Extended Process Similarity Review Panel

(EPSRP)

Report for

Country: Greece

A-Label: xn--qxam

U-Label: ελ

Unicode Code Points: U+03B5, U+03BB

String in English: el

String Language: Greek, Modern (1453)

Script: Greek

September 2014

Contents

Executive Summary

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

Corresponding ISO3166 Entry: GR
A-Label: xn--qxam
U-Label: ελ
Unicode Code Points: U+03B5, U+03BB
String in English: el
String Language: Greek, Modern (1453)
Script: Greek

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: GR
A-Label: xn--qxam
U-Label: ελ
Unicode Code Points: U+03B5, U+03BB
String in English: el
String Language: Greek, Modern (1453)
Script: Greek

5 Documents provided to the panel by ICANN

Submitted to the panel by ICANN:

- EPSRP Application form
- Letter_of_Support_Greece.pdf
- ALLOWED_GREEK_CHARACTERS_FOR_IDN_REGISTRATIONS_UNDER_.gr.pdf

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

- Email by Panagiotis Papaspiliopoulos on 5 April 2014 providing additional explanations for the application.

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: ελ/ ΕΛ (.el/.EL in Greek)

	Serif lowercase **Times New Roman**	**Sans serif lowercase** **Segoe UI**
Evaluation target	**ελ**	**ελ**
Similar Latin	ey, sy, ex, ev	ey, sy, ex, ev
Dissimilar Latin comparisons:	ab,gn,zq,fr	ab,gn,zq,fr
Other Highly similar comparisons	none evaluated	none evaluated

Evaluation Target	**Serif uppercase** **Times new roman**	**Sans serif uppercase** **Segoe UI Uppercase**
	ΕΛ	ΕΛ
Similar Latin	EV, FV,EA,FA	EV,FV,EA,FA
Dissimilar Latin comparisons:	SG,UB,CR,QJ	SG UB,CR,QJ
Other Highly similar comparisons	None evaluated	None evaluated

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**
EA	*Sans Uppercase*	*0.829*	*0.914*
FA	*Sans Uppercase*	*0.899*	*0.956*
EV	*Serif Uppercase*	*0.855*	*0.909*
FV	*Serif Uppercase*	*0.891*	*0.943*
EA	*Serif Uppercase*	*0.844*	*0.934*
FA	*Serif Uppercase*	*0.86*	*0.911*

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
EV	Sans Uppercase	0.753	0.878
EA	Sans Uppercase	0.663	0.865
EV	Serif Uppercase	0.699	0.807
EA	**Serif Uppercase**	**0.529**	**0.711**
FA	Serif Uppercase	0.705	0.841

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

Pair:	Fontface	Mean:	Confidence interval
EA	**Serif Uppercase**	**0.643**	**0.472**

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: GR
A-Label: xn--qxam
U-Label: ελ
Unicode Code Points: U+03B5, U+03BB
String in English: el
String Language: Greek, Modern (1453)
Script: Greek

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: EL in Greek. (epsilon lambda)

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

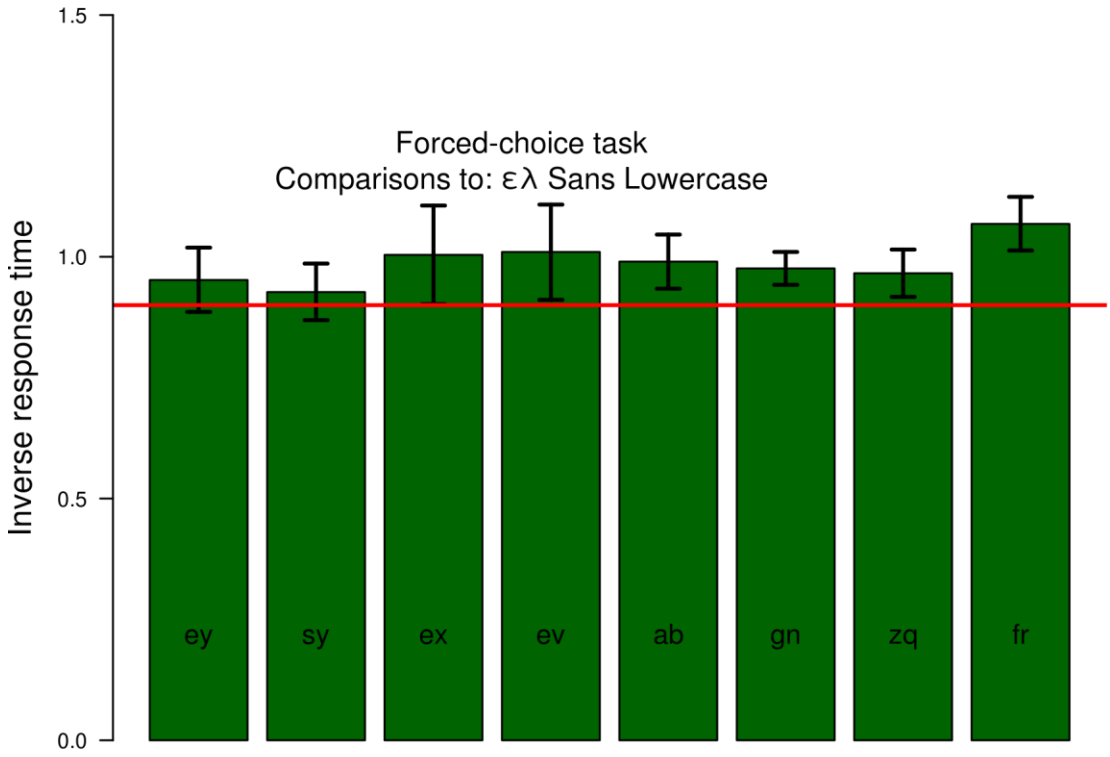
- Testing used two procedures: 1. A delayed match-to-sample forced-choice identification task, and 2. A go/no-go response same-different judgment task. The advantage of method 1 is that it tends to produce differences in response time based on confusability that are highly reliable with minimal observations,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

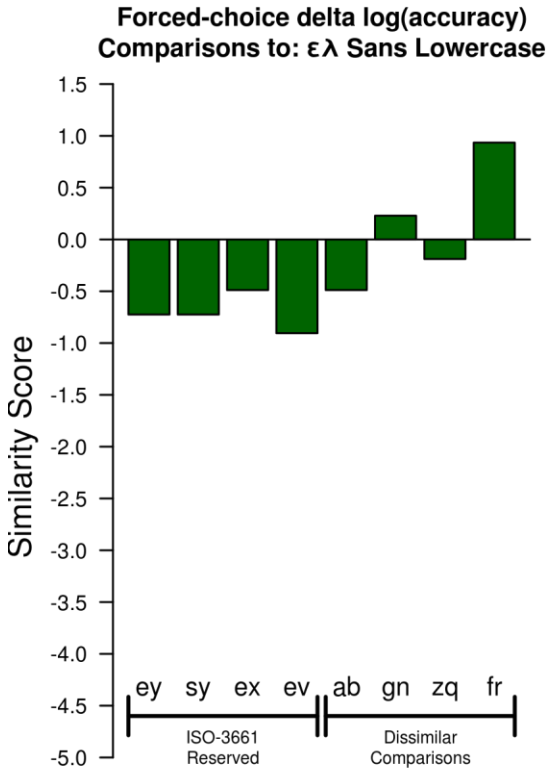
Critical value: 0.9

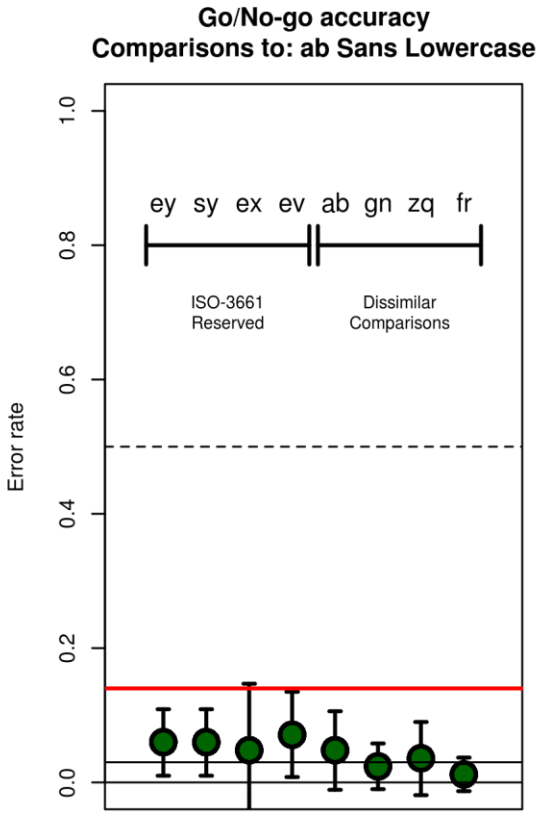


	mean:	sd:	N:	se:	5%	95%
ey	0.952	0.147	21	0.032	0.886	1.019
sy	0.927	0.129	21	0.028	0.869	0.986
ex	1.004	0.224	21	0.049	0.902	1.106
ev	1.01	0.216	21	0.047	0.911	1.108
ab	0.99	0.123	21	0.027	0.934	1.046
gn	0.976	0.074	21	0.016	0.942	1.01
zq	0.966	0.107	21	0.023	0.917	1.015
fr	1.068	0.122	21	0.027	1.013	1.124

Error rate: Sans Lowercase

Critical value: 0.14



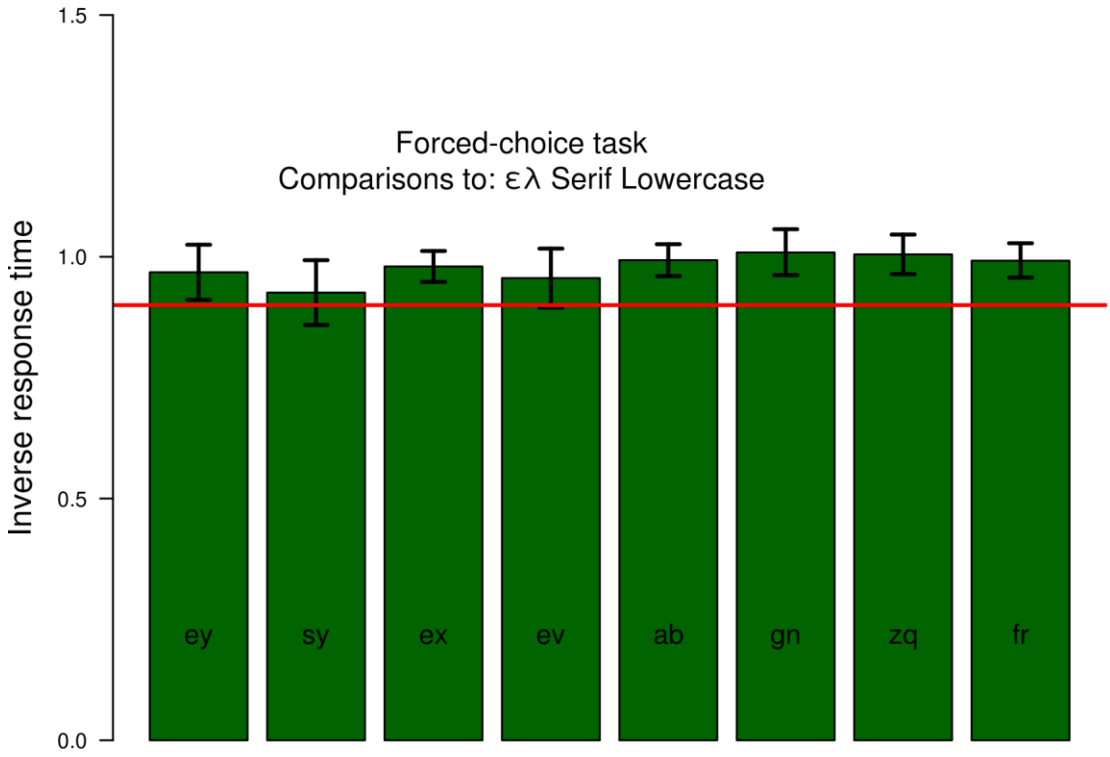


	mean:	sd:	N:	se:	5%	95%
ey	0.06	0.109	21	0.024	0.01	0.109
sy	0.06	0.109	21	0.024	0.01	0.109
ex	0.048	0.218	21	0.048	-0.052	0.147
ev	0.071	0.14	21	0.031	0.008	0.135
ab	0.048	0.128	21	0.028	-0.011	0.106
gn	0.024	0.075	21	0.016	-0.01	0.058
zq	0.036	0.12	21	0.026	-0.019	0.09
fr	0.012	0.055	21	0.012	-0.013	0.037

Correlation between error rate and inverse RT: -0.5092

Inverse response time: Serif Lowercase

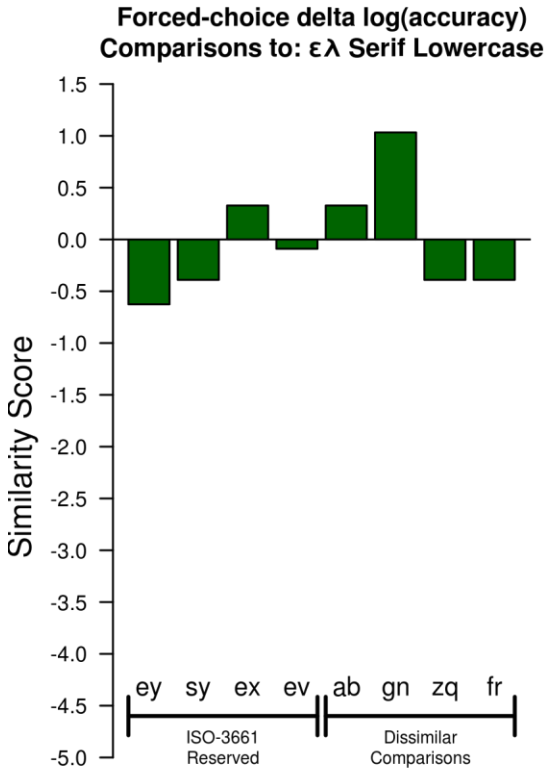
Critical value: 0.9

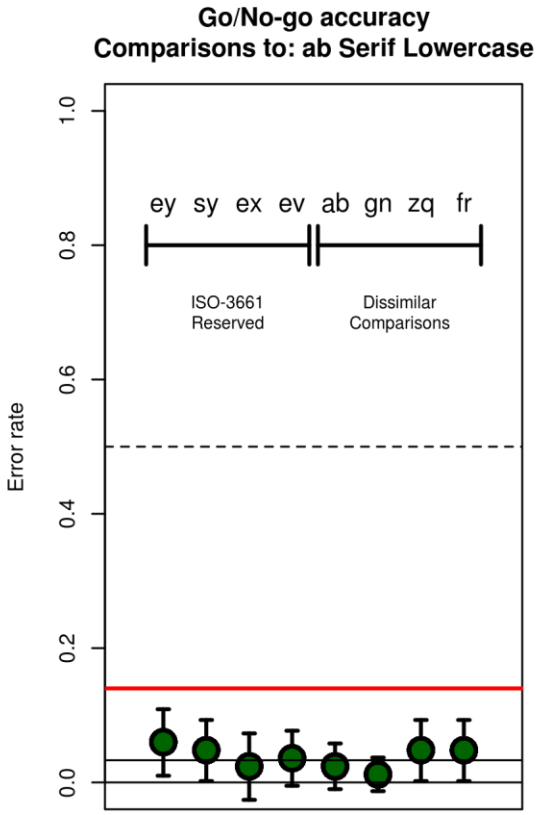


	mean:	sd:	N:	se:	5%	95%
ey	0.968	0.125	21	0.027	0.911	1.025
sy	0.926	0.148	21	0.032	0.859	0.993
ex	0.98	0.07	21	0.015	0.948	1.012
ev	0.956	0.134	21	0.029	0.895	1.017
ab	0.993	0.073	21	0.016	0.96	1.026
gn	1.009	0.105	21	0.023	0.962	1.057
zq	1.005	0.09	21	0.02	0.964	1.046
fr	0.992	0.078	21	0.017	0.957	1.028

