Still all Greeklish to me: Greeklish to Greek Transliteration

Anastasios Toumazatos, John Pavlopoulos,
Ion Androutsopoulos, Stavros Vassos

Greeklish

- Greek written in the Latin alphabet.
- Very prevalent in online informal settings.

Greeklish

Greeklish	Greek	Category
e	αι	phonetic
e	ε	key-sharing
ai	αι	optical
x	χ	optical
x	ξ	phonetic
ch	χ	phonetic
th	θ	phonetic
u	θ	key-sharing
8	θ	optical

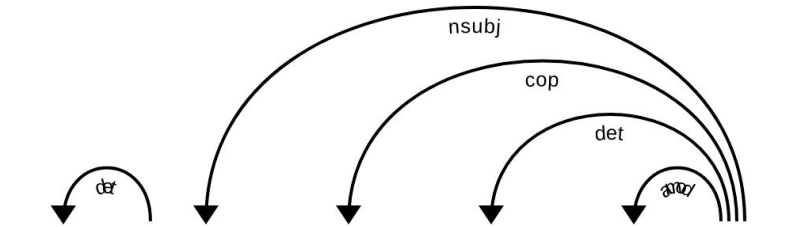
- Greek written in the Latin alphabet.
- Very prevalent in online informal settings.
- No standardized mapping between Greeklish and Greek.

Greeklish

- Lack of standardization makes Greeklish hard to handle in various applications.

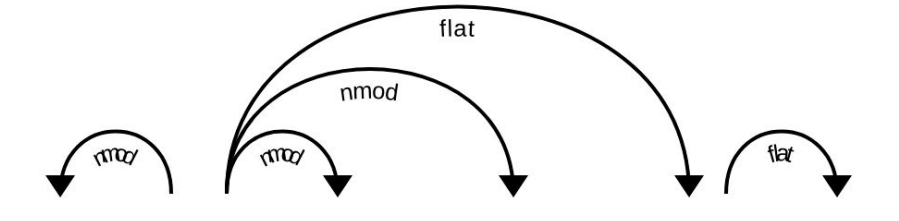
Greeklish

- Lack of standardization makes Greeklish hard to handle in various applications.
- Example: Dependency parsing with **spaCy** (same sentence in Greeklish and Greek).



(English: “Greek is a hard language.”)

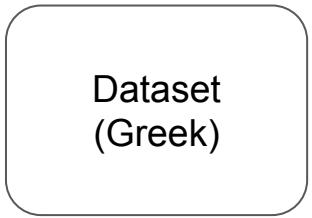
(a) Greek



(a) Greeklish

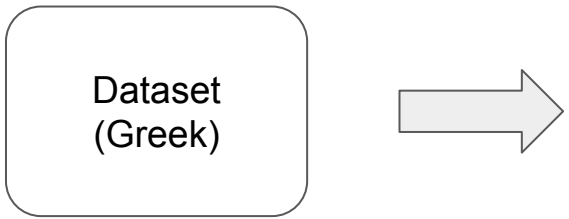
Data (synthetic)

- No existing datasets for Greeklish-to-Greek (G2G) transliteration.
- We created new synthetic datasets, by converting Greek text data to Greeklish using a mapping of all plausible transliterations of Greek letters and choosing randomly each time.
- Data from **Europarl (1996-2011) .**
- Subtitles from two TV Shows, **Para 5** and **Friends (1st season).**



Data (synthetic)

- No existing datasets for Greeklish-to-Greek (G2G) transliteration.
- We created new synthetic datasets, by converting Greek text data to Greeklish using a mapping of all plausible transliterations of Greek letters and choosing randomly each time.
- Data from **Europarl (1996-2011) .**
- Subtitles from two TV Shows, **Para 5** and **Friends (1st season).**

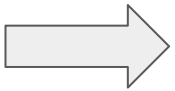


Greeklish	Greek	Greeklish	Greek
A	Α	N	Ν
Ai	Αι	Nt	Ντ
B	{Β, Μπ}	O	{Ο, Ω}
D	{Δ, Ντ}	Oi	Οι
E	{Ε, Αι}	Ou	Ου
Ei	Ει	P	{Π, Ψ}
F	Φ	Q	Θ
G	Γ	R	Ρ
H	{Η, Χ}	S	Σ
I	{Η, Ι, Υ, Ει, Οι, Υι}	T	Τ
K	Κ	Th	Θ
Ks	Ξ	U	{Θ, Ου, Υ}
L	Λ	V	Β
M	Μ	W	Ω
Mp	Μπ	X	{Ξ, Χ}
Y	Υ	o	{ο, ω}
Yi	Υι	oi	οι
Z	Ζ	ou	ου
a	α	p	π
ai	αι	ps	ψ
b	{β, μπ}	r	ρ
d	{δ, ντ}	s	{σ, ς}
e	{ε, αι}	t	τ
ei	ει	th	θ
f	φ	u	{υ, θ, ου}
g	γ	ui	υι
h	{η, χ}	v	β
i	{η, ι, υ, ει, οι, υι}	w	ω
k	κ	x	{ξ, χ}
ks	ξ	y	υ
l	λ	z	ζ
m	μ	n	ν
mp	μπ	nt	ντ
8	{Θ, θ}	3	{Ξ, ξ}

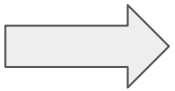
Data (synthetic)

- No existing datasets for Greeklish-to-Greek (G2G) transliteration.
- We created new synthetic datasets, by converting Greek text data to Greeklish using a mapping of all plausible transliterations of Greek letters and choosing randomly each time.
- Data from **Europarl (1996-2011) .**
- Subtitles from two TV Shows, **Para 5** and **Friends (1st season).**

Dataset (Greek)



