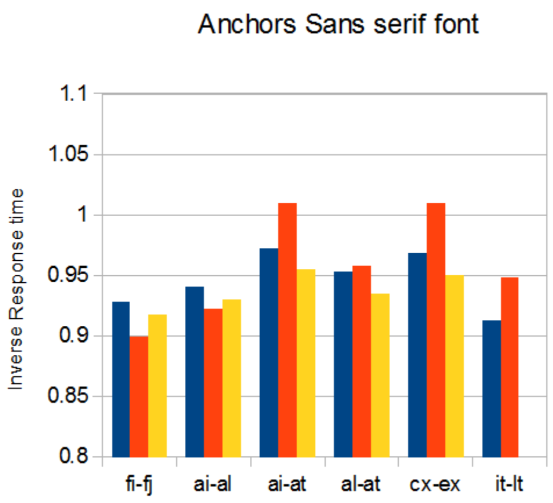
Error Rate

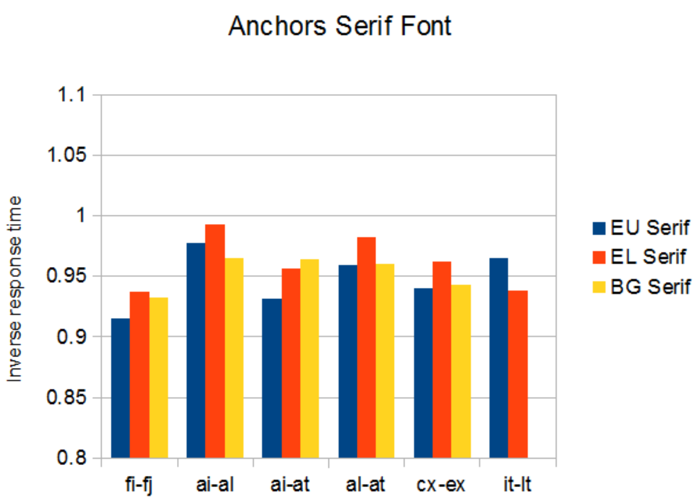
Option	EU Sans	EL Sans	BG Sans
fi-fj	0.039	0.048	0.0787
ai-al	0.055	0.083	0.0833
ai-at	0.039	0.054	0.0509
al-at	0.047	0.06	0.0602
cx-ex	0.016	0.06	0.0827
it-lt	0.141	0.113	
Between	**0.033**	**0.021**	**0.0217**

Option	EU Serif	EL Serif	BG Serif
fi-fj	0.078	0.077	0.0787
ai-al	0.047	0.054	0.0741
ai-at	0.047	0.036	0.0694
al-at	0.078	0.018	0.037
cx-ex	0.047	0.06	0.0694
it-lt	0.109	0.077	
Between	**0.03**	**0.031**	**0.0306**

- In the tables and figures, EU/EL/BG indicate the study in which the data were collected, the stimuli were not visually different and design differed minimally.
- it-lt has the highest error rate (average .127; max .14). Overall dissimilar error rate is 2-3%, but this tends to be a bit higher for it-lt. This is 3-4 times the baseline error rate.
- Test-retest reliability for Sans is .90 ; serif is .36
- Adjusting accuracy (by subtracting or dividing by baseline) reduces test-retest reliability.
- **Recommendation: use .14 as criterion.**

Inverse Response Time

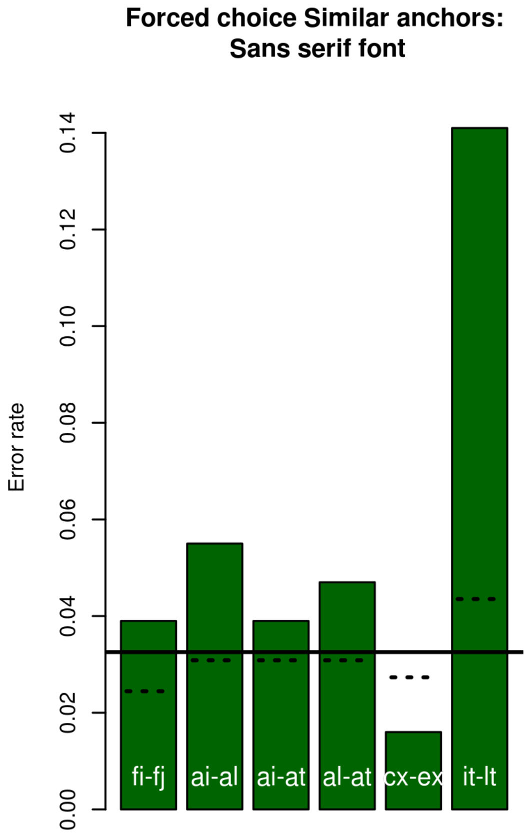




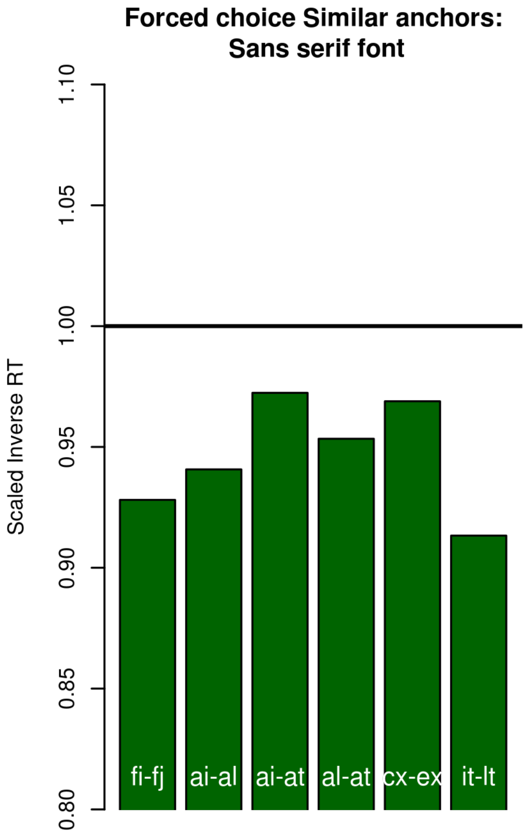
Option	EU Sans	EL Sans	BG Sans
fi-fj	0.9281	0.8995	0.918
ai-al	0.9407	0.9225	0.93
ai-at	0.9724	1.0096	0.955
al-at	0.9534	0.9584	0.935
cx-ex	0.9689	1.01	0.95
it-lt	0.9133	0.9483	-

Option	EU Serif	EL Serif	BG Serif
fi-fj	0.9155	0.9371	0.932
ai-al	0.9773	0.9925	0.965
ai-at	0.9316	0.9561	0.964
al-at	0.9596	0.9826	0.96
cx-ex	0.9401	0.962	0.943
it-lt	0.9648	0.9382	-

- Overall lowest Inverse RT (worst performance) is fi-fj Sans, averaging .915, with lowest of .8995.
- For sans, test-retest reliability was {.78, .98,.99}; for serif, {.63,.76,.72}.
- **Recommendation: Use 0.9 as criterion.**