Error rate: Serif Lowercase

Critical value: 0.14



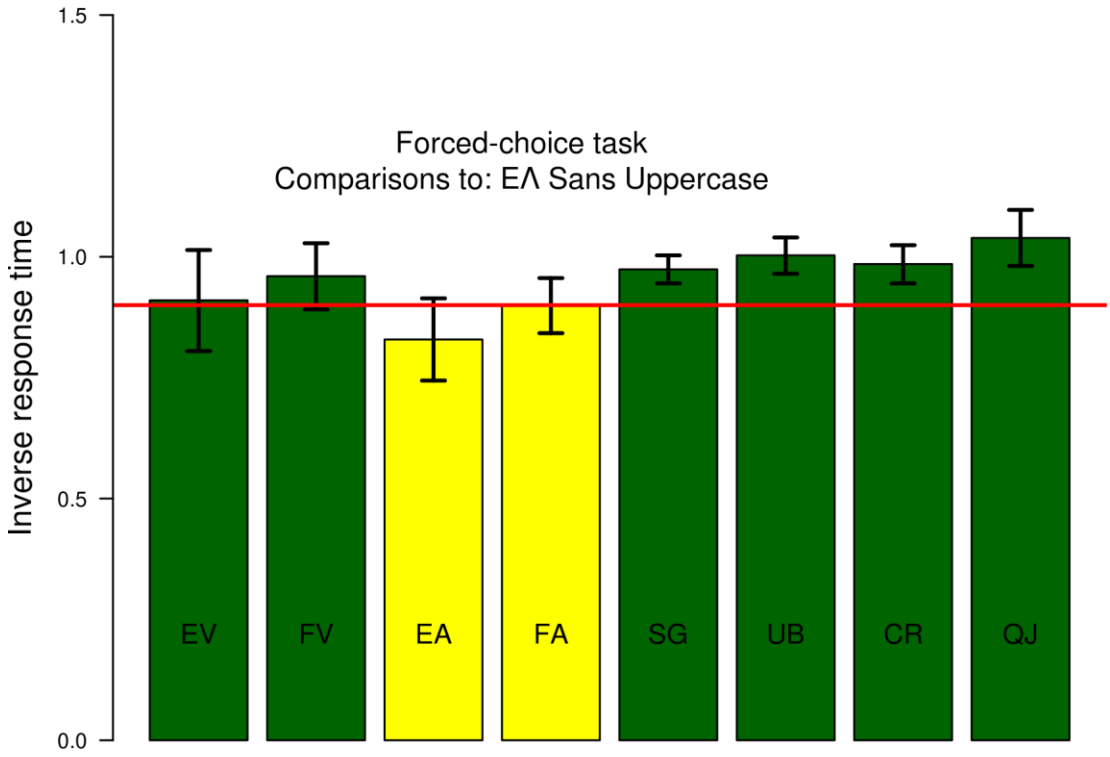


	mean:	sd:	N:	se:	5%	95%
ey	0.06	0.109	21	0.024	0.01	0.109
sy	0.048	0.101	21	0.022	0.002	0.093
ex	0.024	0.109	21	0.024	-0.026	0.073
ev	0.036	0.09	21	0.02	-0.005	0.077
ab	0.024	0.075	21	0.016	-0.01	0.058
gn	0.012	0.055	21	0.012	-0.013	0.037
zq	0.048	0.101	21	0.022	0.002	0.093
fr	0.048	0.101	21	0.022	0.002	0.093

Correlation between error rate and inverse RT: -0.4196

Inverse response time: Sans Uppercase

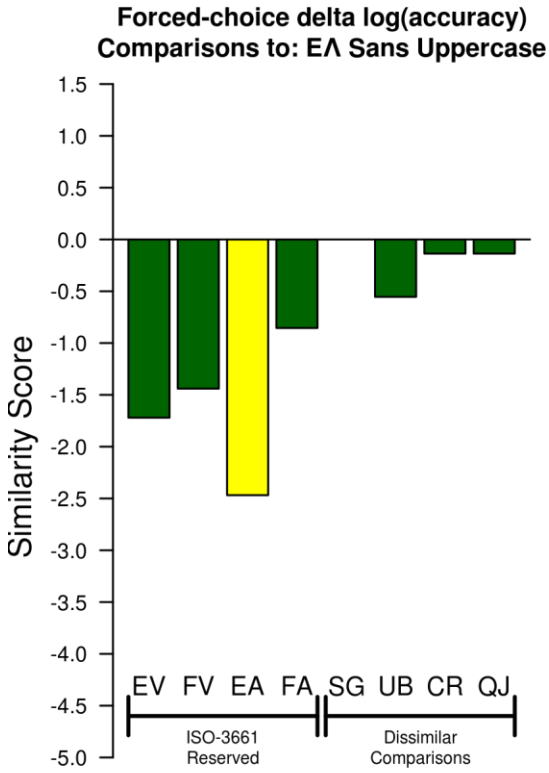
Critical value: 0.9

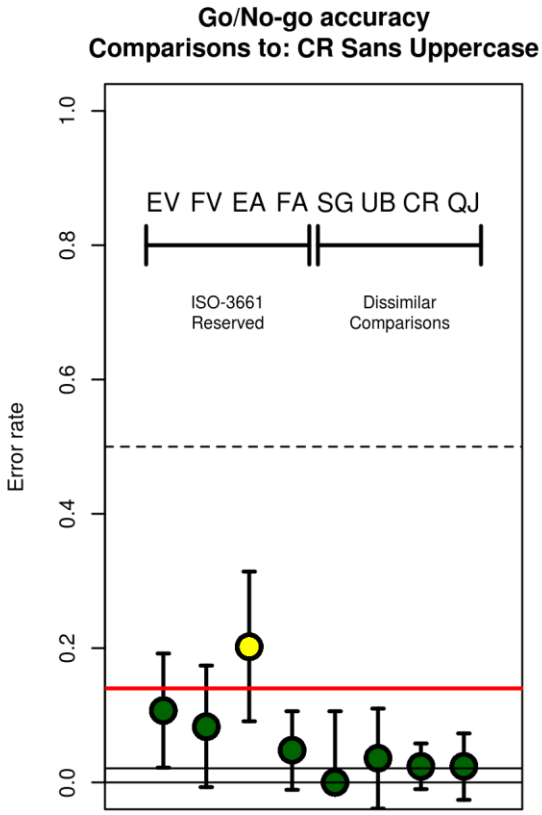


	mean:	sd:	N:	se:	5%	95%
EV	0.91	0.23	21	0.05	0.805	1.014
FV	0.96	0.15	21	0.033	0.891	1.028
EA	0.829	0.187	21	0.041	0.744	0.914
FA	0.899	0.125	21	0.027	0.842	0.956
SG	0.974	0.065	21	0.014	0.945	1.003
UB	1.003	0.082	21	0.018	0.965	1.04
CR	0.985	0.087	21	0.019	0.945	1.024
QJ	1.039	0.127	21	0.028	0.981	1.097

Error rate: Sans Uppercase

Critical value: 0.14



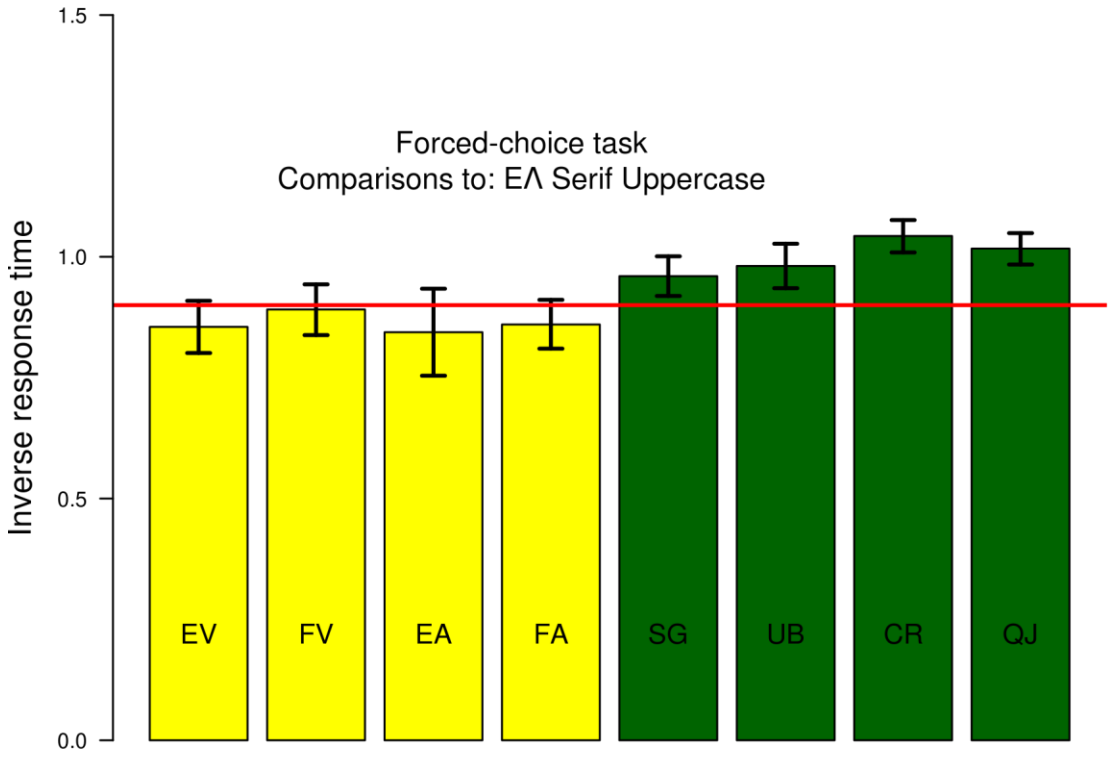


	mean:	sd:	N:	se:	5%	95%
EV	0.107	0.187	21	0.041	0.022	0.192
FV	0.083	0.199	21	0.043	-0.007	0.174
EA	0.202	0.245	21	0.054	0.091	0.314
FA	0.048	0.128	21	0.028	-0.011	0.106
SG	0	0	21	0	-0.011	0.106
UB	0.036	0.164	21	0.036	-0.039	0.11
CR	0.024	0.075	21	0.016	-0.01	0.058
QJ	0.024	0.109	21	0.024	-0.026	0.073

Correlation between error rate and inverse RT: -0.8323

Inverse response time: Serif Uppercase

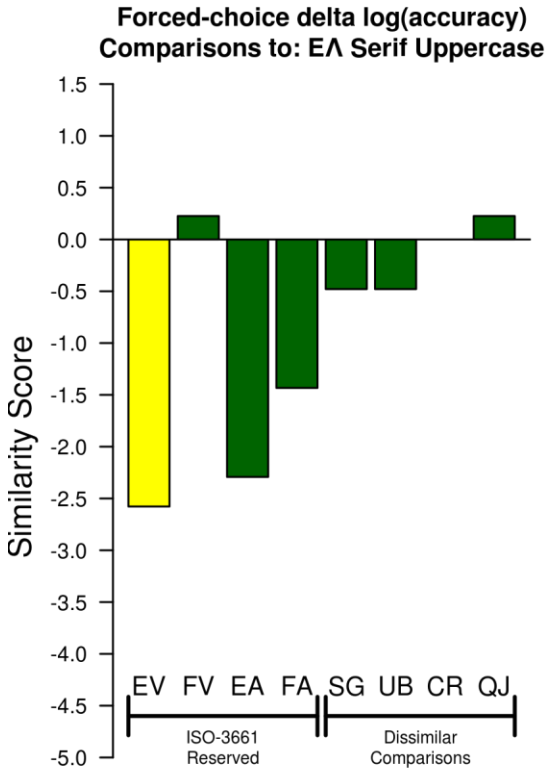
Critical value: 0.9

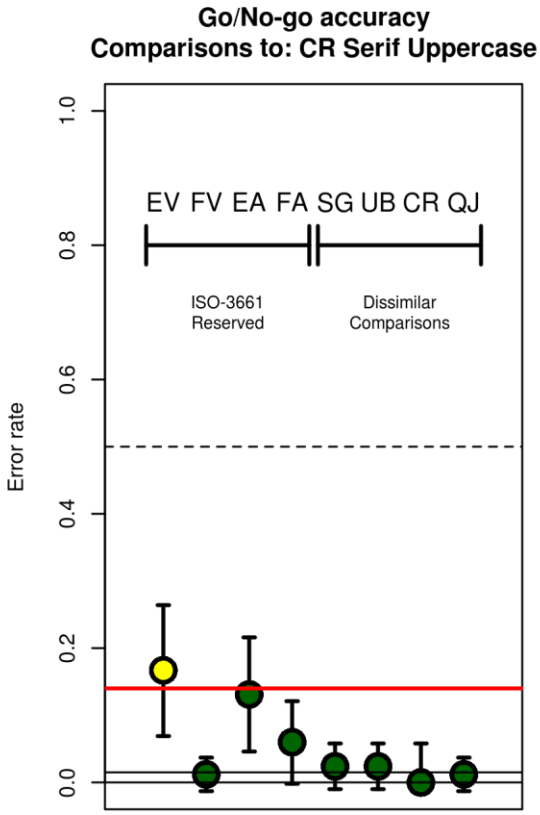


	mean:	sd:	N:	se:	5%	95%
EV	0.855	0.118	21	0.026	0.801	0.909
FV	0.891	0.115	21	0.025	0.838	0.943
EA	0.844	0.197	21	0.043	0.754	0.934
FA	0.86	0.111	21	0.024	0.81	0.911
SG	0.96	0.09	21	0.02	0.919	1.001
UB	0.981	0.1	21	0.022	0.935	1.027
CR	1.043	0.074	21	0.016	1.009	1.076
QJ	1.017	0.071	21	0.015	0.984	1.049