Greeklish	Greek	Greeklish	Greek
A	Α	N	Ν
Ai	Αι	Nt	Ντ
B	{Β, Μπ}	O	{Ο, Ω}
D	{Δ, Ντ}	Oi	Οι
E	{Ε, Αι}	Ou	Ου
Ei	Ει	P	{Π, Ψ}
F	Φ	Q	Θ
G	Γ	R	Ρ
H	{Η, Χ}	S	Σ
I	{Η, Ι, Υ, Ει, Οι, Υι}	T	Τ
K	Κ	Th	Θ
Ks	Ξ	U	{Θ, Ου, Υ}
L	Λ	V	Β
M	Μ	W	Ω
Mp	Μπ	X	{Ξ, Χ}
Y	Υ	o	{ο, ω}
Yi	Υι	oi	οι
Z	Ζ	ou	ου
a	α	p	π
ai	αι	ps	ψ
b	{β, μπ}	r	ρ
d	{δ, ντ}	s	{σ, ς}
e	{ε, αι}	t	τ
ei	ει	th	θ
f	φ	u	{υ, θ, ου}
g	γ	ui	υι
h	{η, χ}	v	β
i	{η, ι, υ, ει, οι, υι}	w	ω
k	κ	x	{ξ, χ}
ks	ξ	y	υ
l	λ	z	ζ
m	μ	n	ν
mp	μπ	nt	ντ
8	{Θ, θ}	3	{Ξ, ξ}



Dataset (Greeklish)

Data (real-world)

- We also created smaller test datasets by manually transliterating real-world Greeklish content to Greek.
- Transcribed by MSc students trained for annotation.

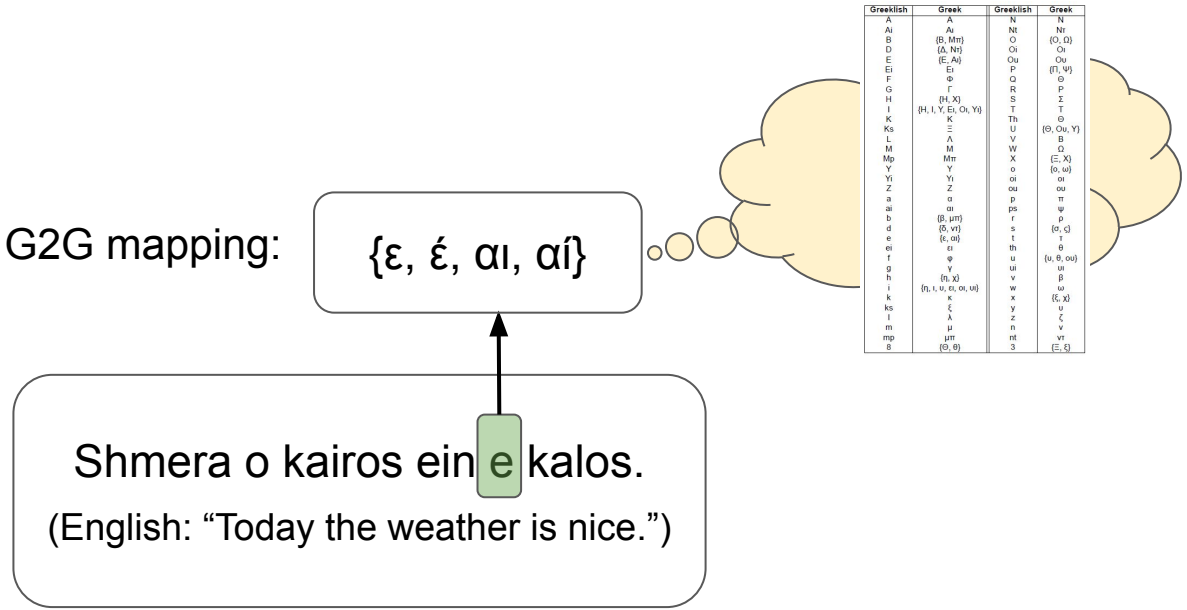
Name	**Description**
Survivorbot	Greeklish data from an online chatbot deployed during the run of a popular TV show in Greece (‘Survivor’).
Gazzetta	Greeklish data from the discussion forum of a popular sports website (‘Gazzetta’).

RBSLM

Shmera o kairos eine kalos.

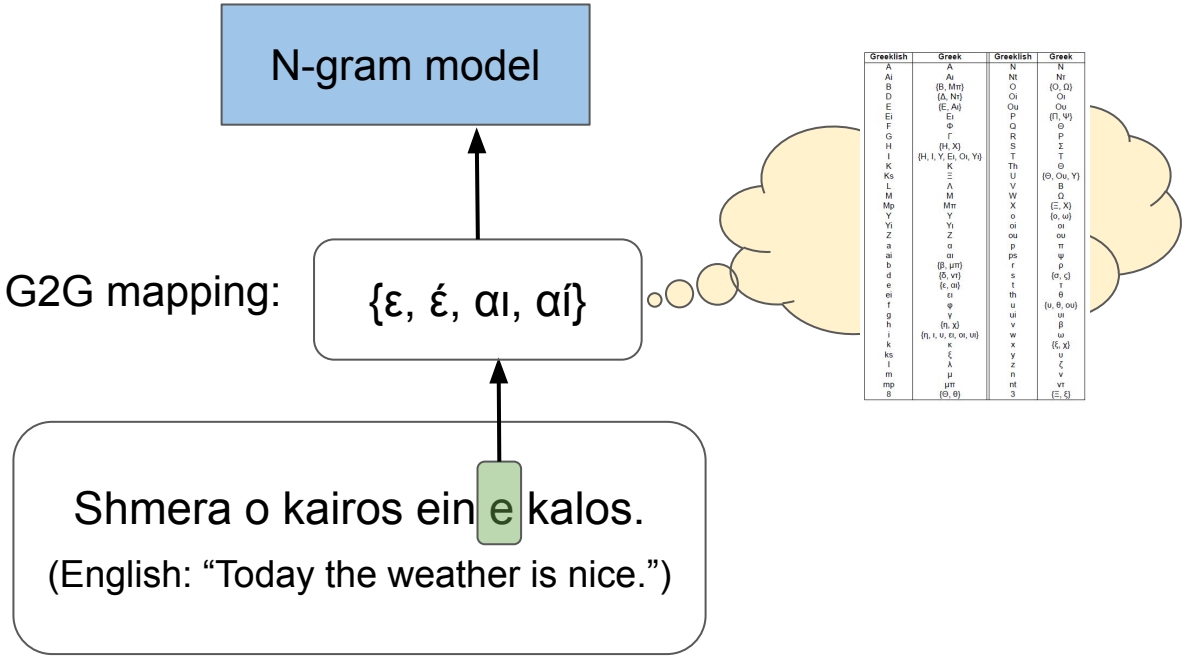
(English: “Today the weather is nice.”)