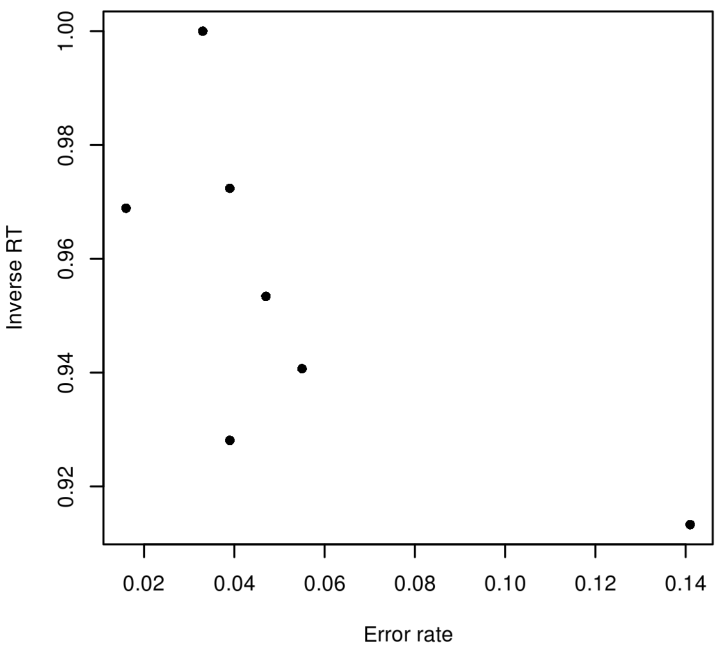


Candidate: EU in Greek. (epsilon upsilon)

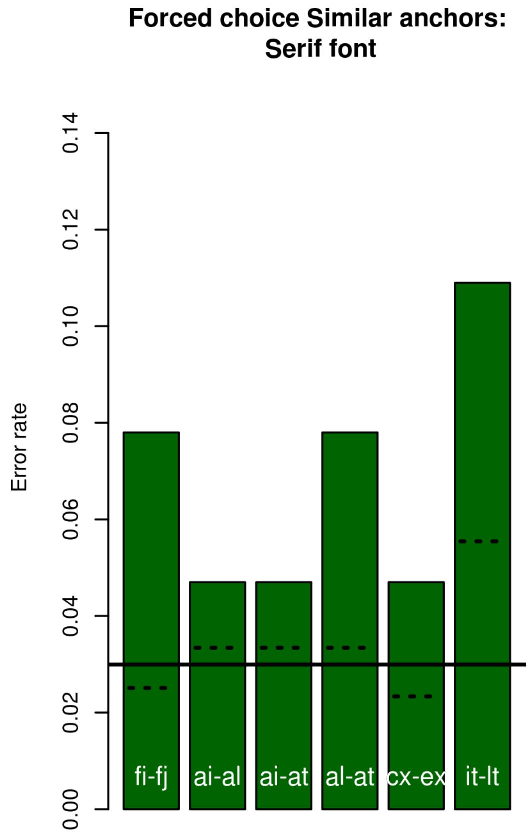


Candidate: EU in Greek. (epsilon upsilon)

Option	**Error rate**	**Between error rate**	**Inverse RT**	**Log-odds delta accuracy**
fi-fj	0.039	0.024	0.9281	-0.484
ai-al	0.055	0.031	0.9407	-0.597
ai-at	0.039	0.031	0.9724	-0.244
al-at	0.047	0.031	0.9534	-0.597
cx-ex	0.016	0.027	0.9689	0.571
it-lt	0.141	0.044	0.9133	-1.28
Between	0.033	0.033	1	0

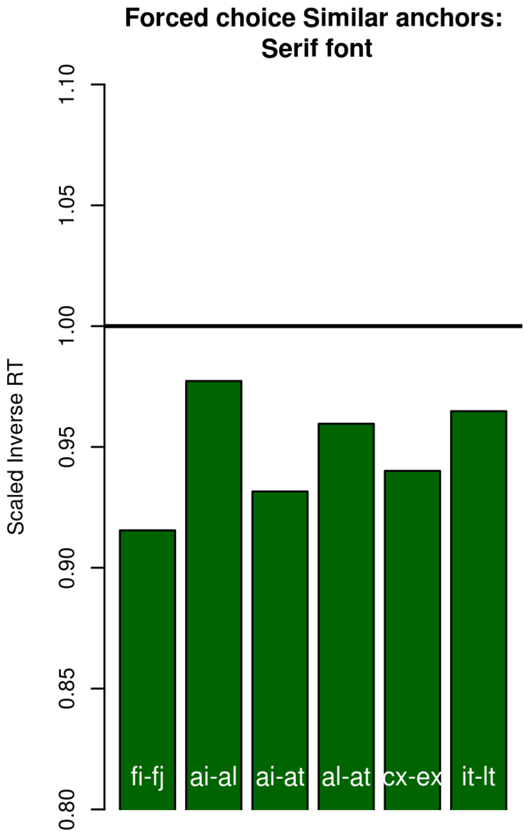


Correlation between error rate and inverse RT: -0.6925



Forced choice Similar anchors: Serif font

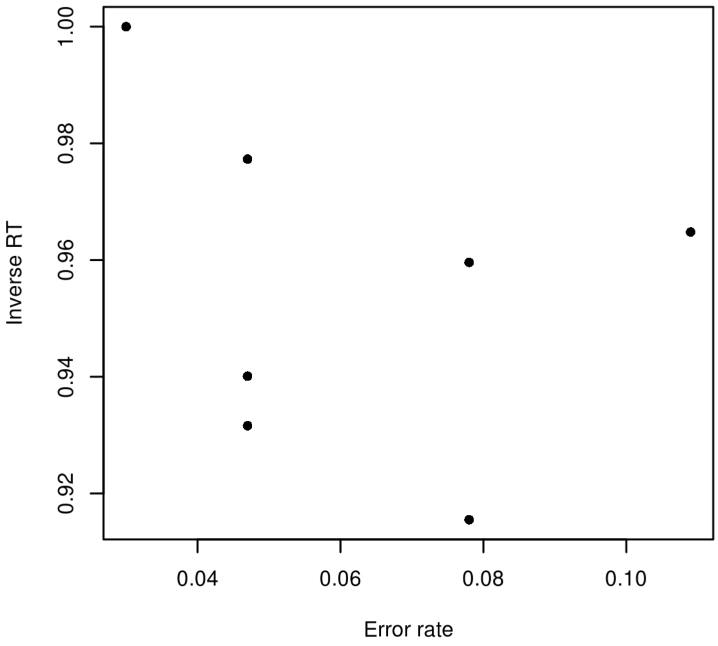
Candidate: EU in Greek. (epsilon upsilon)



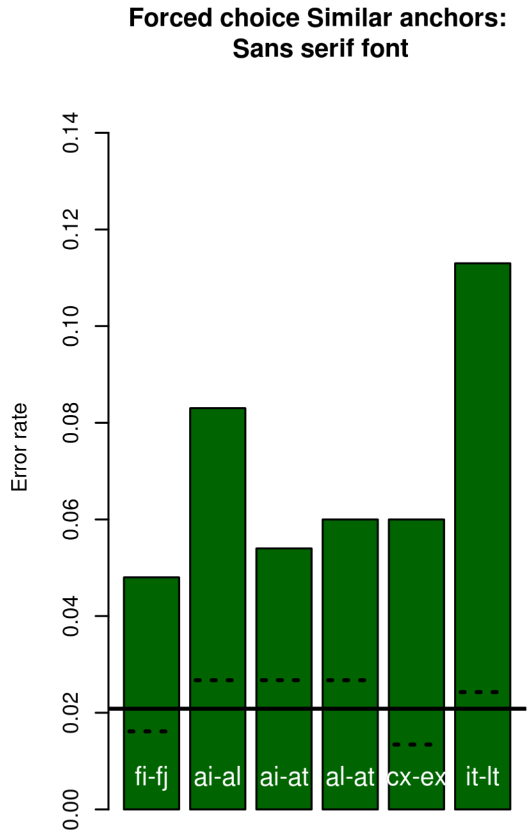
Forced choice Similar anchors: Serif font

Candidate: EU in Greek. (epsilon upsilon)