Error rate: Serif Uppercase

Critical value: 0.14





	mean:	sd:	N:	se:	5%	95%
EV	0.167	0.214	21	0.047	0.069	0.264
FV	0.012	0.055	21	0.012	-0.013	0.037
EA	0.131	0.187	21	0.041	0.046	0.216
FA	0.06	0.135	21	0.029	-0.002	0.121
SG	0.024	0.075	21	0.016	-0.01	0.058
UB	0.024	0.075	21	0.016	-0.01	0.058
CR	0	0	21	0	-0.01	0.058
QJ	0.012	0.055	21	0.012	-0.013	0.037

Correlation between error rate and inverse RT: -0.7628

Summary of RT below threshold

Pair:	Fontface	Mean:	Confidence interval	< 0.9
EA	*Sans Uppercase*	*0.829*	*0.914*	
FA	*Sans Uppercase*	*0.899*	*0.956*	
EV	*Serif Uppercase*	*0.855*	*0.909*	
FV	*Serif Uppercase*	*0.891*	*0.943*	
EA	*Serif Uppercase*	*0.844*	*0.934*	
FA	*Serif Uppercase*	*0.86*	*0.911*	

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: EL in Greek. (epsilon lambda)

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

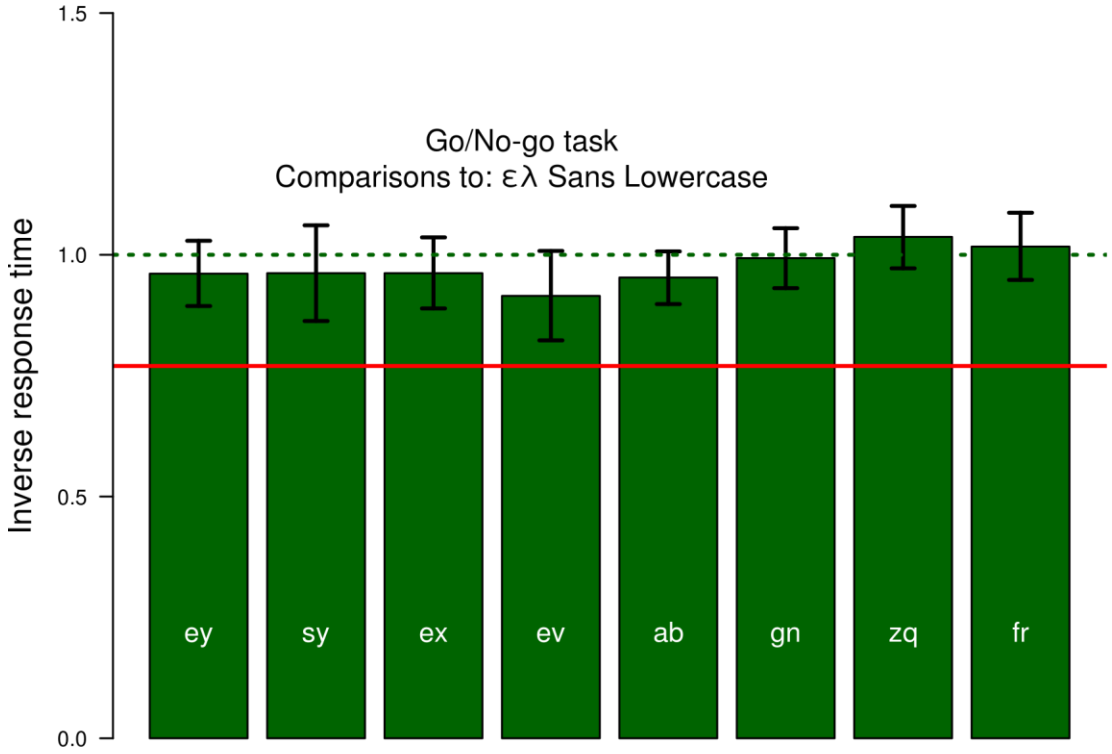
- Testing used two procedures: 1. A delayed match-to-sample forced-choice identification task, and 2. A go/no-go response same-different judgment task. The advantage of method 1 is that it tends to produce differences in response time based on confusability that are highly reliable with minimal observations,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

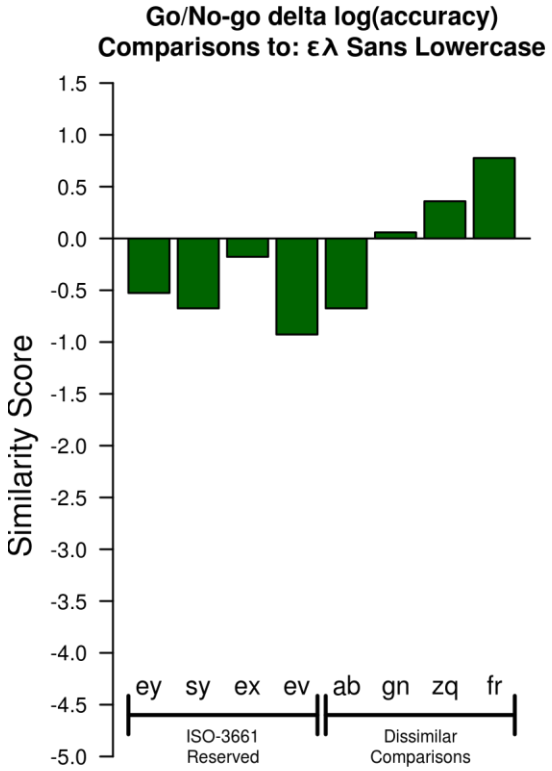
- **Inverse response time: Sans Lowercase**
- Critical value: 0.77

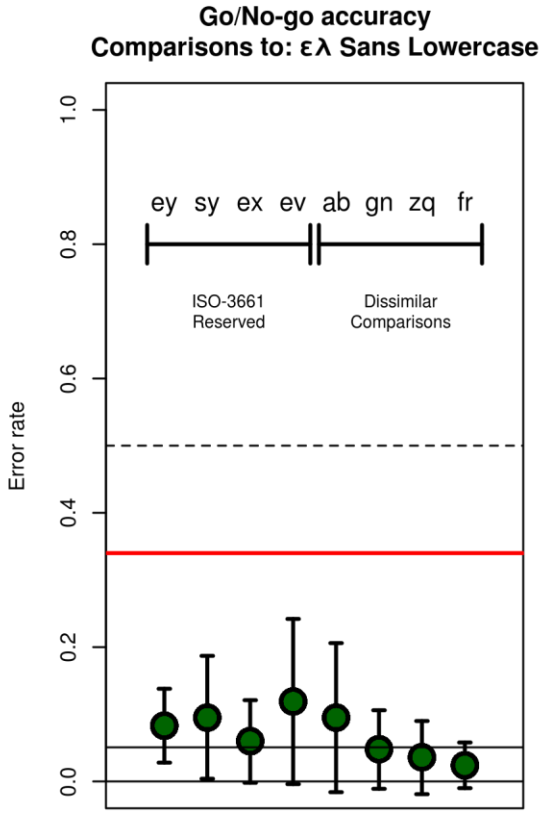


-

	mean:	sd:	N:	se:	5%	95%
ey	0.961	0.148	21	0.032	0.894	1.029
sy	0.962	0.217	21	0.047	0.863	1.061
ex	0.962	0.161	21	0.035	0.889	1.036
ev	0.915	0.202	21	0.044	0.823	1.008
ab	0.953	0.12	21	0.026	0.898	1.007
gn	0.993	0.136	21	0.03	0.931	1.055
zq	1.037	0.142	21	0.031	0.972	1.101
fr	1.017	0.154	21	0.034	0.948	1.087

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



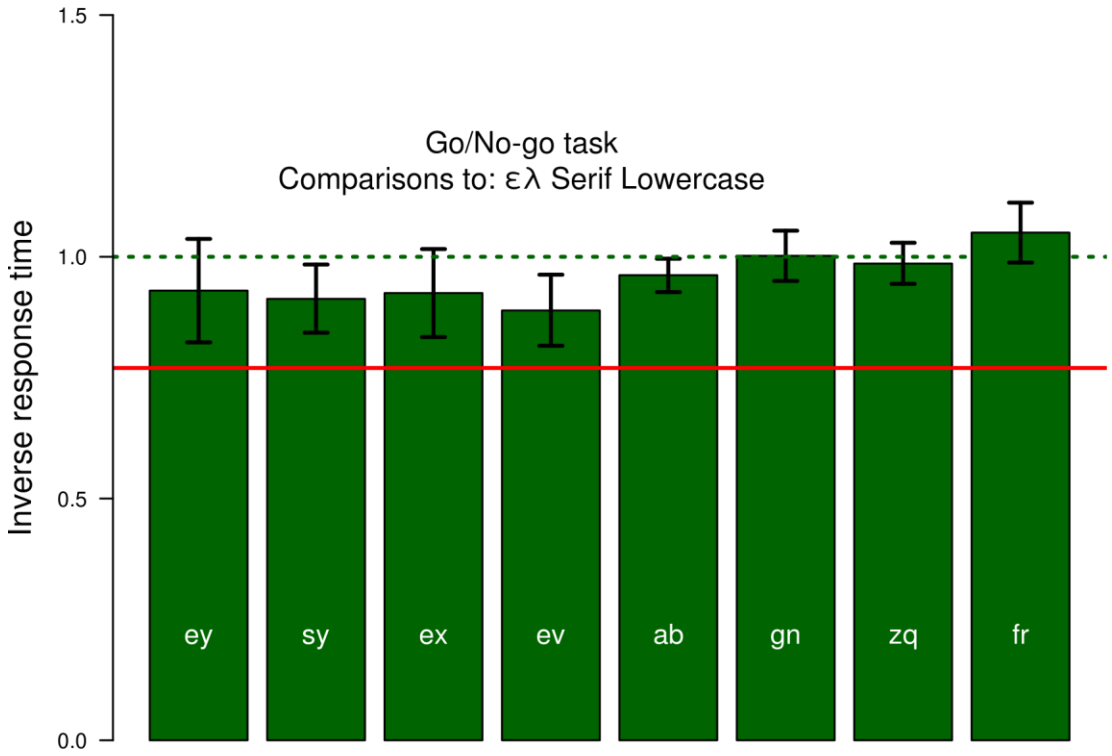


-

	mean:	sd:	N:	se:	5%	95%
ey	0.083	0.121	21	0.026	0.028	0.138
sy	0.095	0.201	21	0.044	0.004	0.187
ex	0.06	0.135	21	0.029	-0.002	0.121
ev	0.119	0.269	21	0.059	-0.004	0.242
ab	0.095	0.243	21	0.053	-0.016	0.206
gn	0.048	0.128	21	0.028	-0.011	0.106
zq	0.036	0.12	21	0.026	-0.019	0.09
fr	0.024	0.075	21	0.016	-0.01	0.058

- Correlation between error rate and inverse RT: -0.9227

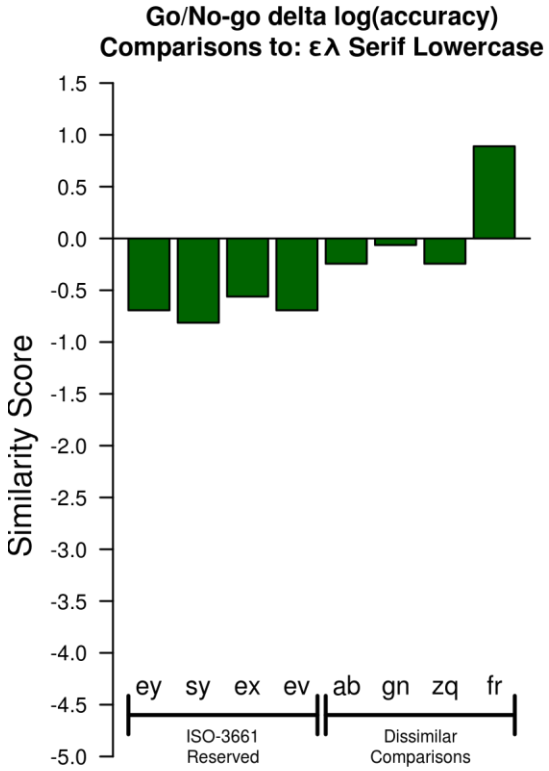
- **Inverse response time: Serif Lowercase**
- Critical value: 0.77

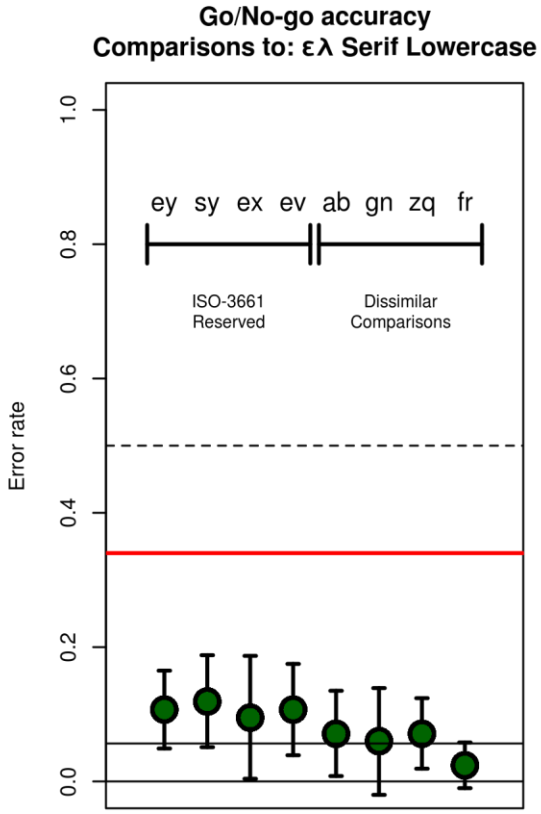


-

	mean:	sd:	N:	se:	5%	95%
ey	0.93	0.236	21	0.051	0.823	1.037
sy	0.913	0.155	21	0.034	0.843	0.984
ex	0.925	0.2	21	0.044	0.834	1.016
ev	0.889	0.162	21	0.035	0.816	0.963
ab	0.962	0.076	21	0.017	0.927	0.996
gn	1.002	0.114	21	0.025	0.95	1.054
zq	0.986	0.093	21	0.02	0.944	1.029
fr	1.05	0.137	21	0.03	0.988	1.112