RBSLM

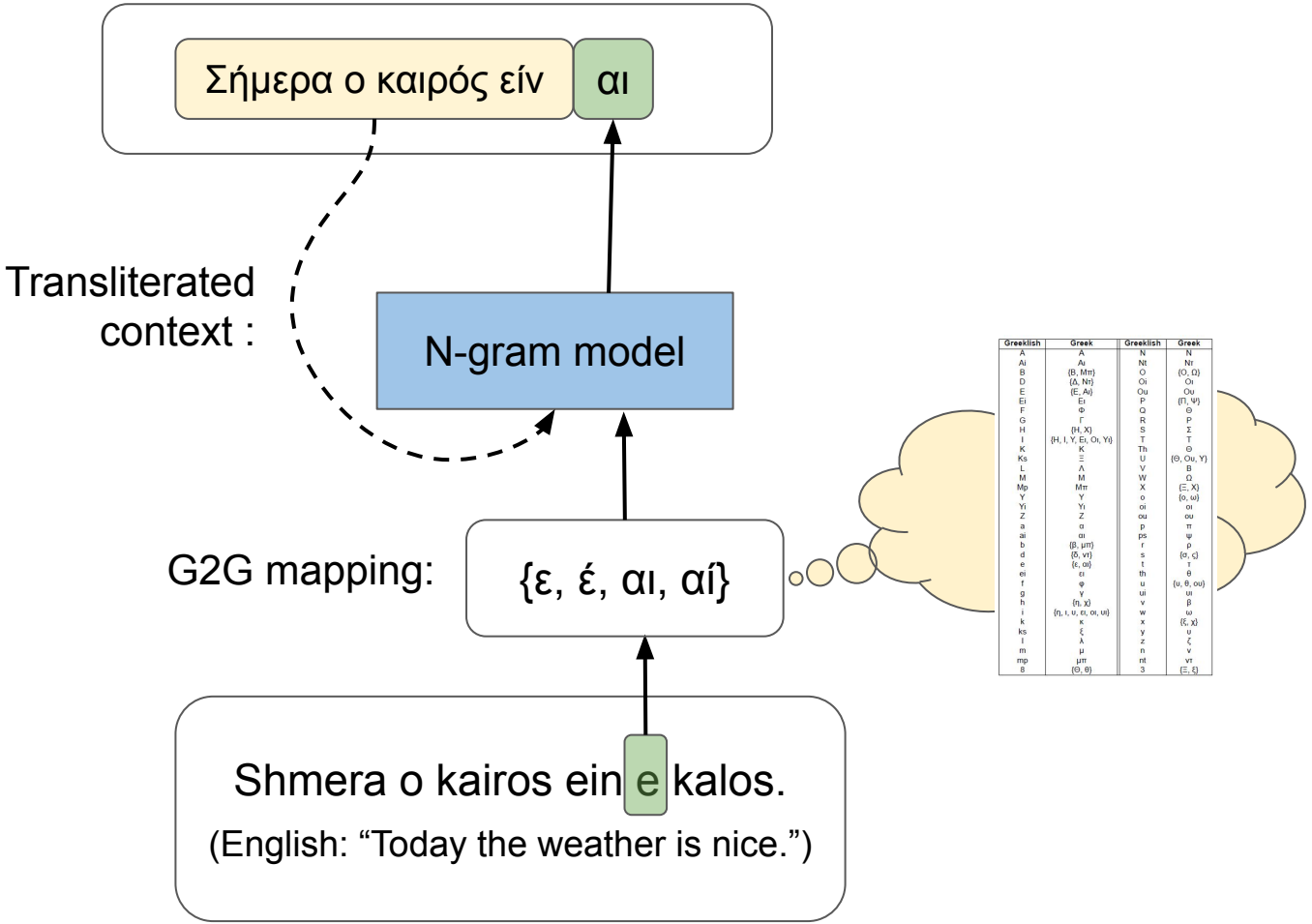


Greeklish	Greek	Greeklish	Greek
A	Α	N	Ν
Ai	Αι	Nt	Ντ
B	{Β, Μπ}	O	{Ο, Ω}
D	{Δ, Ντ}	Oi	Οι
E	{Ε, Αι}	Ou	Ου
Ei	Ει	P	{Π, Ψ}
F	Φ	Q	Θ
G	Γ	R	Ρ
H	{Η, Χ}	S	Σ
I	{Η, Ι, Υ, Ει, Οι, Υι}	T	Τ
K	Κ	Th	Θ
Ks	Ξ	U	{Θ, Ου, Υ}
L	Λ	V	Β
M	Μ	W	Ω
Mp	Μπ	X	{Ξ, Χ}
Y	Υ	o	{ο, ω}
Yi	Υι	oi	οι
Z	Ζ	ou	ου
a	α	p	π
ai	αι	ps	ψ
b	{β, μπ}	r	ρ
d	{δ, ντ}	s	{σ, ς}
e	{ε, αι}	t	τ
ei	ει	th	θ
f	φ	u	{υ, θ, ου}
g	γ	ui	υι
h	{η, χ}	v	β
i	{η, ι, υ, ει, οι, υι}	w	ω
k	κ	x	{ξ, χ}
ks	ξ	y	υ
l	λ	z	ζ
m	μ	n	ν
mp	μπ	nt	ντ
8	{Θ, θ}	3	{Ξ, ξ}