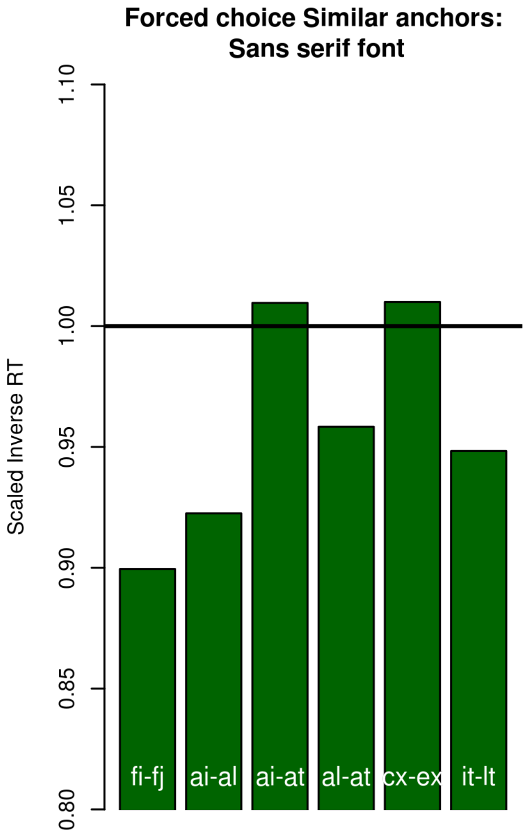
Option	**Error rate**	**Between error rate**	**Inverse RT**	**Log-odds delta accuracy**
fi-fj	0.078	0.025	0.9155	-1.192
ai-al	0.047	0.033	0.9773	-0.352
ai-at	0.047	0.033	0.9316	-0.352
al-at	0.078	0.033	0.9596	-0.352
cx-ex	0.047	0.023	0.9401	-0.721
it-lt	0.109	0.055	0.9648	-0.738
Between	0.03	0.03	1	0



Correlation between error rate and inverse RT: -0.2772

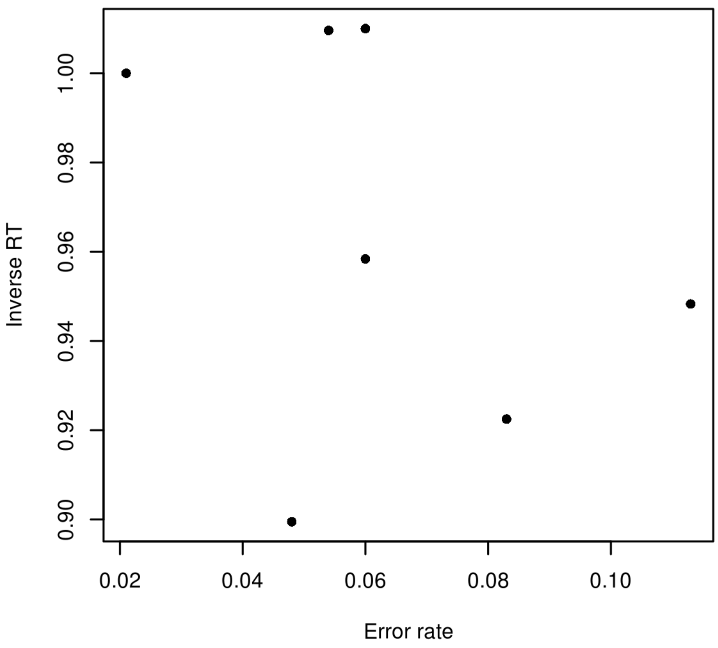


Candidate: EL in Greek. (epsilon lambda)

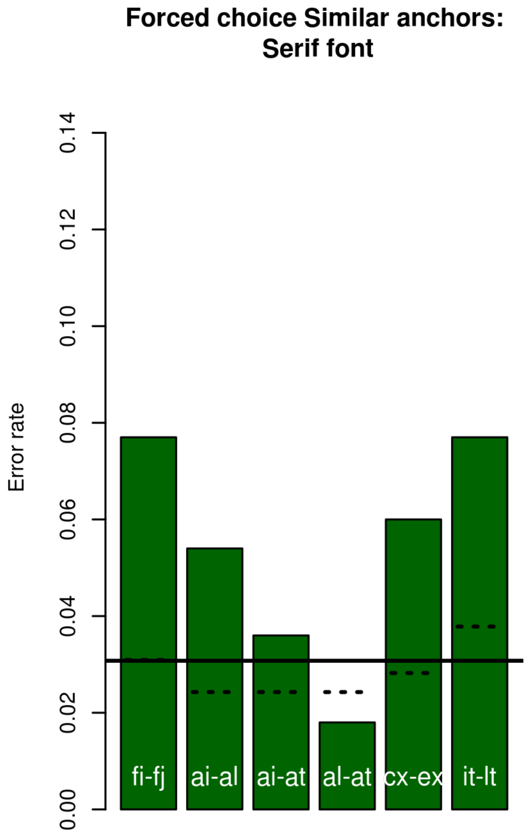


Candidate: EL in Greek. (epsilon lambda)

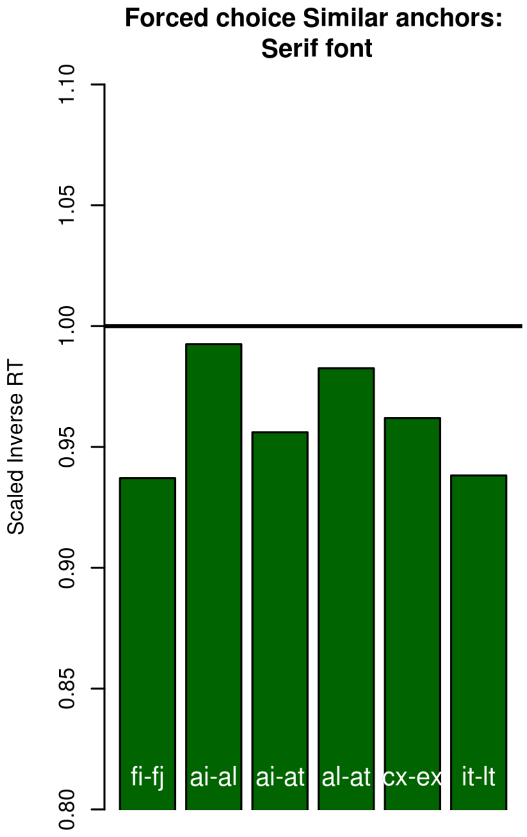
Option	Error rate	Between error rate	Inverse RT	Log-odds delta accuracy
fi-fj	0.048	0.016	0.8995	-1.114
ai-al	0.083	0.027	0.9225	-1.197
ai-at	0.054	0.027	1.0096	-0.723
al-at	0.06	0.027	0.9584	-1.197
cx-ex	0.06	0.013	1.01	-1.537
it-lt	0.113	0.024	0.9483	-1.635
Between	0.021	0.021	1	0



Correlation between error rate and inverse RT: -0.353

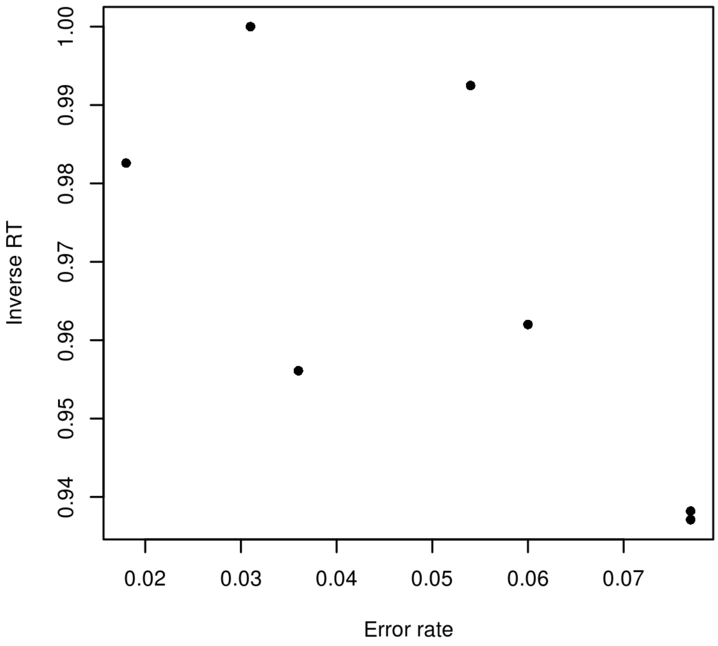


Candidate: EL in Greek. (epsilon lambda)