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



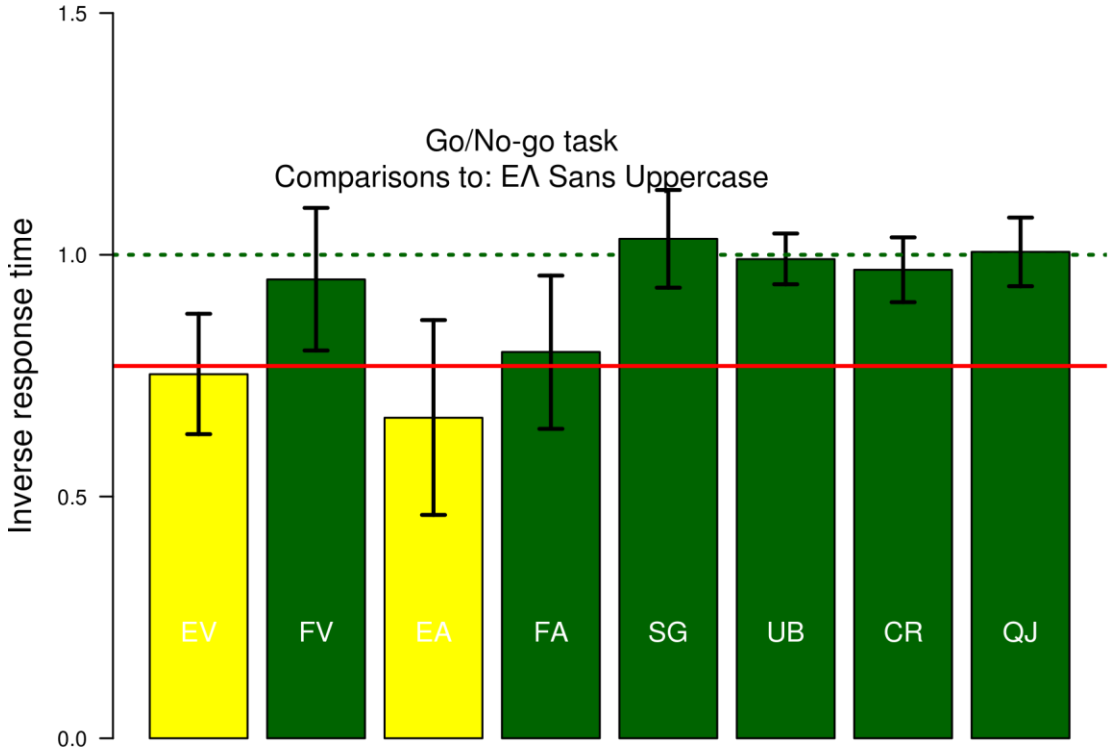


-

	mean:	sd:	N:	se:	5%	95%
ey	0.107	0.127	21	0.028	0.049	0.165
sy	0.119	0.15	21	0.033	0.051	0.188
ex	0.095	0.201	21	0.044	0.004	0.187
ev	0.107	0.149	21	0.033	0.039	0.175
ab	0.071	0.14	21	0.031	0.008	0.135
gn	0.06	0.175	21	0.038	-0.02	0.139
zq	0.071	0.116	21	0.025	0.019	0.124
fr	0.024	0.075	21	0.016	-0.01	0.058

- Correlation between error rate and inverse RT: -0.9565

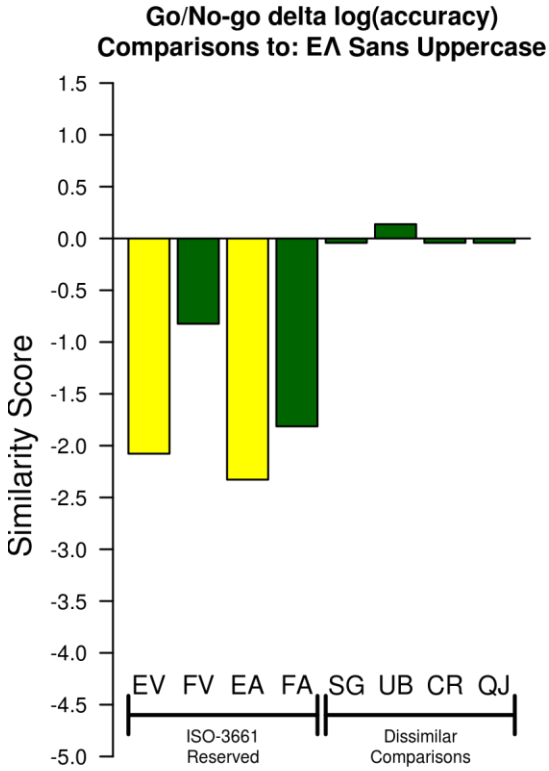
- **Inverse response time: Sans Uppercase**
- Critical value: 0.77

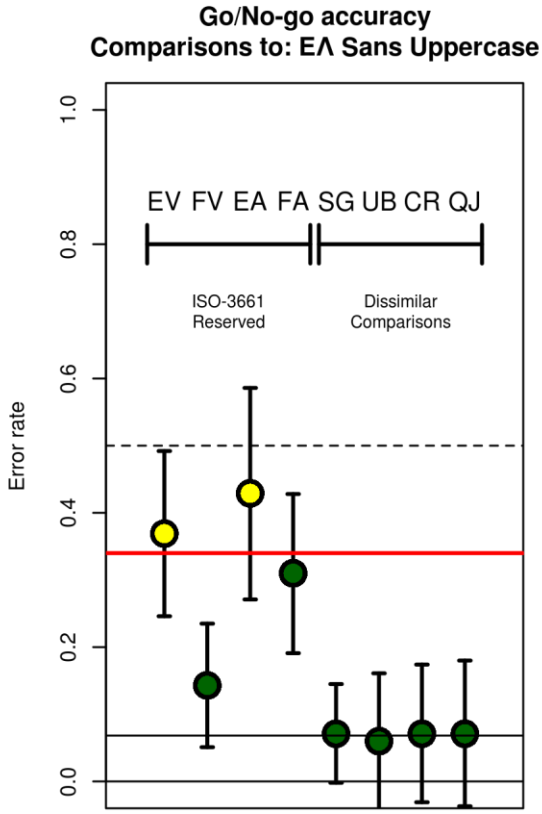


-

	mean:	sd:	N:	se:	5%	95%
EV	0.753	0.273	21	0.06	0.629	0.878
FV	0.949	0.323	21	0.071	0.802	1.097
EA	0.663	0.444	21	0.097	0.462	0.865
FA	0.799	0.348	21	0.076	0.64	0.957
SG	1.033	0.222	21	0.048	0.932	1.134
UB	0.991	0.116	21	0.025	0.939	1.044
CR	0.969	0.147	21	0.032	0.902	1.036
QJ	1.006	0.156	21	0.034	0.935	1.077

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



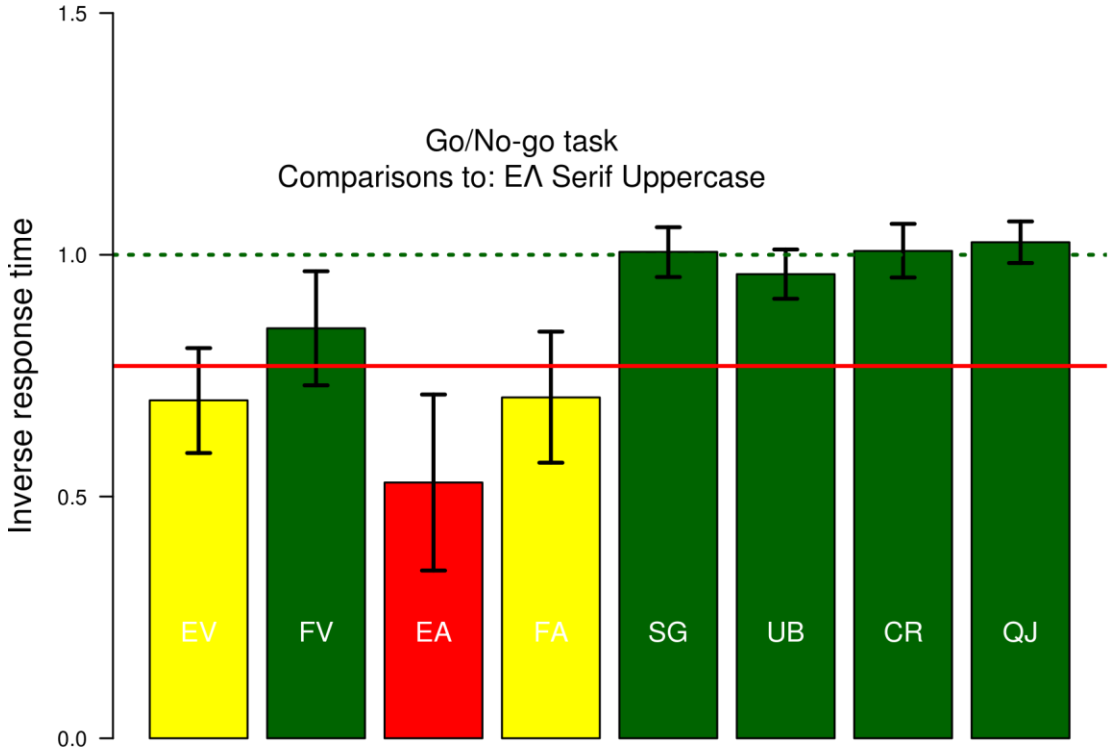


-

	mean:	sd:	N:	se:	5%	95%
EV	0.369	0.269	21	0.059	0.246	0.492
FV	0.143	0.203	21	0.044	0.051	0.235
EA	0.429	0.346	21	0.075	0.271	0.586
FA	0.31	0.261	21	0.057	0.191	0.428
SG	0.071	0.161	21	0.035	-0.002	0.145
UB	0.06	0.222	21	0.049	-0.042	0.161
CR	0.071	0.226	21	0.049	-0.031	0.174
QJ	0.071	0.239	21	0.052	-0.037	0.18

- Correlation between error rate and inverse RT: -0.9871

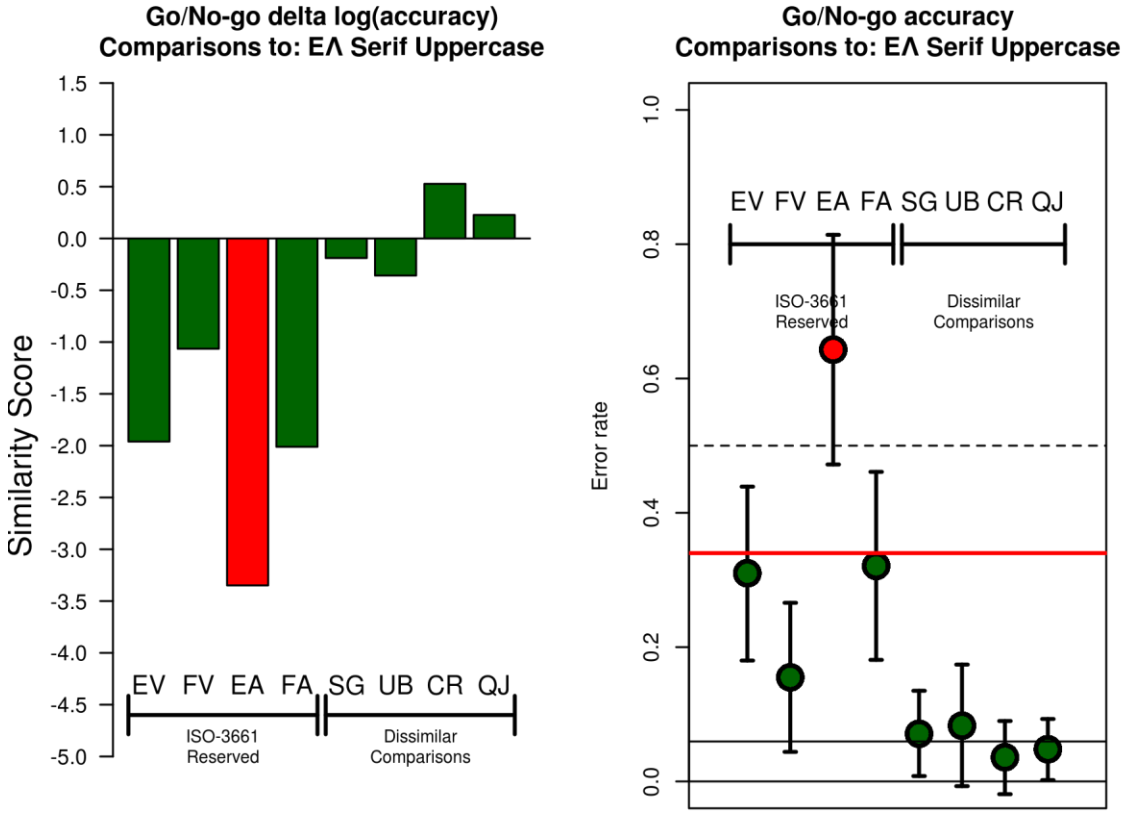
- **Inverse response time: Serif Uppercase**
- Critical value: 0.77



-

	mean:	sd:	N:	se:	5%	95%
EV	0.699	0.238	21	0.052	0.59	0.807
FV	0.848	0.259	21	0.057	0.73	0.966
EA	0.529	0.399	21	0.087	0.347	0.711
FA	0.705	0.297	21	0.065	0.57	0.841
SG	1.006	0.113	21	0.025	0.954	1.057
UB	0.96	0.112	21	0.024	0.909	1.011
CR	1.008	0.122	21	0.027	0.953	1.064
QJ	1.026	0.094	21	0.021	0.983	1.069

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



-

	mean:	sd:	N:	se:	5%	95%
EV	0.31	0.284	21	0.062	0.18	0.439
FV	0.155	0.243	21	0.053	0.044	0.266
EA	0.643	0.376	21	0.082	0.472	0.814
FA	0.321	0.308	21	0.067	0.181	0.461
SG	0.071	0.14	21	0.031	0.008	0.135
UB	0.083	0.199	21	0.043	-0.007	0.174
CR	0.036	0.12	21	0.026	-0.019	0.09
QJ	0.048	0.101	21	0.022	0.002	0.093

- Correlation between error rate and inverse RT: -0.9695

- **Summary of RT below threshold**

- Pair: Fontface Mean: Confidence interval < 0.77
  EV *Sans Uppercase* *0.753* *0.878*
  EA *Sans Uppercase* *0.663* *0.865*
  EV *Serif Uppercase* *0.699* *0.807*
  EA **Serif Uppercase** **0.529** **0.711**
  FA *Serif Uppercase* *0.705* *0.841*
  Italic indicates mean surpasses threshold. Bold indicates mean significantly surpasses threshold.

- **Summary of Error rate above threshold**

- Pair: Fontface Mean: Confidence interval > 0.34
  EA **Serif Uppercase** **0.643** **0.472**
  Italic indicates mean surpasses threshold. Bold indicates mean significantly surpasses threshold.