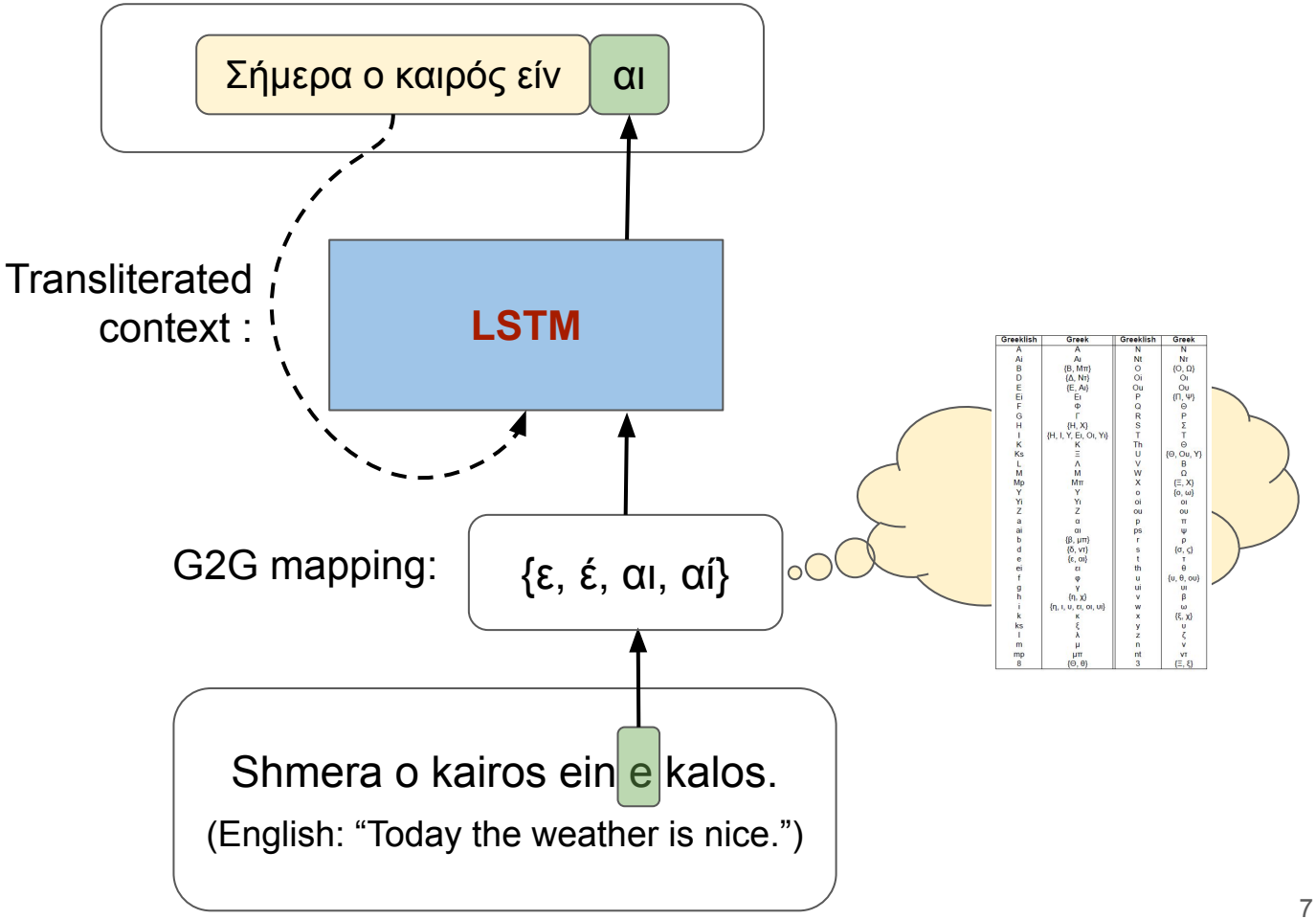
RBSLM


N-gram model
G2G mapping:
{ε, έ, αι, αί}
Shmera o kairos ein e kalos.
(English: “Today the weather is nice.”)