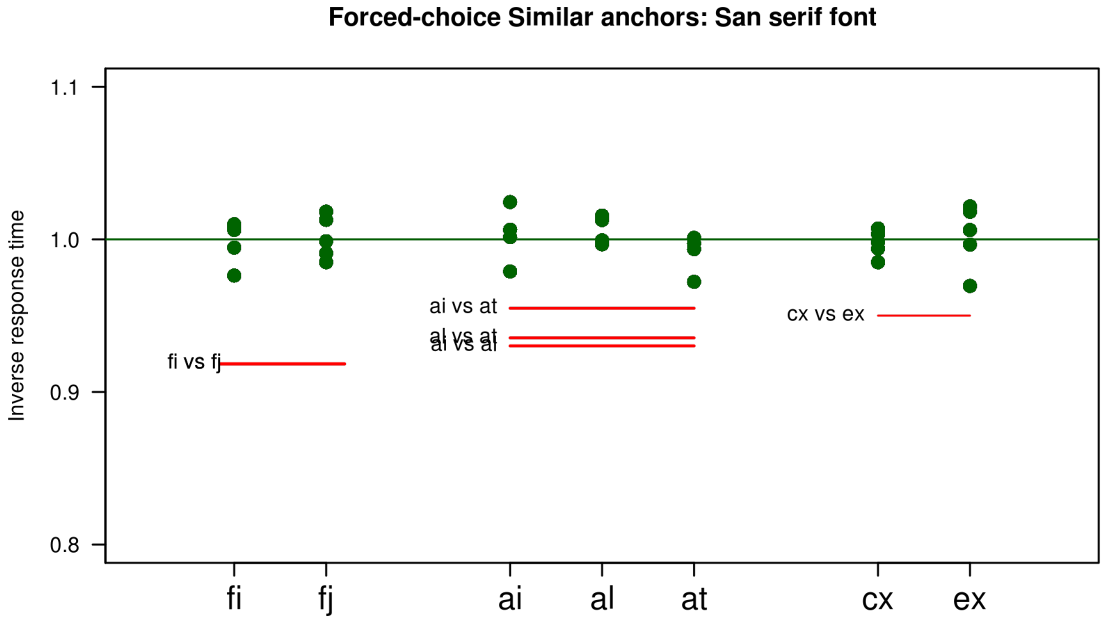
Candidate: EL in Greek. (epsilon lambda)

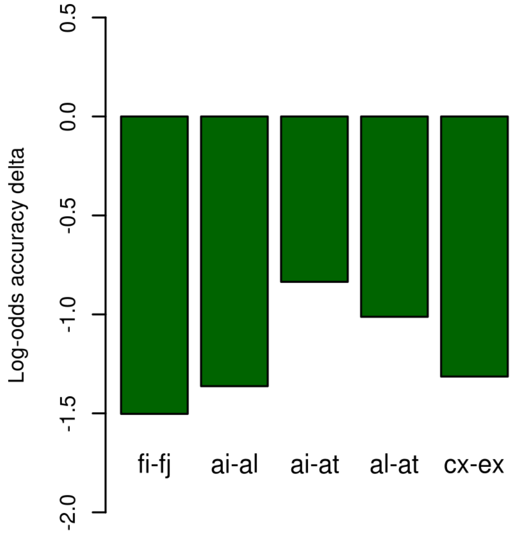
Option	Error rate	Between error rate	Inverse RT	Log-odds delta accuracy
fi-fj	0.077	0.031	0.9371	-0.966
ai-al	0.054	0.024	0.9925	-0.822
ai-at	0.036	0.024	0.9561	-0.398
al-at	0.018	0.024	0.9826	-0.822
cx-ex	0.06	0.028	0.962	-0.779
it-lt	0.077	0.038	0.9382	-0.757
Between	0.031	0.031	1	0

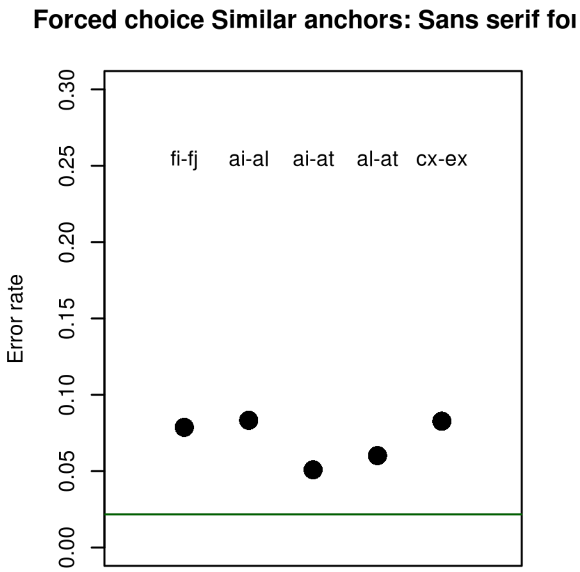


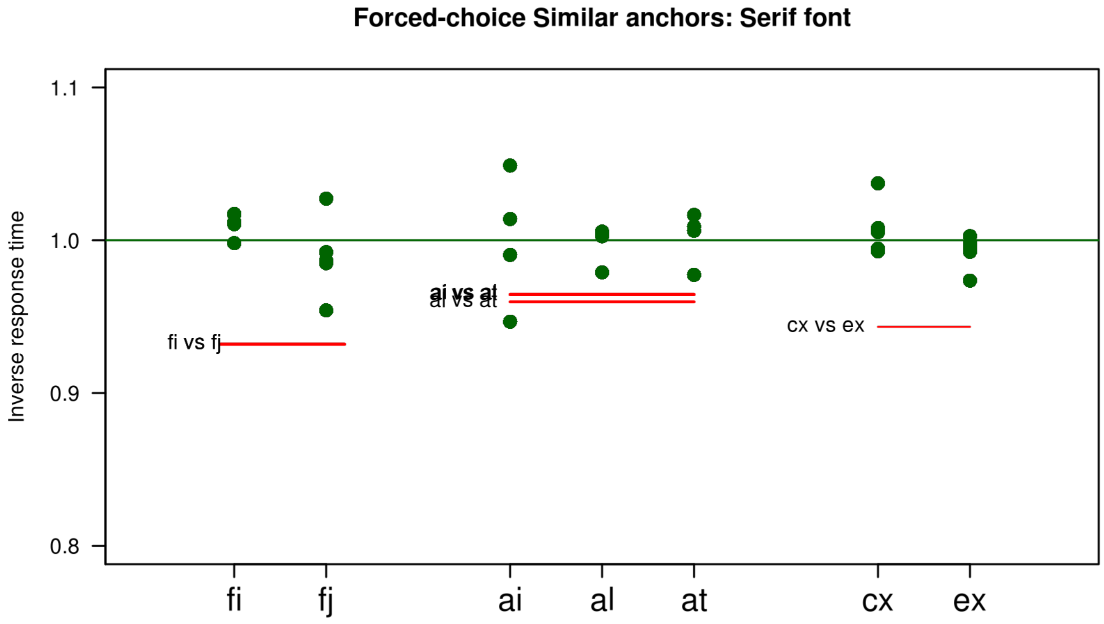
Correlation between error rate and inverse RT: -0.7193

The next figure shows comparisons of similar latin pairs. These serve as a comparison set, with the logic that any new pair evaluated to be less similar than these anchors is justifiably allowable.