Annex B - Final Report of the EPSRP for the application for EL in Greek

Final Report of the EPSRP for the application for EL in Greek

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 ελ/ ΕΛ (.el/.EL in Greek)

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
EA	*Sans Uppercase*	*0.829*	*0.914*
FA	*Sans Uppercase*	*0.899*	*0.956*
EV	*Serif Uppercase*	*0.855*	*0.909*
FV	*Serif Uppercase*	*0.891*	*0.943*
EA	*Serif Uppercase*	*0.844*	*0.934*
FA	*Serif Uppercase*	*0.86*	*0.911*

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
EV	Sans Uppercase	0.753	0.878
EA	Sans Uppercase	0.663	0.865
EV	Serif Uppercase	0.699	0.807
EA	**Serif Uppercase**	**0.529**	**0.711**
FA	Serif Uppercase	0.705	0.841

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

Pair:	Fontface	Mean:	Confidence interval
EA	**Serif Uppercase**	**0.643**	**0.472**

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.

Candidate: ελ/ ΕΛ (.el/.EL in Greek)

	Serif lowercase **Times New Roman**	**Sans serif lowercase** **Segoe UI**
Evaluation target	**ελ**	**ελ**
Similar Latin	ey, sy, ex, ev	ey, sy, ex, ev
Dissimilar Latin comparisons:	ab,gn,zq,fr	ab,gn,zq,fr
Other Highly similar comparisons	none evaluated	none evaluated

Evaluation Target	**Serif uppercase** **Times new roman**	**Sans serif uppercase** **Segoe UI Uppercase**
	ΕΛ	ΕΛ
Similar Latin	EV, FV,EA,FA	EV,FV,EA,FA
Dissimilar Latin comparisons:	SG,UB,CR,QJ	SG UB,CR,QJ
Other Highly similar comparisons	None evaluated	None evaluated

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

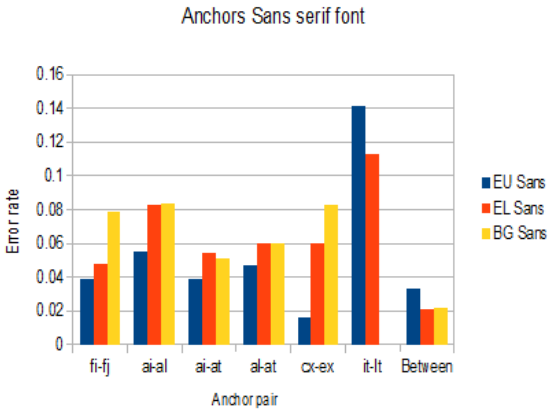
Testing used two procedures: 1. A delayed match-to-sample forced-choice identification task, and 2. A go/no-go response same-different judgment task. The advantage of method 1 is that it tends to produce differences in response time based on confusability that are highly reliable with minimal observations,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

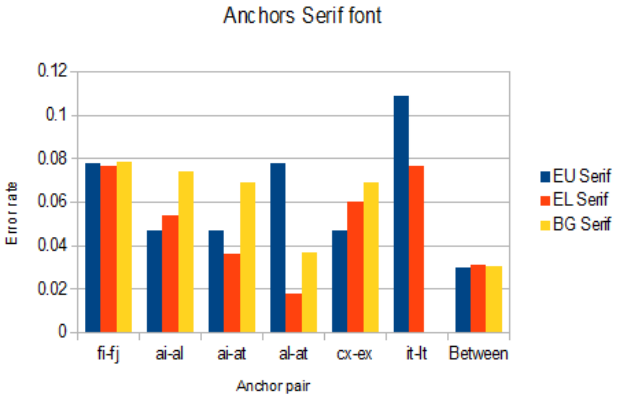
Each test was performed in a blocked design in the same order across participants. Each set of stimuli will appear in a contiguous block. Testing was designed to assess the similarity between the target and (1) any of a set of highly-similar Latin character pairs in the same case (2) a set of 3-4 dissimilar 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





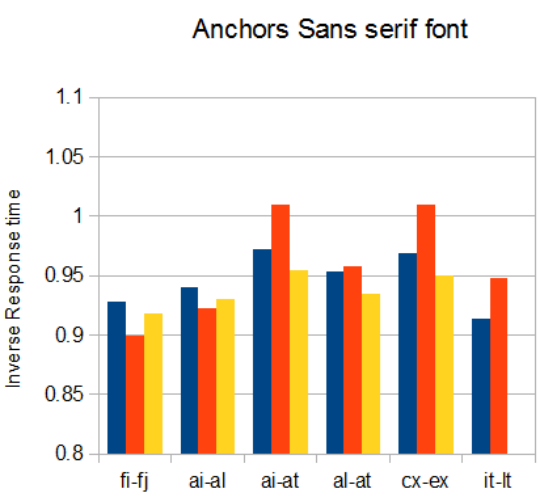
Error Rate

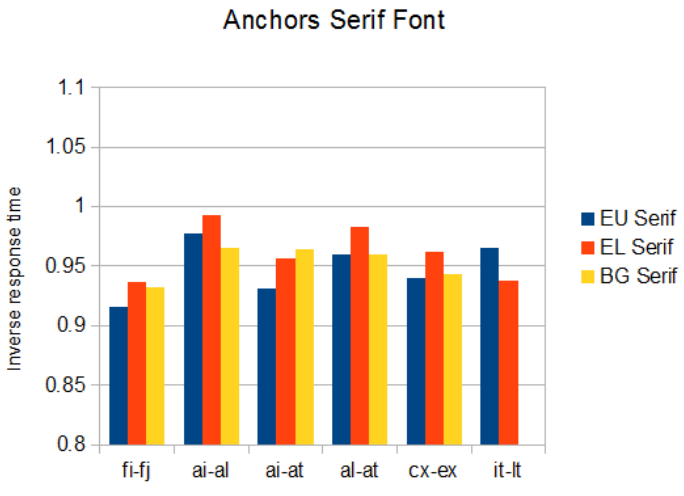
Option	EU Sans	EL Sans	BG Sans
fi-fj	0.039	0.048	0.0787
ai-al	0.055	0.083	0.0833
ai-at	0.039	0.054	0.0509
al-at	0.047	0.06	0.0602
cx-ex	0.016	0.06	0.0827
it-lt	0.141	0.113	
Between	**0.033**	**0.021**	**0.0217**

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

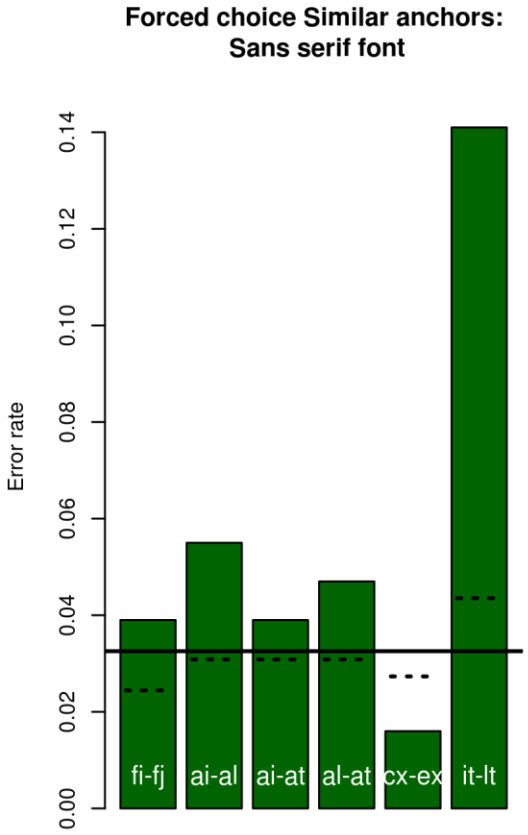


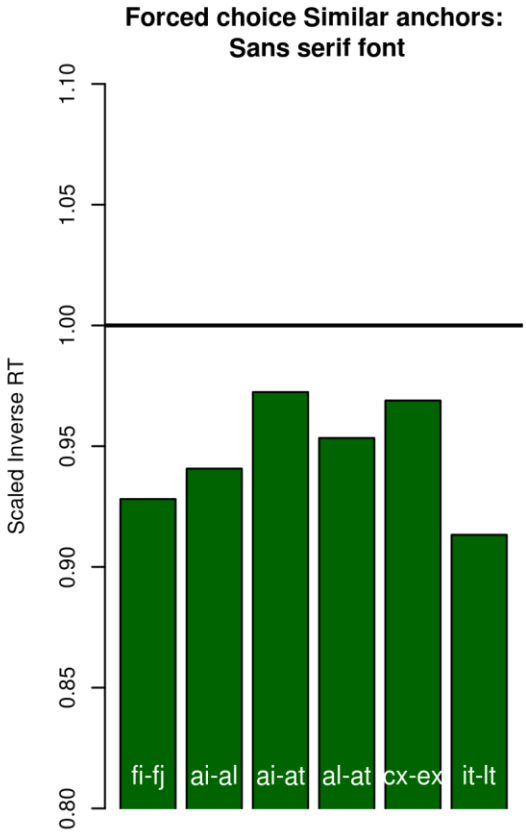


Option	EU Sans	EL Sans	BG Sans
fi-fj	0.9281	0.8995	0.918
ai-al	0.9407	0.9225	0.93
ai-at	0.9724	1.0096	0.955
al-at	0.9534	0.9584	0.935
cx-ex	0.9689	1.01	0.95
it-lt	0.9133	0.9483	-

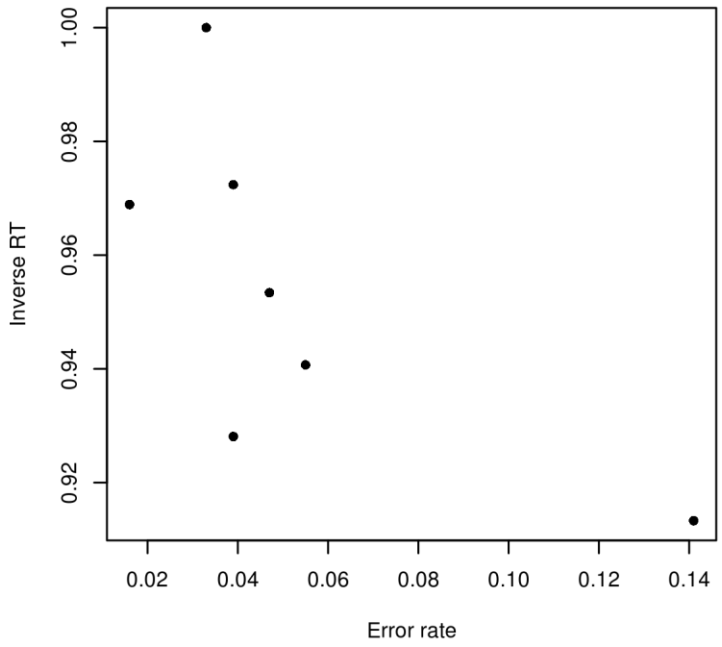
Option	EU Serif	EL Serif	BG Serif
fi-fj	0.9155	0.9371	0.932
ai-al	0.9773	0.9925	0.965
ai-at	0.9316	0.9561	0.964
al-at	0.9596	0.9826	0.96
cx-ex	0.9401	0.962	0.943
it-lt	0.9648	0.9382	-

- 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.**

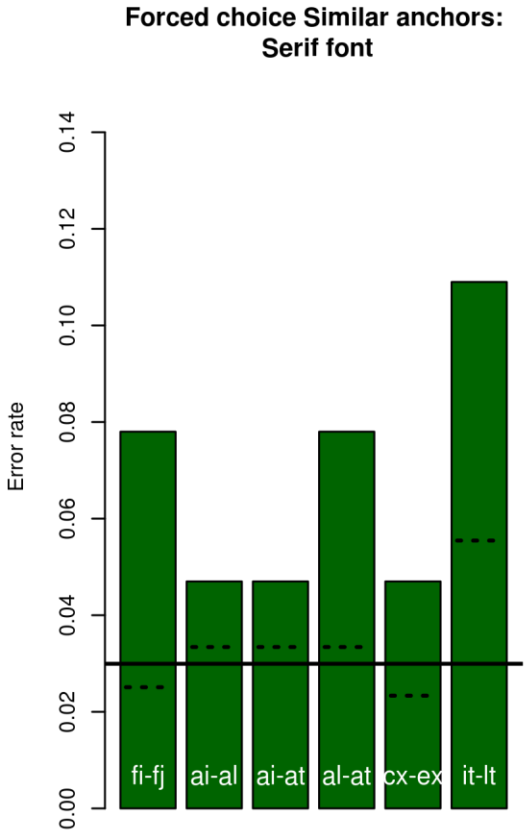


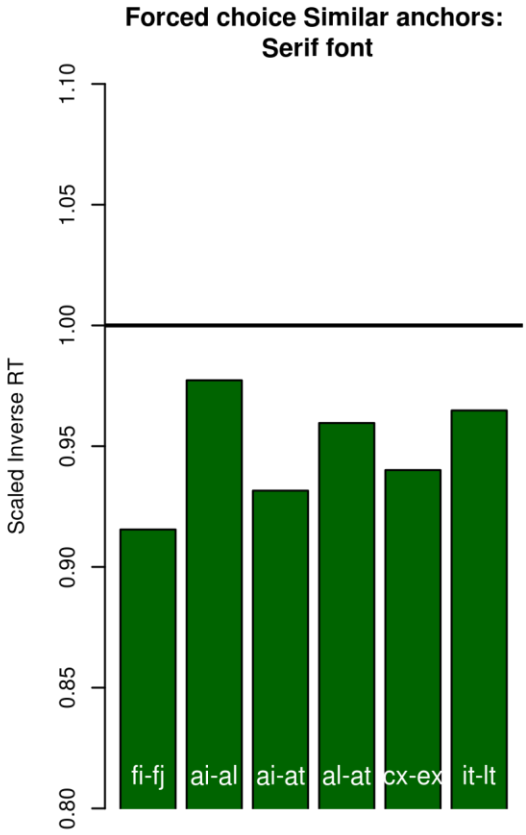


Option	Error rate	Between error rate	Inverse RT	Log-odds delta accuracy
fi-fj	0.039	0.024	0.9281	-0.484
ai-al	0.055	0.031	0.9407	-0.597
ai-at	0.039	0.031	0.9724	-0.244
al-at	0.047	0.031	0.9534	-0.597
cx-ex	0.016	0.027	0.9689	0.571
it-lt	0.141	0.044	0.9133	-1.28
Between	0.033	0.033	1	0