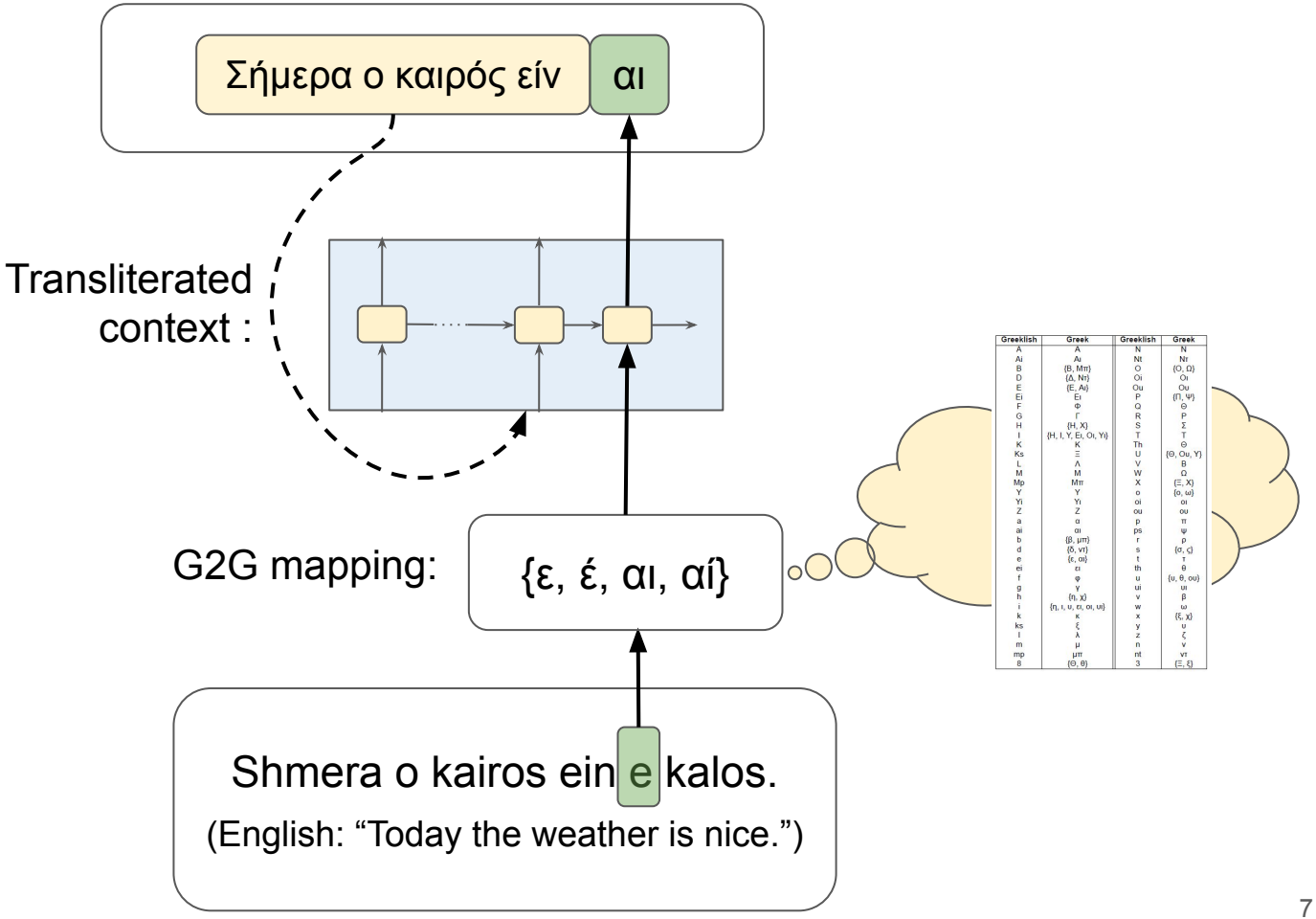
RBSLM


Σήμερα ο καιρός είν
αι
Transliterated context :
N-gram model
G2G mapping:
{ε, έ, αι, αί}
Shmera o kairos ein e kalos.
(English: “Today the weather is nice.”)

RBNLM


Σήμερα ο καιρός είν
αι
Transliterated context :
LSTM
G2G mapping:
{ε, έ, αι, αί}
Shmera o kairos ein e kalos.
(English: “Today the weather is nice.”)

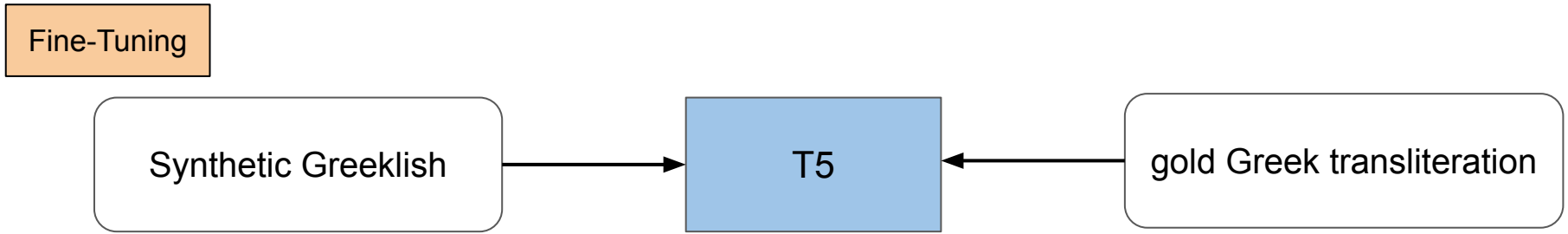
RBNLM



Greeklish	Greek	Greeklish	Greek
A	A	N	N
Ai	Αι	Nt	Ντ
B	{Β, Μπ}	O	{Ο, Ω}
D	{Δ, Ντ}	Oi	Οι
E	{Ε, Αι}	Ou	Ου
Ei	Ει	P	{Π, Ψ}
F	Φ	Q	Θ
G	Γ	R	Ρ
H	{Η, Χ}	S	Σ
I	{Η, Ι, Υ, Ει, Οι, Υι}	T	Τ
K	Κ	Th	Θ
Ks	Ξ	U	{Θ, Ου, Υ}
L	Λ	V	Β
M	Μ	W	Ω
Mp	Μπ	X	{Ξ, Χ}
Y	Υ	o	{ο, ω}
Yi	Υι	oi	οι
Z	Ζ	ou	ου
a	α	p	π
ai	αι	ps	ψ
b	{β, μπ}	r	ρ
d	{δ, ντ}	s	{σ, ς}
e	{ε, αι}	t	τ
ei	ει	th	θ
f	φ	u	{υ, θ, ου}
g	γ	ui	υι
h	{η, χ}	v	β
i	{η, ι, υ, ει, οι, υι}	w	ω
k	κ	x	{ξ, χ}
ks	ξ	y	υ
l	λ	z	ζ
m	μ	n	ν
mp	μπ	nt	ντ
n	{Θ, θ}	3	{Ξ, ξ}