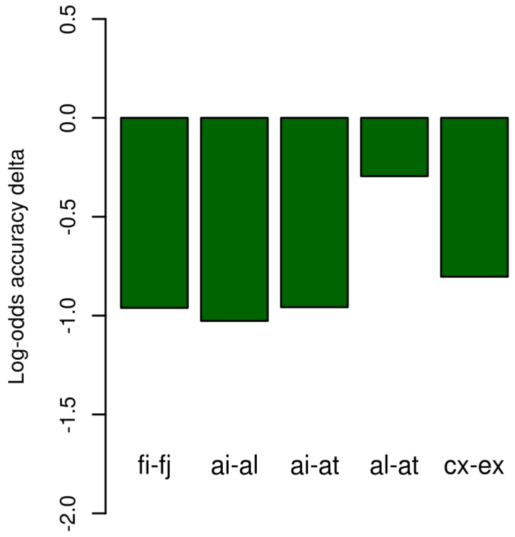


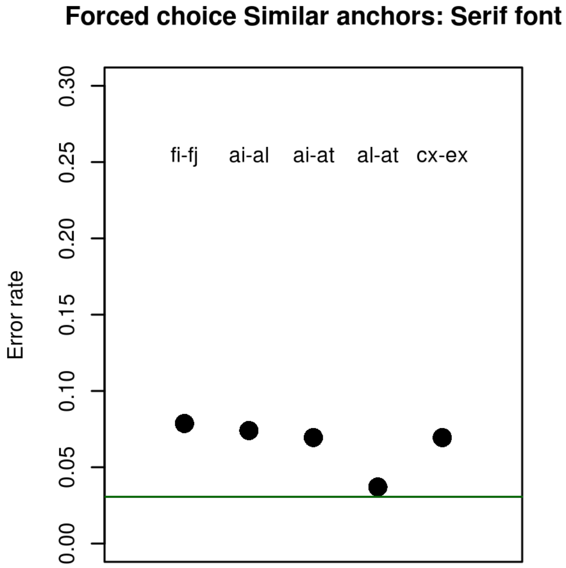






Forced-choice Similar anchors: Serif font





Forced choice Similar anchors: Serif font

Inverse response time

	fi-fj	ai-al	ai-at	al-at	cx-ex
Sans serif	0.918	0.93	0.955	0.935	0.95
Serif	0.932	0.965	0.964	0.96	0.943

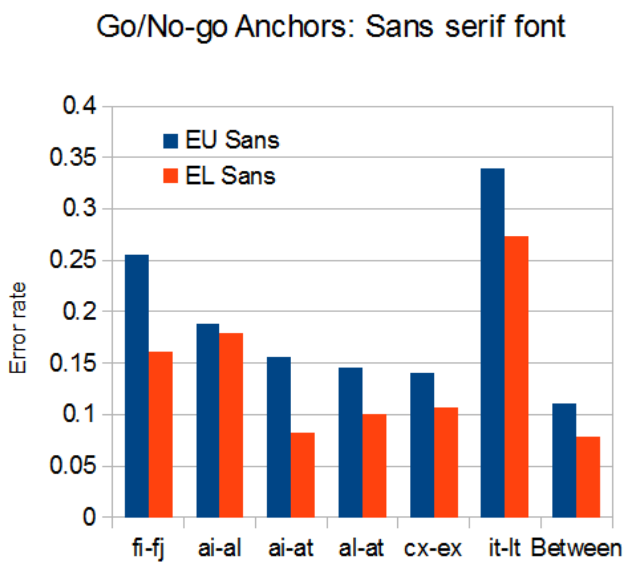
Log-odds difference in accuracy

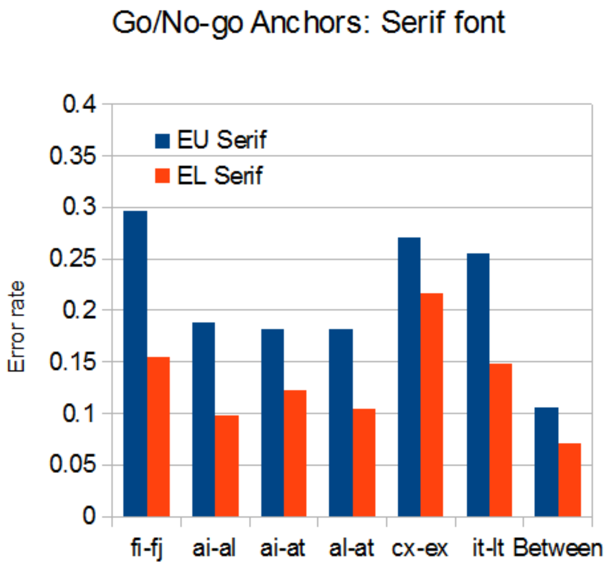
	fi-fj	ai-al	ai-at	al-at	cx-ex
Sans serif	-1.5025	-1.3627	-0.8355	-1.0124	-1.3141
Serif	-0.961	-1.027	-0.9575	-0.2946	-0.8034

Error rate

	Between	fi-fj	ai-al	ai-at	al-at	cx-ex
Sans serif	0.0217	0.0787	0.0833	0.0509	0.0602	0.0827
Serif	0.0306	0.0787	0.0741	0.0694	0.037	0.0694

Go/No-Go Task: Accuracy Metric


Go/No-go Anchors: Sans serif font


Go/No-go Anchors: Serif font

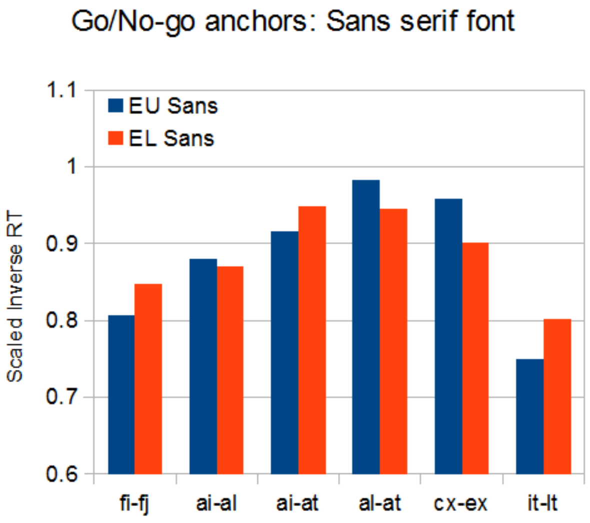
Option	EU Sans	EL Sans
fi-fj	0.255	0.161
ai-al	0.188	0.179
ai-at	0.156	0.083
al-at	0.146	0.101
cx-ex	0.141	0.107
it-lt	0.339	0.274
Between	**0.111**	**0.079**

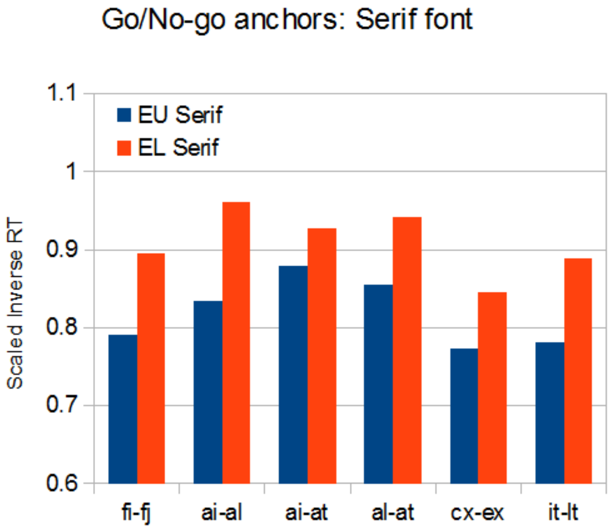
Option	EU Serif	EL Serif
fi-fj	0.297	0.155
ai-al	0.188	0.098
ai-at	0.182	0.122
al-at	0.182	0.104
cx-ex	0.271	0.217
it-lt	0.255	0.149
Between	**0.106**	**0.071**