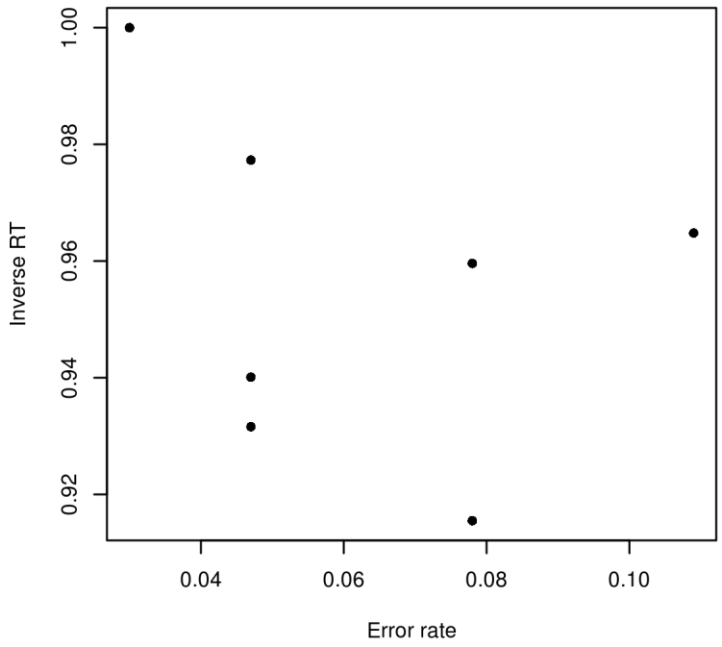


Correlation between error rate and inverse RT: -0.6925

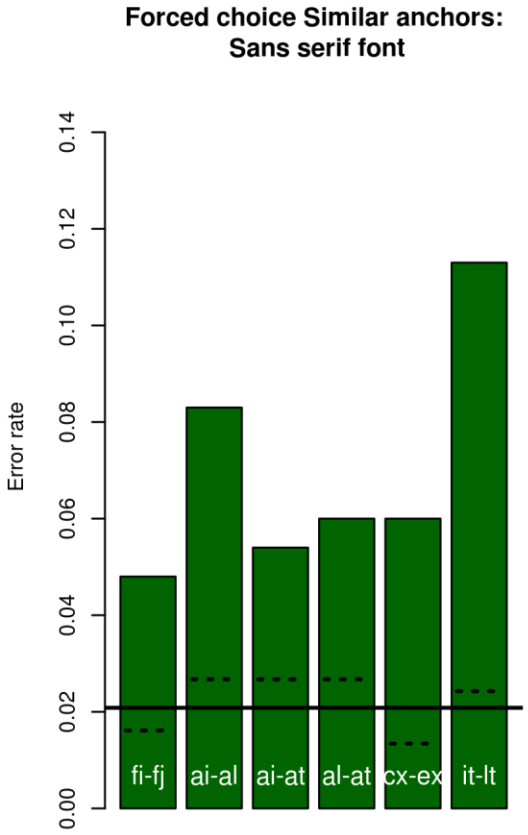


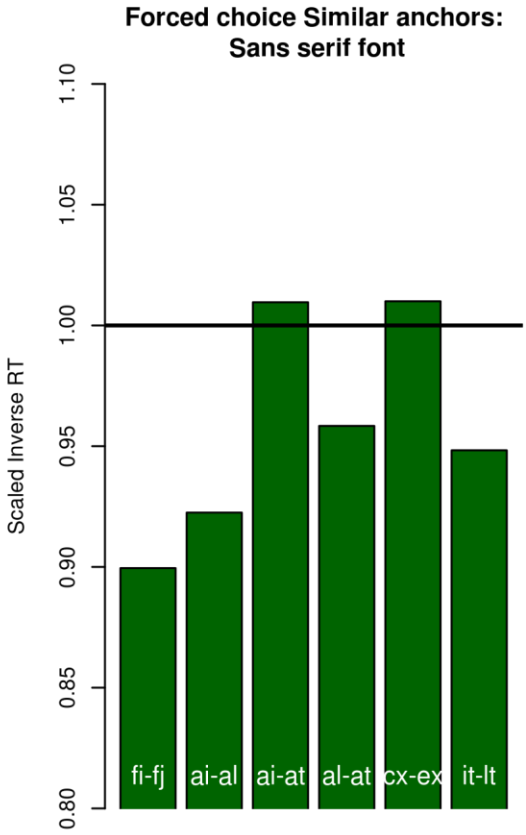


Option	Error rate	Between error rate	Inverse RT	Log-odds delta accuracy
fi-fj	0.078	0.025	0.9155	-1.192
ai-al	0.047	0.033	0.9773	-0.352
ai-at	0.047	0.033	0.9316	-0.352
al-at	0.078	0.033	0.9596	-0.352
cx-ex	0.047	0.023	0.9401	-0.721
it-lt	0.109	0.055	0.9648	-0.738
Between	0.03	0.03	1	0

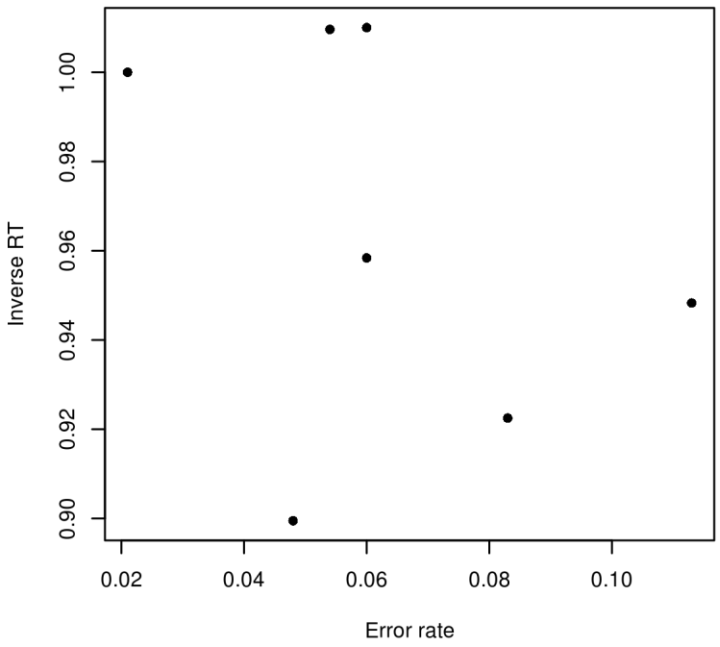


Correlation between error rate and inverse RT: -0.2772

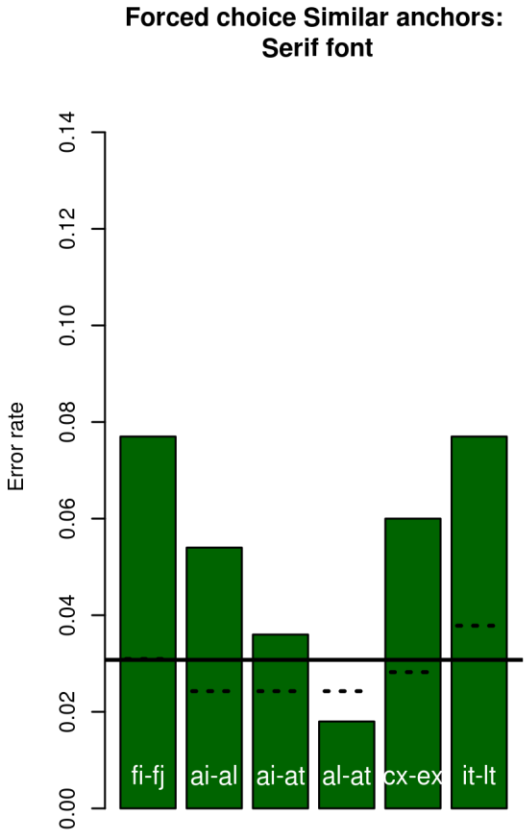


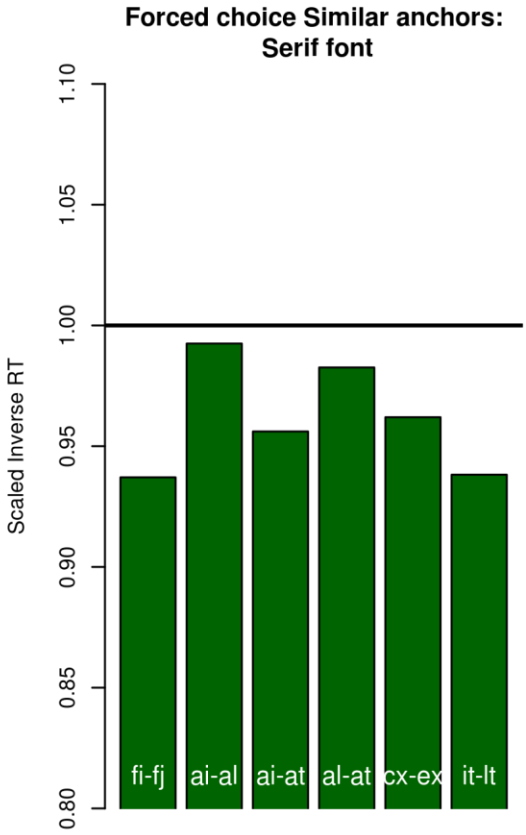


Option	Error rate	Between error rate	Inverse RT	Log-odds delta accuracy
fi-fj	0.048	0.016	0.8995	-1.114
ai-al	0.083	0.027	0.9225	-1.197
ai-at	0.054	0.027	1.0096	-0.723
al-at	0.06	0.027	0.9584	-1.197
cx-ex	0.06	0.013	1.01	-1.537
it-lt	0.113	0.024	0.9483	-1.635
Between	0.021	0.021	1	0

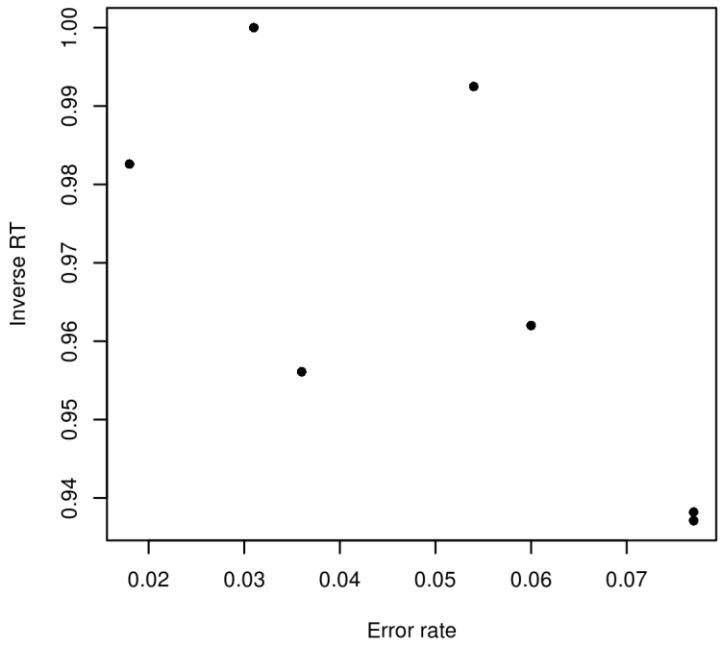


Correlation between error rate and inverse RT: -0.353



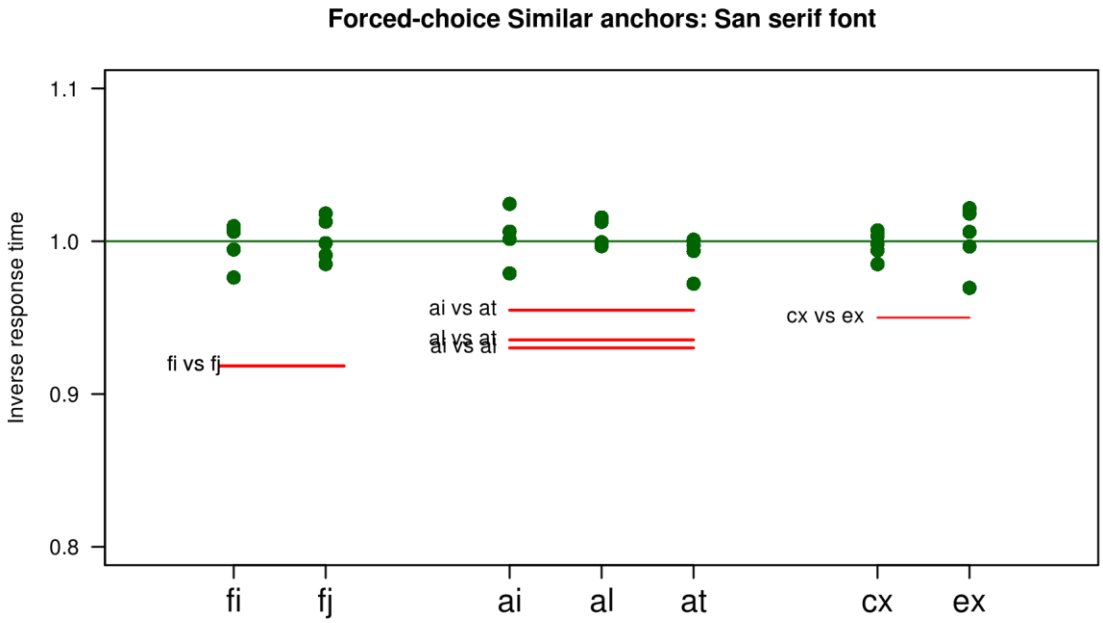


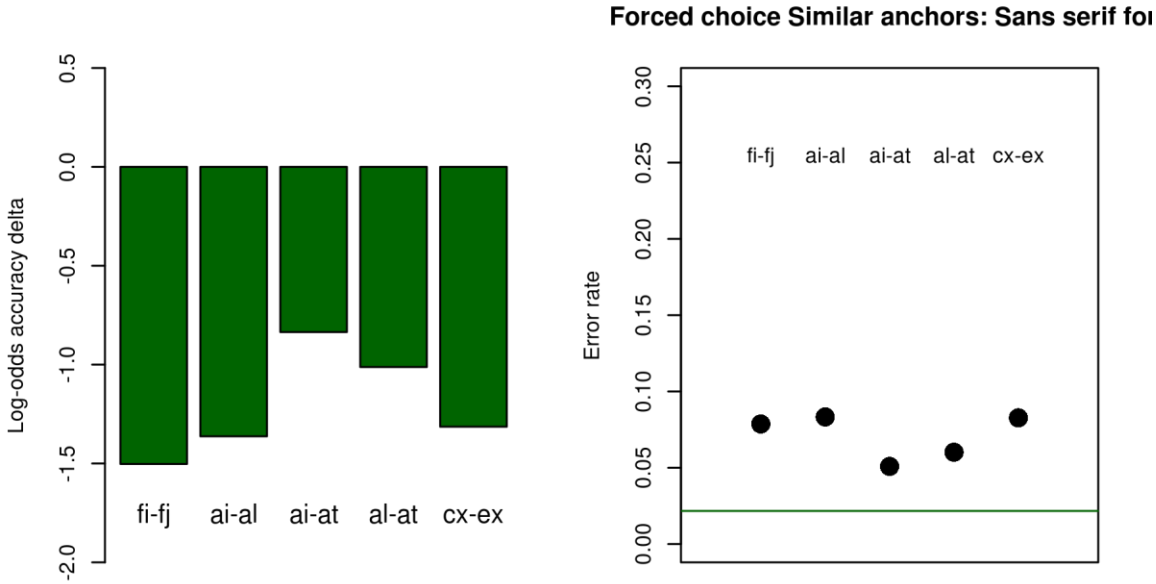
Option	Error rate	Between error rate	Inverse RT	Log-odds delta accuracy
fi-fj	0.077	0.031	0.9371	-0.966
ai-al	0.054	0.024	0.9925	-0.822
ai-at	0.036	0.024	0.9561	-0.398
al-at	0.018	0.024	0.9826	-0.822
cx-ex	0.06	0.028	0.962	-0.779
it-lt	0.077	0.038	0.9382	-0.757
Between	0.031	0.031	1	0