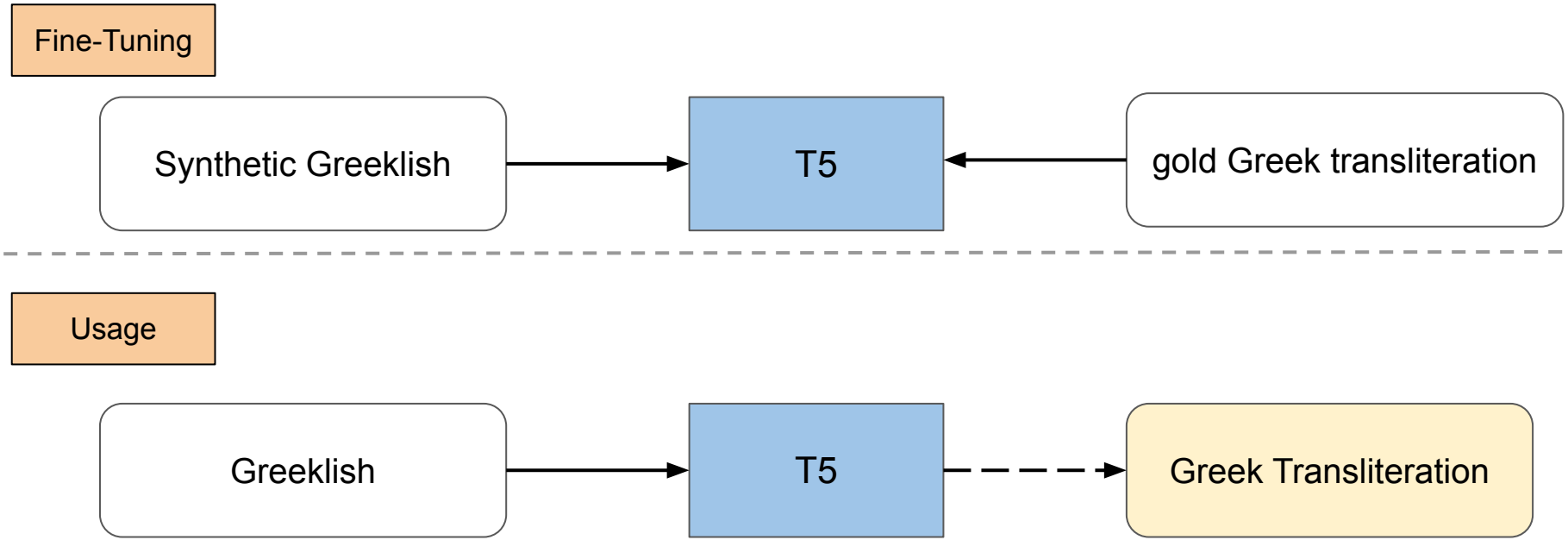
T5-based

- We used Google's **mT5** & **ByT5** models.
- Both are pre-trained on the same multilingual dataset.
- We further fine tune them on data from our *synthetic* datasets.



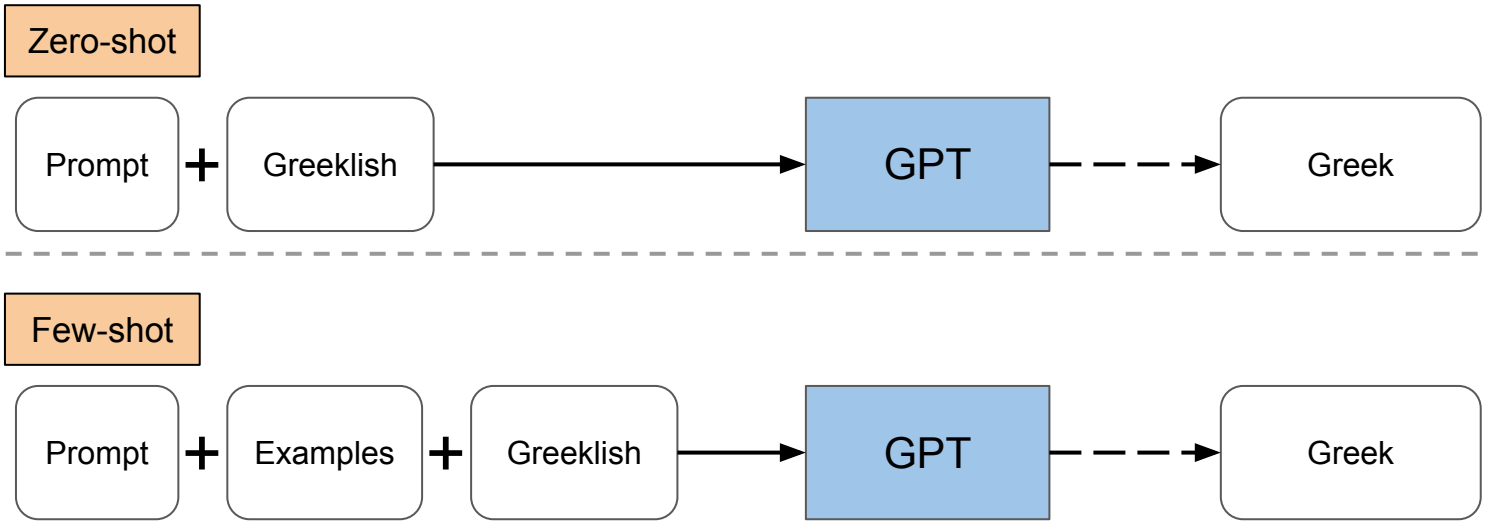
T5-based

- We used Google's **mT5** & **ByT5** models.
- Both are pre-trained on the same multilingual dataset.
- We further fine tune them on data from our *synthetic* datasets.



GPT-based

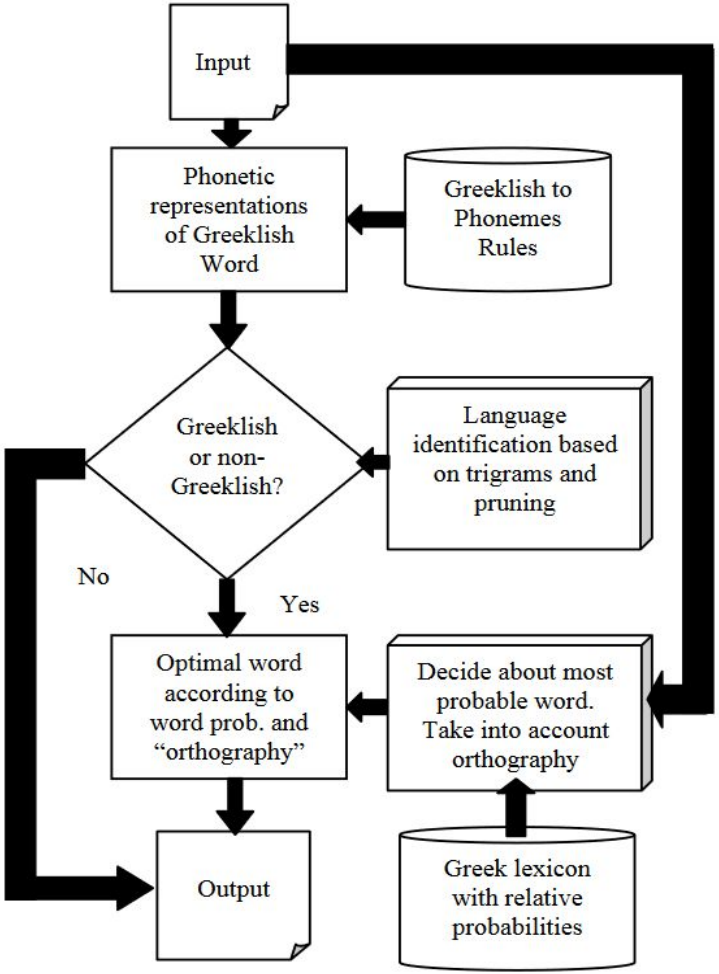
- We used OpenAI's **GPT-3.5-t** and **GPT-4***, also experimenting with the web version of ChatGPT.
- We investigated various prompts, as well as zero-shot and few-shot prompting. We find that few-shot prompting is helpful regardless of the relevance of topic (e.g. Europarl examples for transliterating a sports-related sentence).