- Test-retest reliability is .922 for Sans and .77 for serif.
- EL study produced overall lower error rates; possibly because these anchors were tested at the end of the study and
- Adjusting accuracy by subtracting error rate obtained for each pair changes these to (.91, .91), and by dividing to (.88, .98).
- Adjusting by dividing seems to make highest values most consistent across experiments, but this adjustment cannot be done reliably on an individual basis (because of error rates of 0, relatively small numbers of observations for the comparison cases, and wide binomial error variability)
- Correlations of adjusted to non-adjusted accuracy scores are all above .95, but it seems likely that the increase in reliability is mostly accidental and might not be replicated in future studies (and was did not occur for DMTS task).
- Worst-case is .339 for it-lt; Average of it-lt sans is .306, consistent with fi-fj serif of .297.
- **Recommendation: use error rate of 0.34 as a conservative criterion**

Note: Error rate and Inverse RT were correlated {-.937, -.979, -.965, -.89}, suggesting that the overall decision should agree highly between these two measures and both may not be necessary.

Go/No-Go Task: Inverse RT Metric

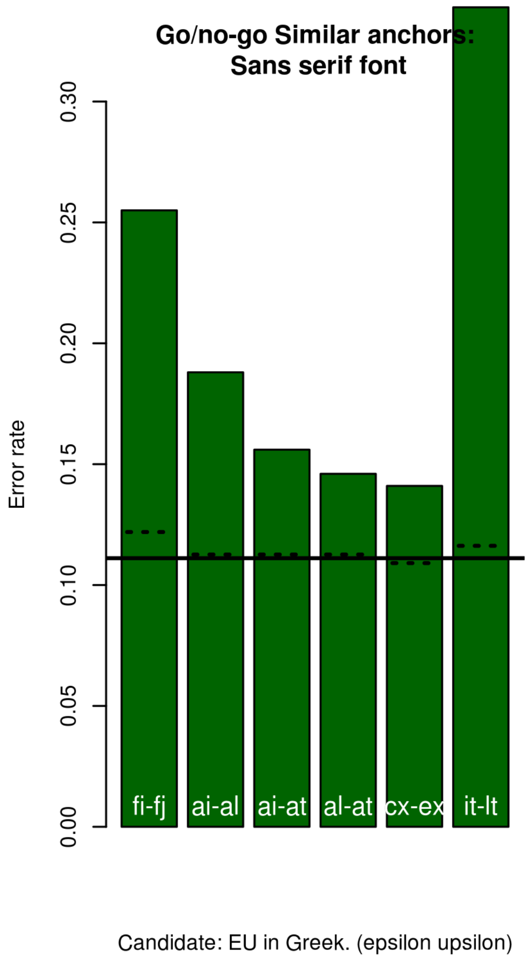


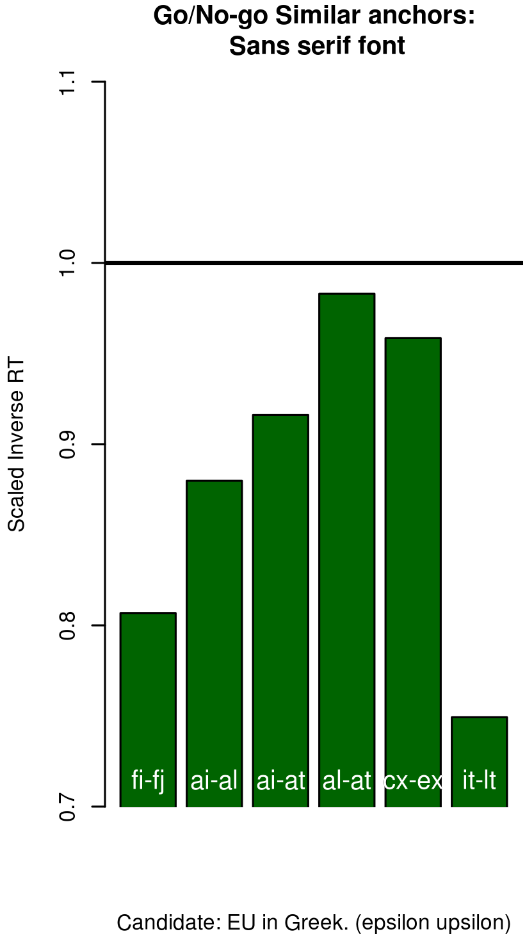


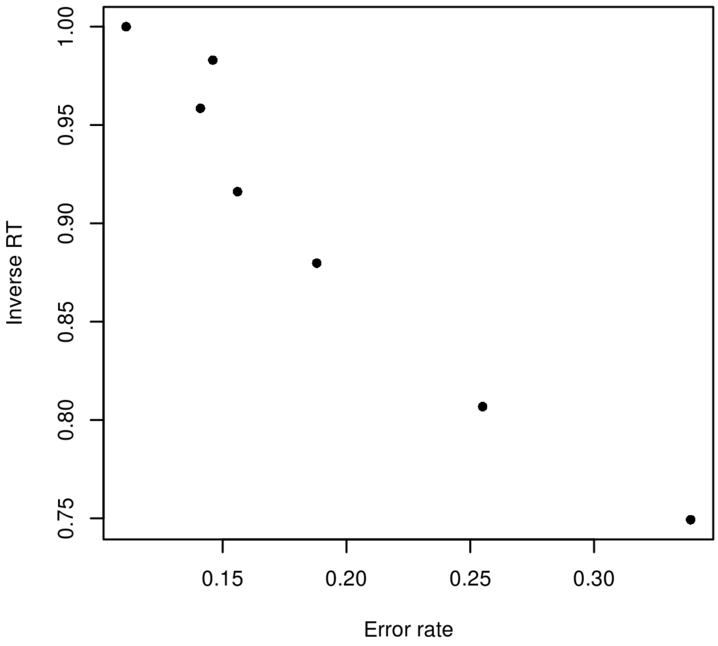
Option	EU Sans	EL Sans
fi-fj	0.8068	0.8472
ai-al	0.8798	0.8704
ai-at	0.9161	0.9486
al-at	0.983	0.9455
cx-ex	0.9585	0.9014
it-lt	0.7493	0.802

Option	EU Serif	EL Serif
fi-fj	0.7907	0.8953
ai-al	0.8344	0.9606
ai-at	0.8796	0.9281
al-at	0.8552	0.9414
cx-ex	0.7723	0.8454
it-lt	0.781	0.8886