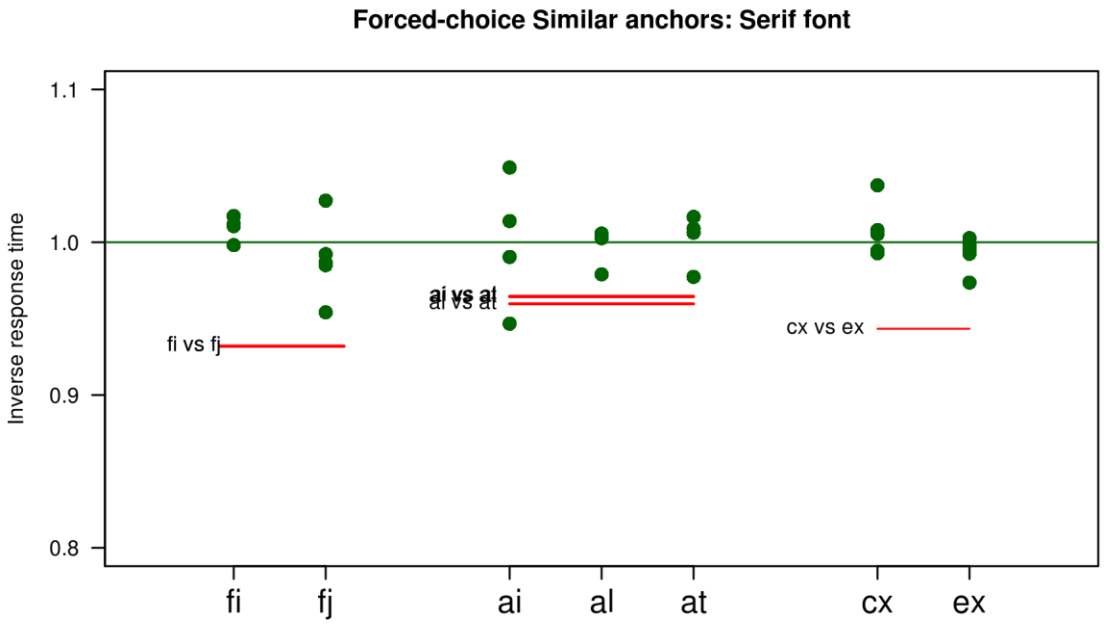


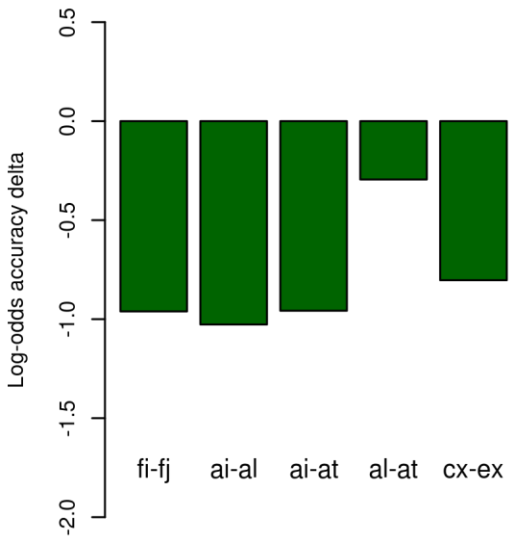
Correlation between error rate and inverse RT: -0.7193

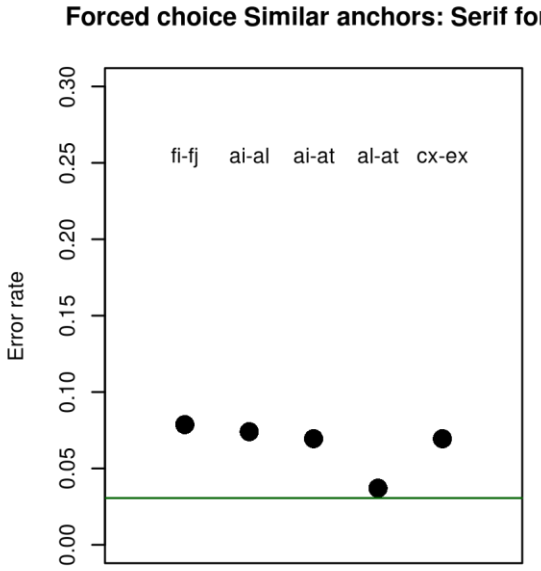
The next figure shows comparisons of similar latin pairs. These serve as a comparison set, with the logic that any new pair evaluated to be less similar than these anchors is justifiably allowable.






Forced-choice Similar anchors: Serif font
Inverse response time
1.1
1.0
0.9
0.8
fi vs fj
ai vs al
al vs at
cx vs ex
fi
fj
ai
al
at
cx
ex


Log-odds accuracy delta
0.5
0.0
-0.5
-1.0
-1.5
-2.0
fi-fj
ai-al
ai-at
al-at
cx-ex


Forced choice Similar anchors: Serif font
Error rate
0.30
0.25
0.20
0.15
0.10
0.05
0.00
fi-fj
ai-al
ai-at
al-at
cx-ex

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

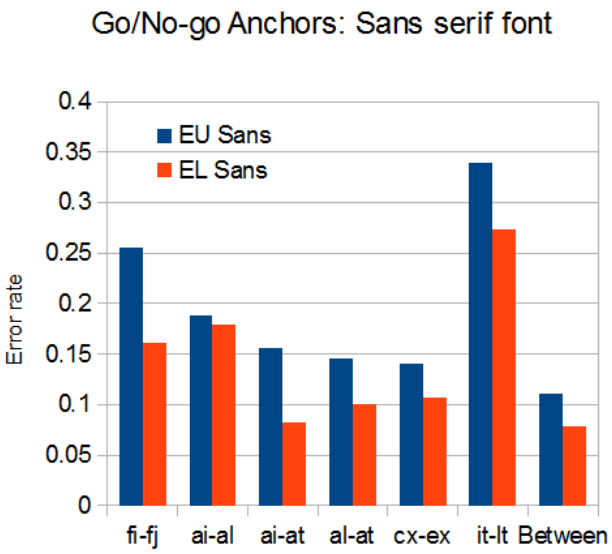
Log-odds difference in accuracy

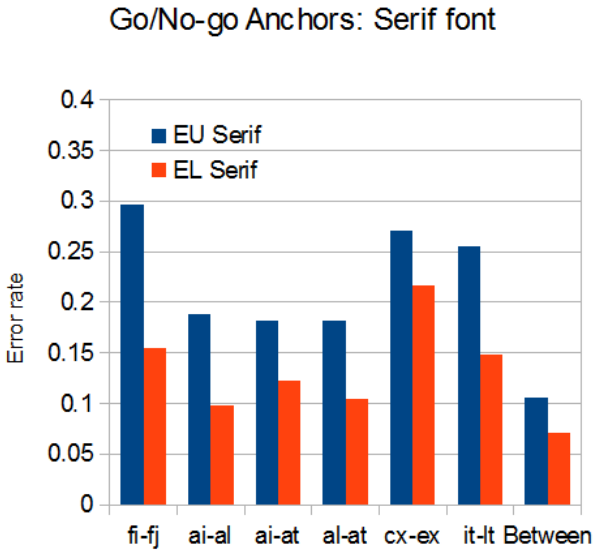
	fi-fj	ai-al	ai-at	al-at	cx-ex
Sans serif	-1.5025	-1.3627	-0.8355	-1.0124	-1.3141
Serif	-0.961	-1.027	-0.9575	-0.2946	-0.8034

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





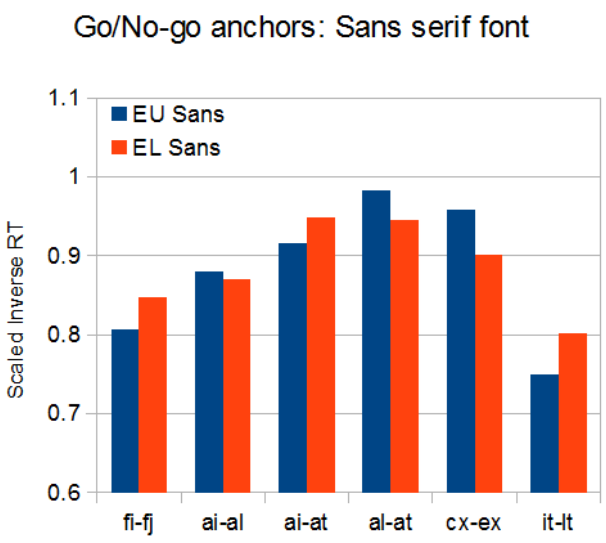
Option	EU Sans	EL Sans
fi-fj	0.255	0.161
ai-al	0.188	0.179
ai-at	0.156	0.083
al-at	0.146	0.101
cx-ex	0.141	0.107
it-lt	0.339	0.274
Between	**0.111**	**0.079**

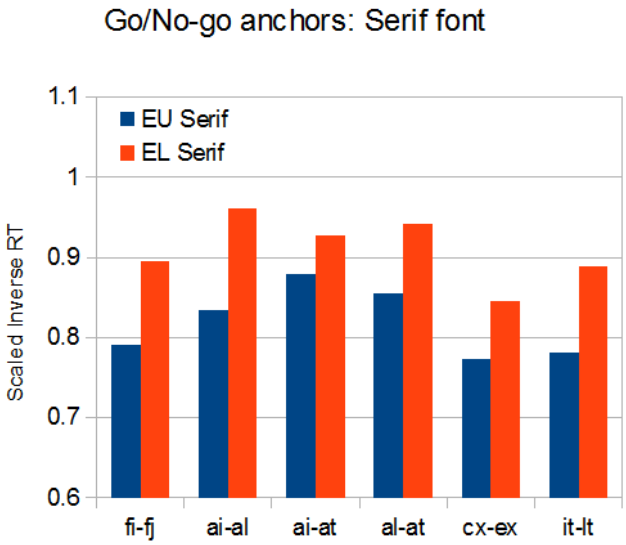
Option	EU Serif	EL Serif
fi-fj	0.297	0.155
ai-al	0.188	0.098
ai-at	0.182	0.122
al-at	0.182	0.104
cx-ex	0.271	0.217
it-lt	0.255	0.149
Between	**0.106**	**0.071**

- 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

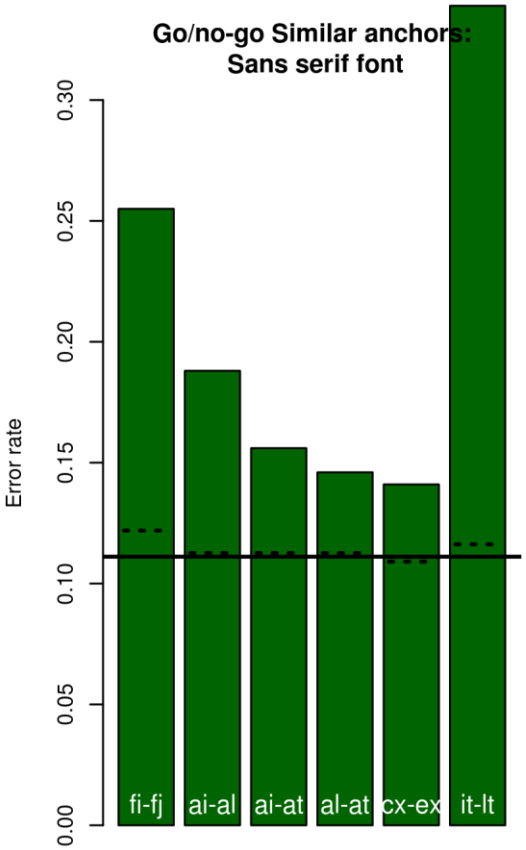




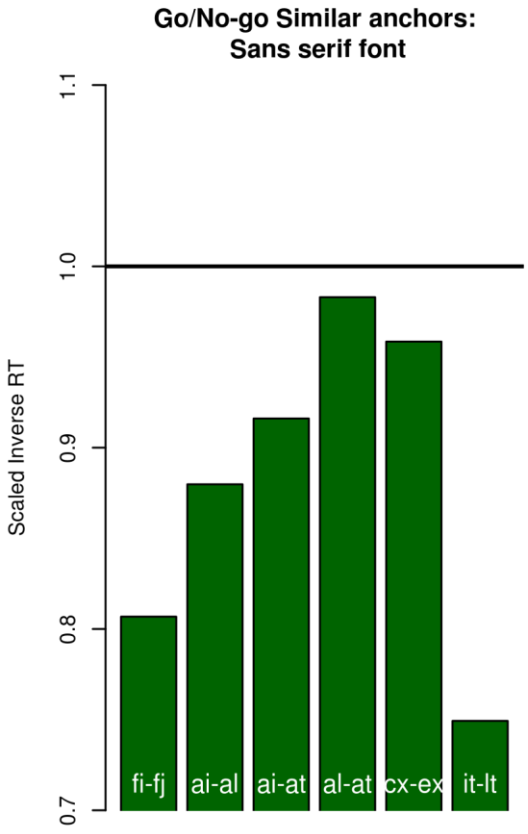
Option	EU Sans	EL Sans
fi-fj	0.8068	0.8472
ai-al	0.8798	0.8704
ai-at	0.9161	0.9486
al-at	0.983	0.9455
cx-ex	0.9585	0.9014
it-lt	0.7493	0.802