*Exact versions used were **gpt-3.5-turbo-0314** and **gpt-4-0314.**

All Greek to me!

- Proposed by Chalamandaris et al. (2006)
- Trained on real-world web-collected data.
- Based on a statistical tri-gram model & lexicons from large corpora.
- Figure of the architecture used from the original paper by Chalamandaris et al. (https://aclanthology.org/L06-1229/)



Results (synthetic data)

Model	Europarl		Friends (TV)		Para5 (TV)	
	CER	WER	CER	WER	CER	WER
AllGreek	2.98	11.54	7.56	19.62	5.32	16.32
RBSLM	4.71	15.92	14.29	39.67	12.98	36.53
RBNLM	1.34	4.45	9.68	27.03	7.36	20.55
mT5	6.06	10.14	32.60	48.53	31.56	43.86
ByT5	1.30	4.11	14.36	38.20	13.93	34.28
ByT5-EU	0.64	2.30	11.36	29.74	9.80	25.04
ByT5-TV	1.37	3.92	3.77	11.69	3.51	9.93
GPT-4-6shot	2.27	5.21	9.74	21.41	8.34	25.15

Results (synthetic data)

mT5 performs worst overall, much lower than ByT5.

Model	Europarl		Friends (TV)		Para5 (TV)	
	CER	WER	CER	WER	CER	WER
AllGreek	2.98	11.54	7.56	19.62	5.32	16.32
RBSLM	4.71	15.92	14.29	39.67	12.98	36.53
RBNLM	1.34	4.45	9.68	27.03	7.36	20.55
mT5	6.06	10.14	32.60	48.53	31.56	43.86
ByT5	1.30	4.11	14.36	38.20	13.93	34.28
ByT5-EU	0.64	2.30	11.36	29.74	9.80	25.04
ByT5-TV	1.37	3.92	3.77	11.69	3.51	9.93
GPT-4-6shot	2.27	5.21	9.74	21.41	8.34	25.15

Results (synthetic data)

mT5 performs worst overall, much lower than ByT5.

ByT5-EU is the best model in its specific domain.

Model	Europarl		Friends (TV)		Para5 (TV)	
	CER	WER	CER	WER	CER	WER
AllGreek	2.98	11.54	7.56	19.62	5.32	16.32
RBSLM	4.71	15.92	14.29	39.67	12.98	36.53
RBNLM	1.34	4.45	9.68	27.03	7.36	20.55
mT5	6.06	10.14	32.60	48.53	31.56	43.86
ByT5	1.30	4.11	14.36	38.20	13.93	34.28
ByT5-EU	**0.64**	**2.30**	11.36	29.74	9.80	25.04
ByT TV	1.37	3.92	3.77	11.69	3.51	9.93
T-4-6shot	2.27	5.21	9.74	21.41	8.34	25.15

Results (synthetic data)

Model	Europarl		Friends (TV)		Para5 (TV)	
	CER	WER	CER	WER	CER	WER
AllGreek	2.98	11.54	7.56	19.62	5.32	16.32
RBSLM	4.71	15.92	14.29	39.67	12.98	36.53
RBNLM	1.34	4.45	9.68	27.03	7.36	20.55
mT5	6.06	10.14	32.60	48.53	31.56	43.86
ByT5	1.30	4.11	14.36	38.20	13.93	34.28
ByT5-EU	**0.64**	**2.30**	11.36	29.74	9.80	25.04
ByT TV	1.37	3.92	**3.77**	**11.69**	**3.51**	**9.93**
T-4-6shot	2.27	5.21	9.74	21.41	8.34	25.15

mT5 performs worst overall, much lower than ByT5.

ByT5-EU is the best model in its specific domain.

ByT5-TV is also the best in its domain.

Results (real-world)

Model	Survivorbot		Gazzetta (accepted)		Gazzetta (rejected)	
	CER	WER	CER	WER	CER	WER
AllGreek	14.66	30.48	9.71	24.54	11.17	28.94
RBSLM	22.70	55.16	18.55	43.01	20.70	43.07
RBNLM	19.99	50.21	14.59	33.10	17.41	35.42
mT5	40.99	59.67	30.02	43.70	39.26	50.03
ByT5	29.45	59.89	18.23	35.01	27.63	42.49
ByT5-EU	22.96	51.18	15.57	32.08	24.65	37.70
ByT5-TV	17.70	39.78	13.19	31.11	22.30	35.63
ByT5-90shot	16.41	38.22	11.17	25.98	17.69	29.61
GPT-4-6shot	9.44	22.74	8.02	18.36	10.80	21.76

Results (real-world)

Model	Survivorbot		Gazzetta (accepted)		Gazzetta (rejected)	
	CER	WER	CER	WER	CER	WER
AllGreek	14.66	30.48	9.71	24.54	11.17	28.94
RBSLM	22.70	55.16	18.55	43.01	20.70	43.07
RBNLM	19.99	50.21	14.59	33.10	17.41	35.42
mT5	40.99	59.67	30.02	43.70	39.26	50.03
ByT5	29.45	59.89	18.23	35.01	27.63	42.49
ByT5-EU	22.96	51.18	15.57	32.08	24.65	37.70
ByT5-TV	17.70	39.78	13.19	31.11	22.30	35.63
ByT5-90shot	16.41	38.22	11.17	25.98	17.69	29.61
GPT-4-6shot	**9.44**	**22.74**	**8.02**	**18.36**	**10.80**	**21.76**

GPT-4 is the best performing model in all datasets.