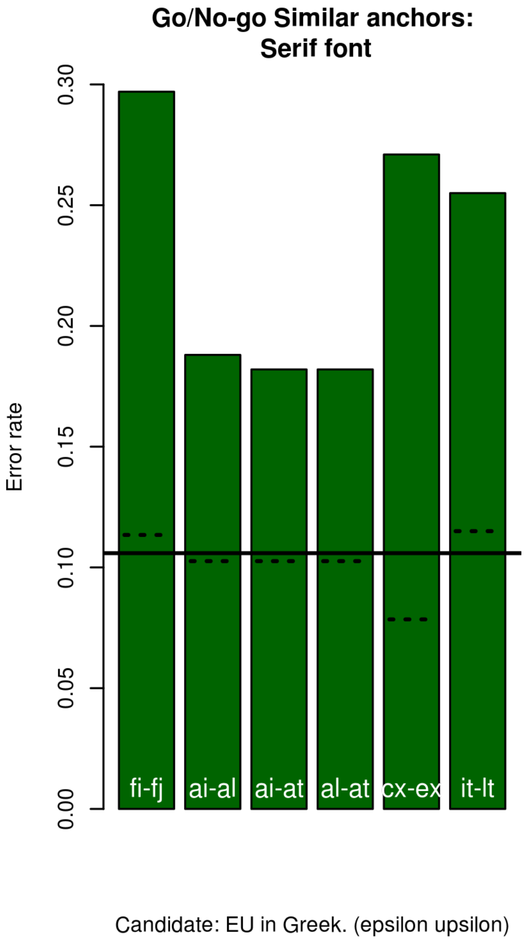
- Test-retest reliability was .906 for sans and .79 for serif. These values are already scaled, so that 1.0 is the average 'different' value.
- EL study produced higher values in the serif font. This is consistent with the overall higher accuracy, and is not a speed-accuracy tradeoff.**.**
- Several cases in each font and each experiment produce scaled RT below 0.8; lowest is 0.77.
- **Recommendation: use 0.77 as criterion.**

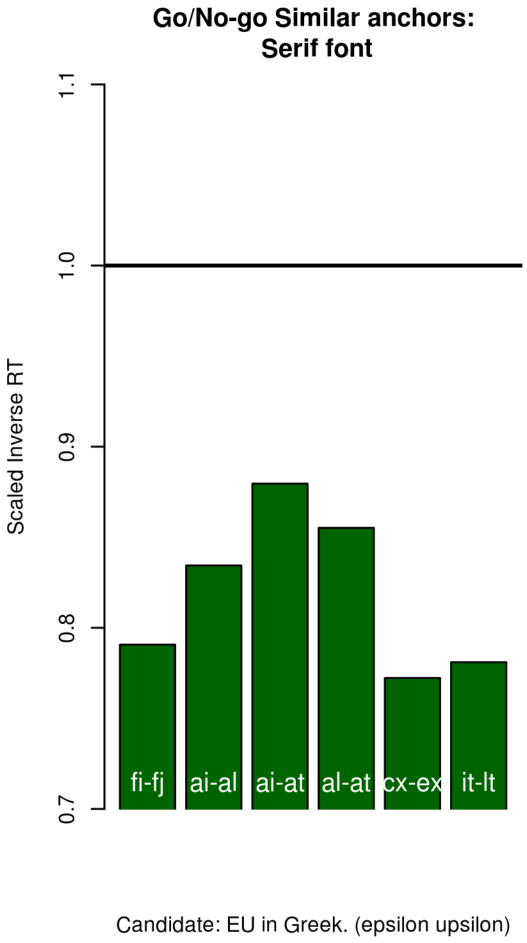

Go/no-go Similar anchors:
Sans serif font
Error rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
fi-fj
ai-al
ai-at
al-at
cx-ex
it-lt
Candidate: EU in Greek. (epsilon upsilon)

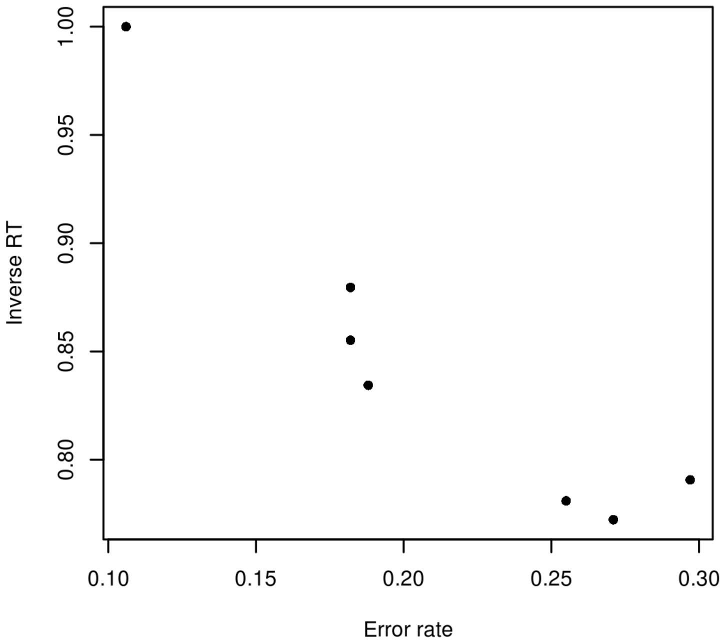

Go/No-go Similar anchors:
Sans serif font
Scaled Inverse RT
0.7
0.8
0.9
1.0
1.1
fi-fj
ai-al
ai-at
al-at
cx-ex
it-lt
Candidate: EU in Greek. (epsilon upsilon)



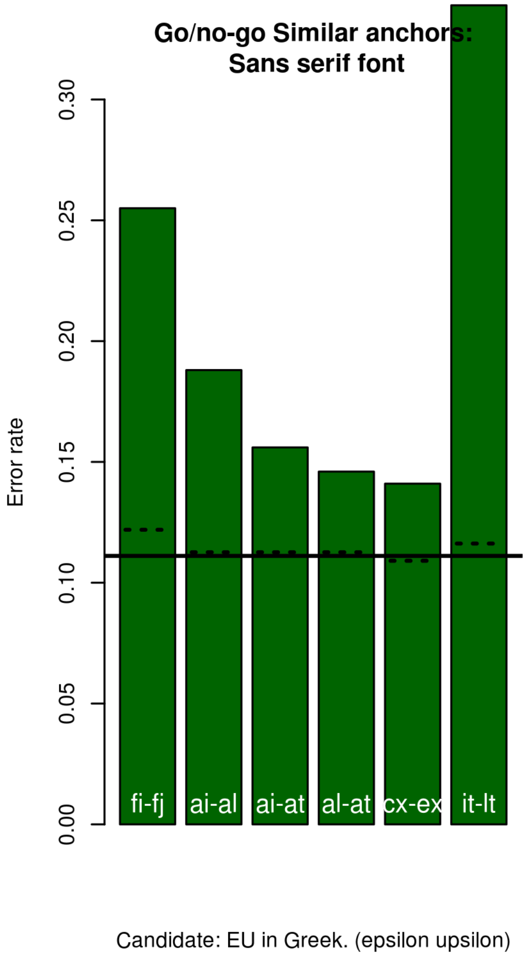
Correlation between error rate and inverse RT: -0.9716


Go/No-go Similar anchors:
Serif font
Error rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
fi-fj
ai-al
ai-at
al-at
cx-ex
it-lt
Candidate: EU in Greek. (epsilon upsilon)

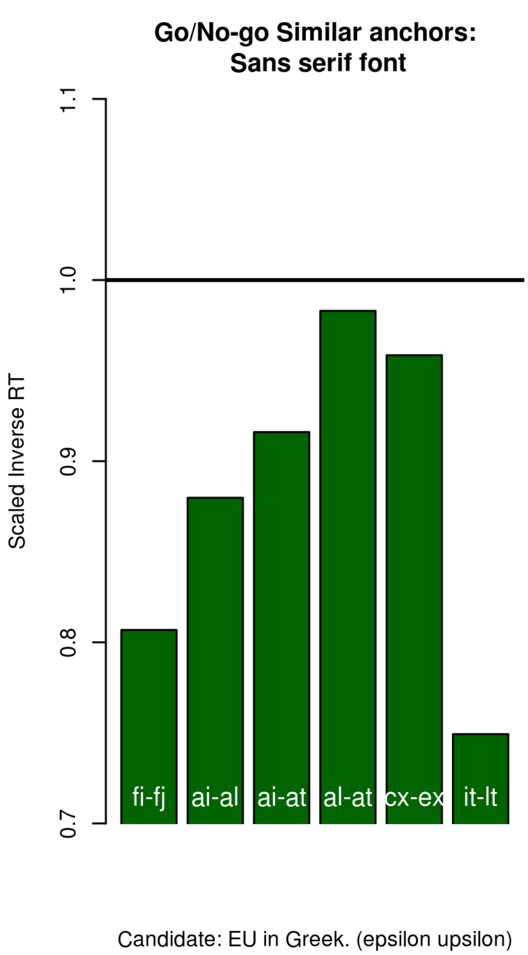

Go/No-go Similar anchors:
Serif font
Scaled Inverse RT
0.7
0.8
0.9
1.0
1.1
fi-fj
ai-al
ai-at
al-at
cx-ex
it-lt
Candidate: EU in Greek. (epsilon upsilon)