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

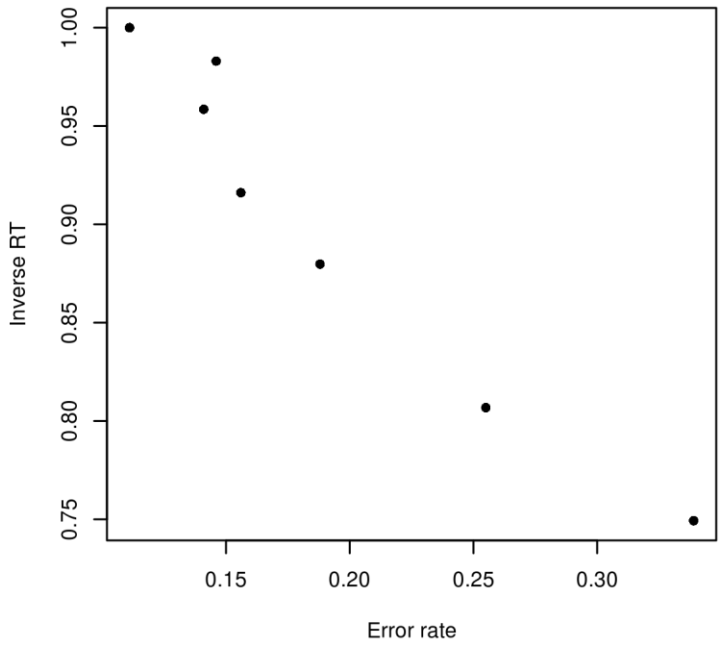
- Test-retest reliability was .906 for sans and .79 for serif. These values are already scaled, so that 1.0 is the average 'different' value.
- EL study produced higher values in the serif font. This is consistent with the overall higher accuracy, and is not a speed-accuracy tradeoff..
- Several cases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



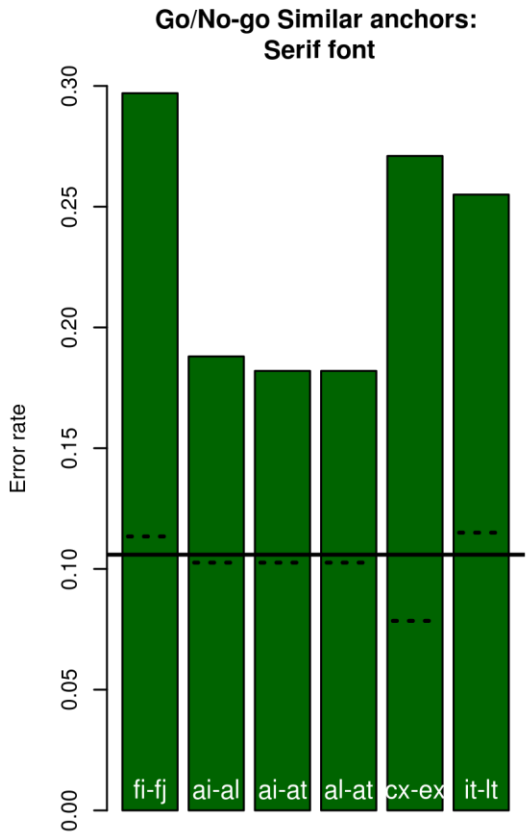
Candidate: EU in Greek. (epsilon upsilon)



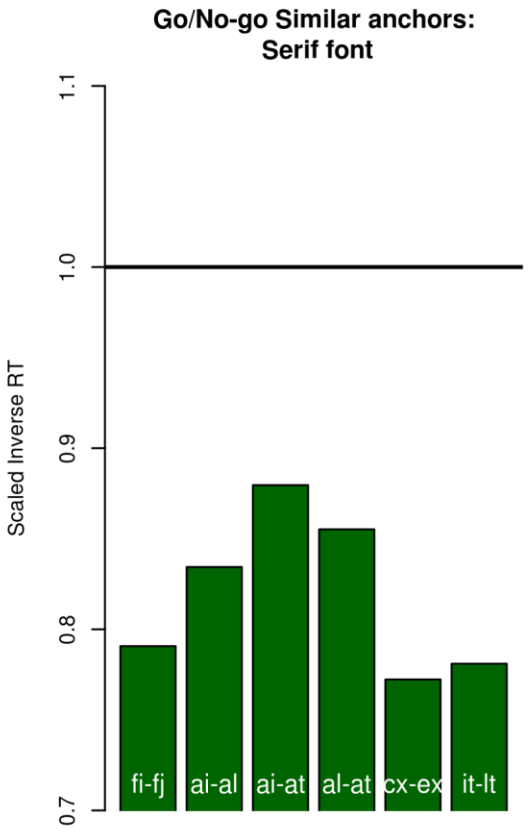
Candidate: EU in Greek. (epsilon upsilon)



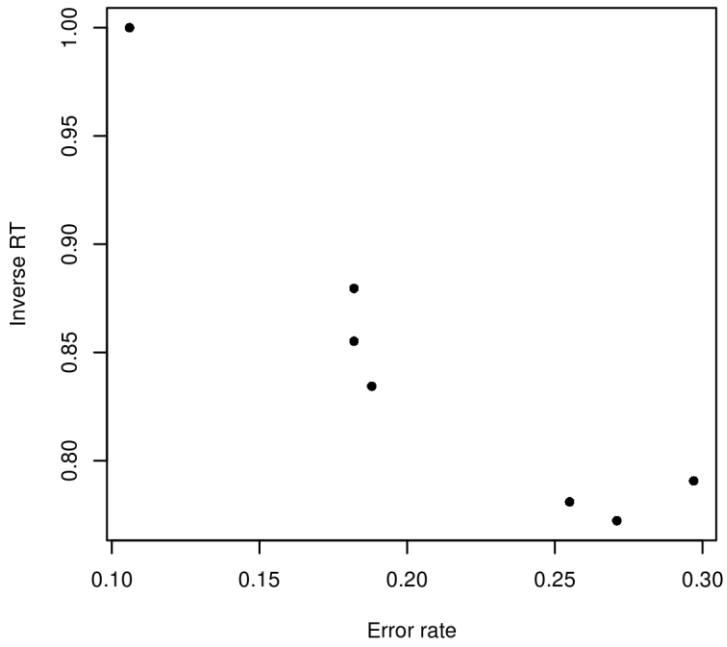
Correlation between error rate and inverse RT: -0.9716



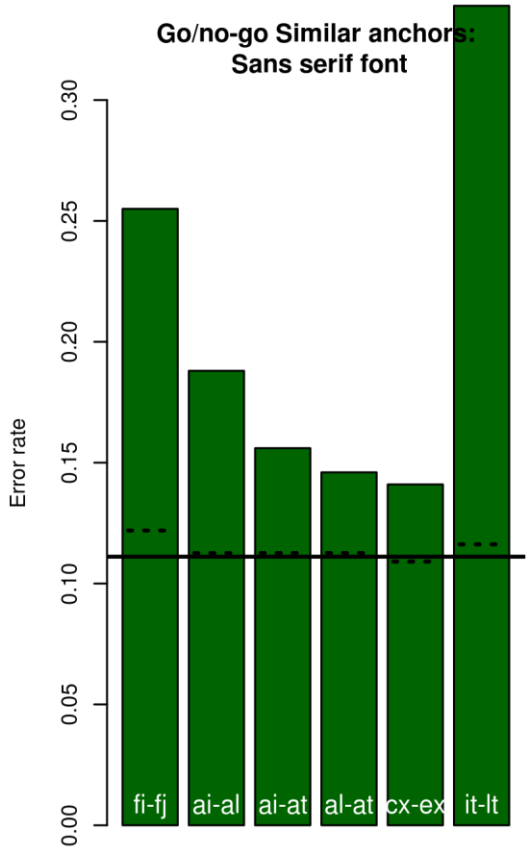
Candidate: EU in Greek. (epsilon upsilon)

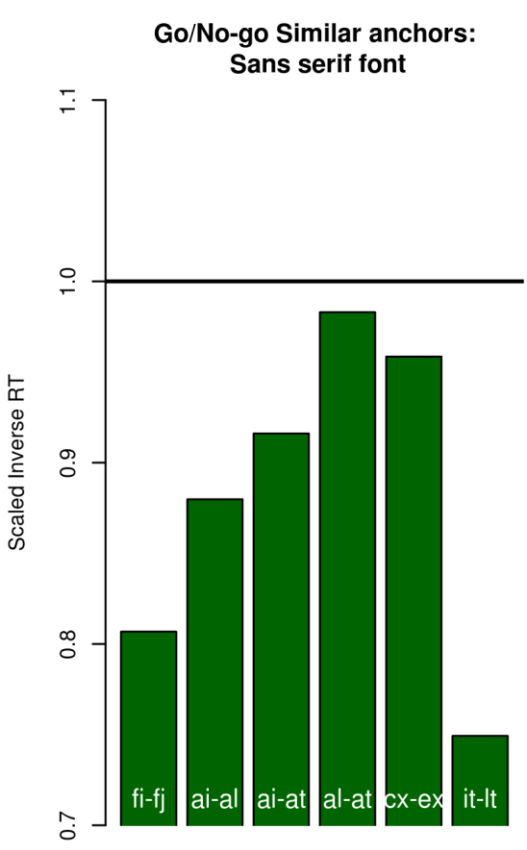


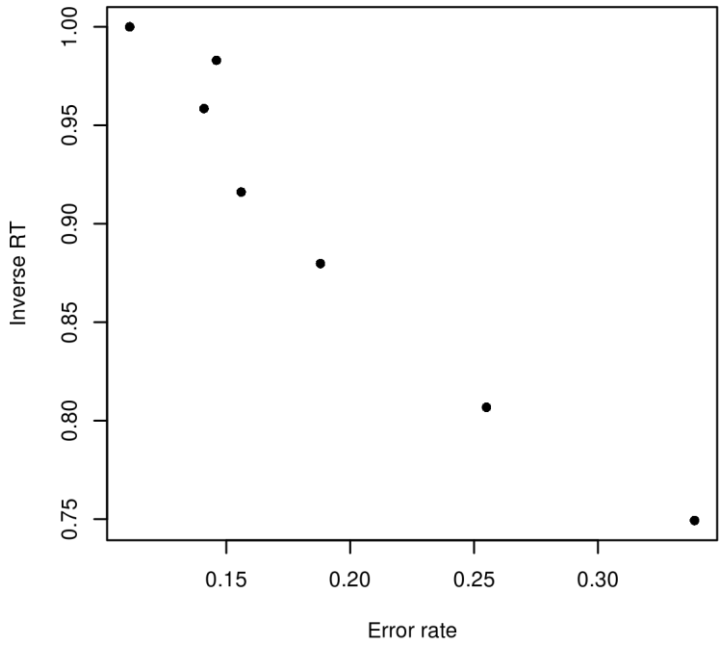
Candidate: EU in Greek. (epsilon upsilon)



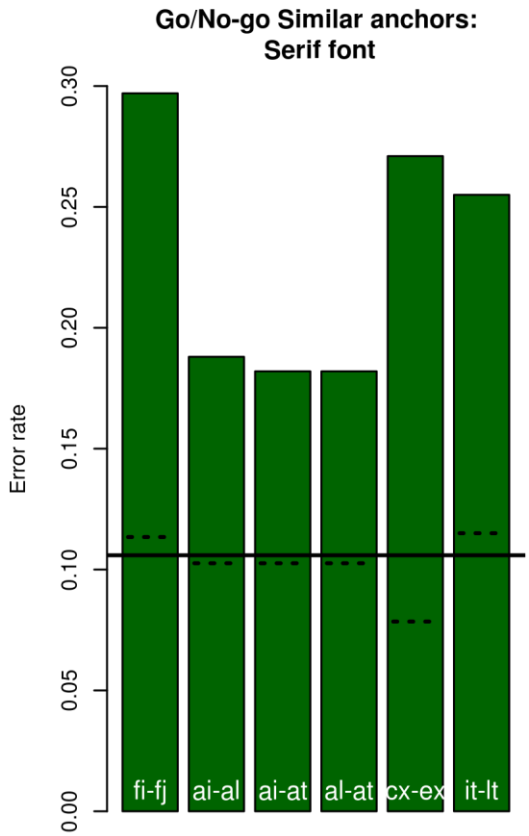
Correlation between error rate and inverse RT: -0.9281



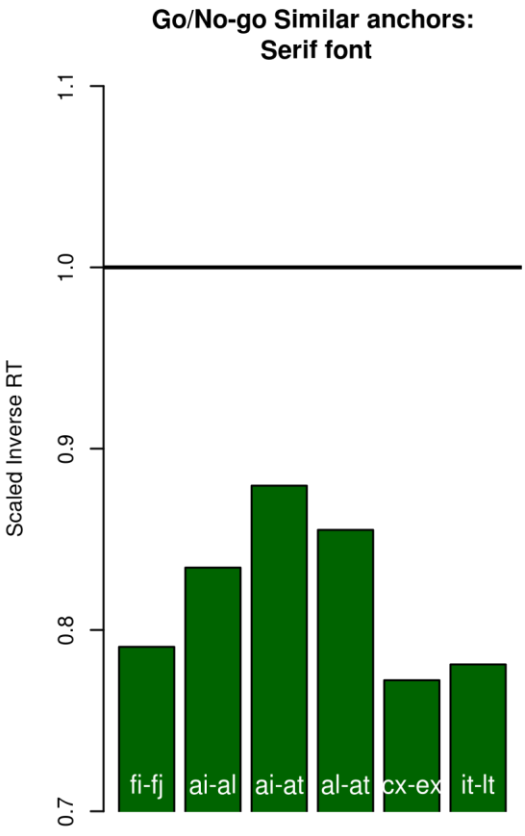




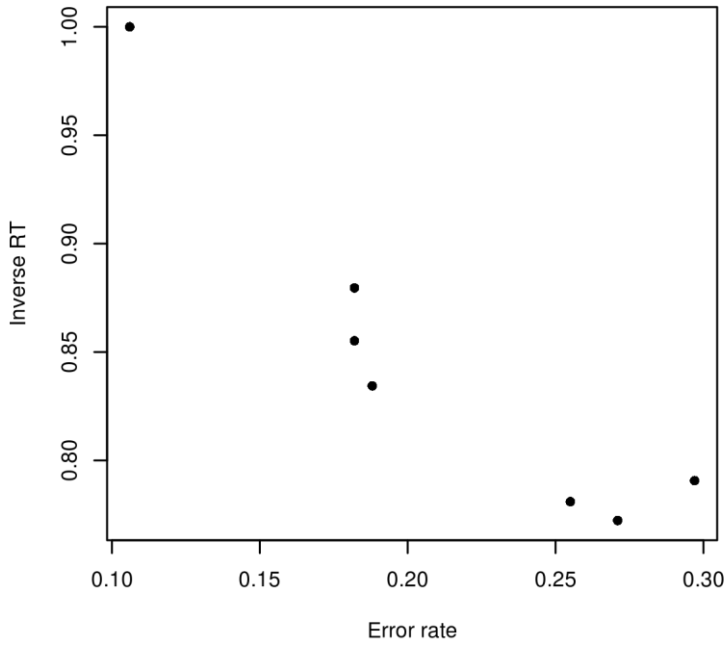
Correlation between error rate and inverse RT: -0.9716



Candidate: EU in Greek. (epsilon upsilon)



Candidate: EU in Greek. (epsilon upsilon)



Correlation between error rate and inverse RT: -0.9281