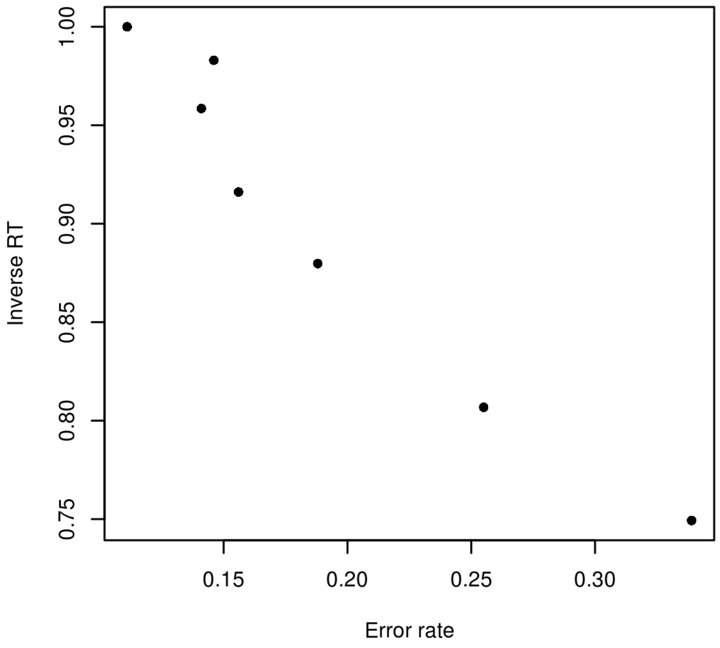
Correlation between error rate and inverse RT: -0.9281



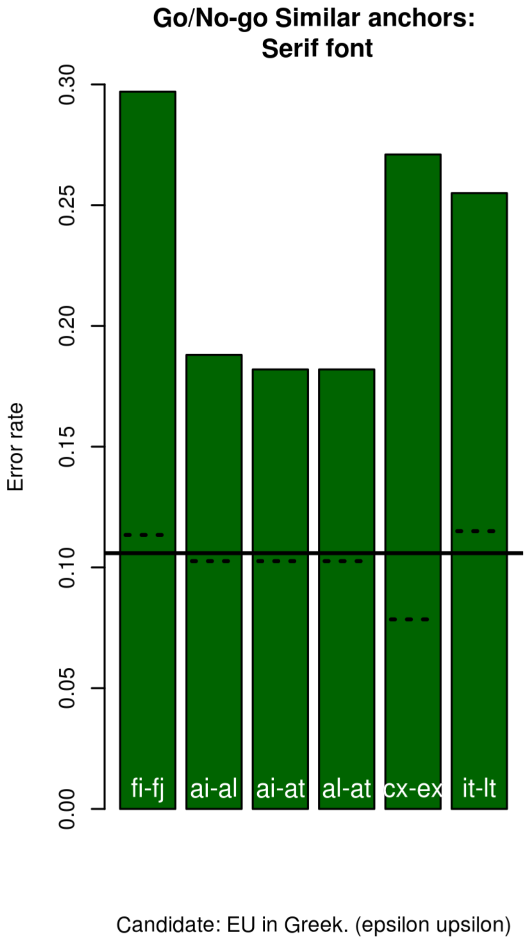
Go/no-go Similar anchors:
Sans serif font

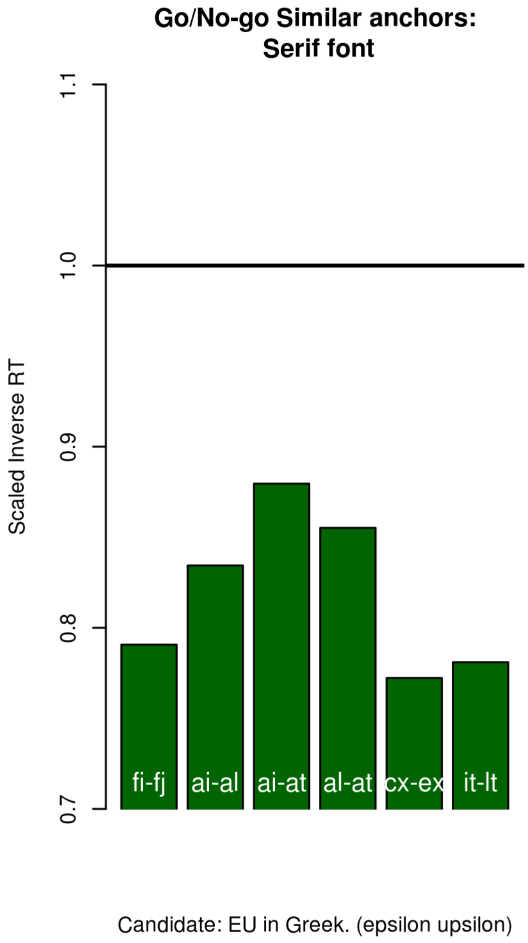


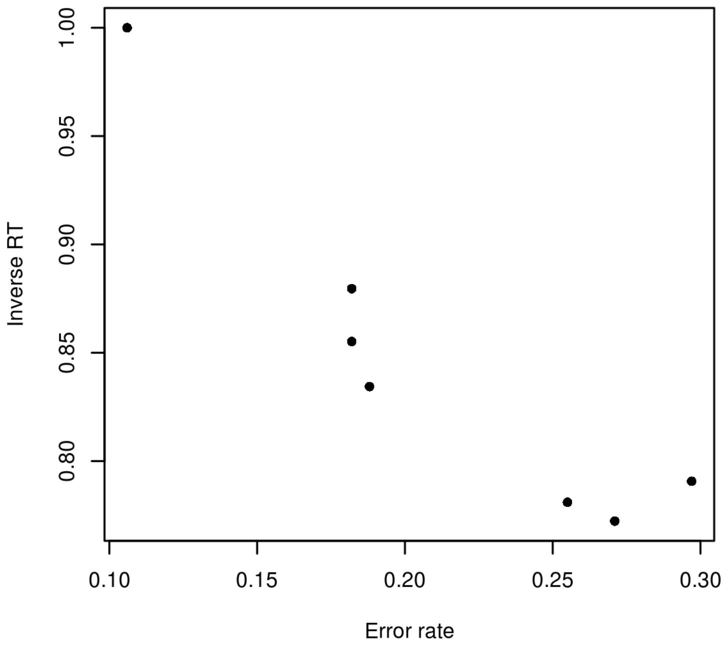
Go/No-go Similar anchors:
Sans serif font



Correlation between error rate and inverse RT: -0.9716


Go/No-go Similar anchors:
Serif font
Error rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
fi-fj
ai-al
ai-at
al-at
cx-ex
it-lt
Candidate: EU in Greek. (epsilon upsilon)


Go/No-go Similar anchors:
Serif font
Scaled Inverse RT
0.7
0.8
0.9
1.0
1.1
fi-fj
ai-al
ai-at
al-at
cx-ex
it-lt
Candidate: EU in Greek. (epsilon upsilon)



Correlation between error rate and inverse RT: -0.9281