Results (real-world)

Model	Survivorbot		Gazzetta (accepted)		Gazzetta (rejected)	
	CER	WER	CER	WER	CER	WER
AllGreek	14.66	30.48	9.71	24.54	11.17	28.94
RBSLM	22.70	55.16	18.55	43.01	20.70	43.07
RBNLM	19.99	50.21	14.59	33.10	17.41	35.42
mT5	40.99	59.67	30.02	43.70	39.26	50.03
ByT5	29.45	59.89	18.23	35.01	27.63	42.49
ByT5-EU	22.96	51.18	15.57	32.08	24.65	37.70
ByT5-TV	17.70	39.78	13.19	31.11	22.30	35.63
ByT5-90shot	16.41	38.22	11.17	25.98	17.69	29.61
GPT-4-6shot	**9.44**	**22.74**	**8.02**	**18.36**	**10.80**	**21.76**

GPT-4 is the best performing model in all datasets.

Fine-tuning on limited real world data drastically improves ByT5.

Results (real-world)

Model	Survivorbot		Gazzetta (accepted)		Gazzetta (rejected)	
	CER	WER	CER	WER	CER	WER
AllGreek	14.66	30.48	9.71	24.54	11.17	28.94
RBSLM	22.70	55.16	18.55	43.01	20.70	43.07
RBNLM	19.99	50.21	14.59	33.10	17.41	35.42
mT5	40.99	59.67	30.02	43.70	39.26	50.03
ByT5	29.45	59.89	18.23	35.01	27.63	42.49
ByT5-EU	22.96	51.18	15.57	32.08	24.65	37.70
ByT5-TV	17.70	39.78	13.19	31.11	22.30	35.63
ByT5-90shot	16.41	38.22	11.17	25.98	17.69	29.61
GPT-4-6shot	**9.44**	**22.74**	**8.02**	**18.36**	**10.80**	**21.76**

AllGreek is the second best model in all real-world experiments.

GPT-4 is the best performing model in all datasets.

Fine-tuning on limited real world data drastically improves ByT5.

Results (real-world)

Observation: more real-world data (regardless of source) leads to better performance.

Model	Survivorbot		Gazzetta (accepted)		Gazzetta (rejected)	
	CER	WER	CER	WER	CER	WER
AllGreek	14.66	30.48	9.71	24.54	11.17	28.94
RBSLM	22.70	55.16	18.55	43.01	20.70	43.07
RBNLM	19.99	50.21	14.59	33.10	17.41	35.42
mT5	40.99	59.67	30.02	43.70	39.26	50.03
ByT5	29.45	59.89	18.23	35.01	27.63	42.49
ByT5-EU	22.96	51.18	15.57	32.08	24.65	37.70
ByT5-TV	17.70	39.78	13.19	31.11	22.30	35.63
ByT5-90shot	16.41	38.22	11.17	25.98	17.69	29.61
GPT-4-6shot	**9.44**	**22.74**	**8.02**	**18.36**	**10.80**	**21.76**

AllGreek is the second best model in all real-world experiments.

GPT-4 is the best performing model in all datasets.

Fine-tuning on limited real world data drastically improves ByT5.

Error Analysis

other-language characters	GREEKLISH	gia na to doume to allani to keno mpogiobits an kalucei
	GOLD	Για να το δούμε το αλάνι το κενό Bojovic αν καλύψει
	MT5	για να το δούμε το άλνενο το κένο μπορντोरअईΣ αν καλύπτει
incorrect spelling	GREEKLISH	to gipedo tis aek to eidane? :P
	GOLD	Το γήπεδο της ΑΕΚ το είδανε; :P
	BYT5	Το γυπέδο της αεκ το είδανε? :P
	MT5	Το γύρω πόλων της άμες το έθατε?
non-tranlsiterated words	GREEKLISH	ginei ke files, boris na mu peis;
	GOLD	γίνει και φίλες, μπορείς να μου πεις;
	ALLGREEK	γίνει και φίλες, boris να μου πεις;

Error Analysis

other-language characters	GREEKLISH	gia na to doume to allani to keno mpogiobits an kalucei
	GOLD	Για να το δούμε το αλάνι το κενό Bojovic αν καλύψει
	MT5	για να το δούμε το άλνενο το κένο μπορντ[illegible] αν καλύπτει
incorrect spelling	GREEKLISH	to gipedo tis aek to eidane? :P
	GOLD	Το γήπεδο της ΑΕΚ το είδανε; :P
	BYT5	Το γυπέδο της αεκ το είδανε? :P
	MT5	Το γύρω πόλων της άμες το έθατε?
non-tranlsiterated words	GREEKLISH	ginei ke files, boris na mu peis;
	GOLD	γίνει και φίλες, μπορείς να μου πεις;
	ALLGREEK	γίνει και φίλες, boris να μου πεις;

mT5's multilingual sub-word tokens might not be optimal for Greeklish.

Error Analysis

other-language characters	GREEKLISH	gia na to doume to allani to keno mpogiobits an kalucei
	GOLD	Για να το δούμε το αλάνι το κενό Bojovic αν καλύψει
	MT5	για να το δούμε το άλνενο το κένο μπορντ[illegible] αν καλύπτει
incorrect spelling	GREEKLISH	to gipedo tis aek to eidane? :P
	GOLD	Το γήπεδο της ΑΕΚ το είδανε; :P
	BYT5	Το γυπέδο της αεκ το είδανε? :P
	MT5	Το γύρω πόλων της άμες το έθατε?
non-tranlsiterated words	GREEKLISH	ginei ke files, boris na mu peis;
	GOLD	γίνει και φίλες, μπορείς να μου πεις;
	ALLGREEK	γίνει και φίλες, boris να μου πεις;

When AllGreek comes across input that it struggles with, it often skips transliterating it (maybe identifying it as code-switching).

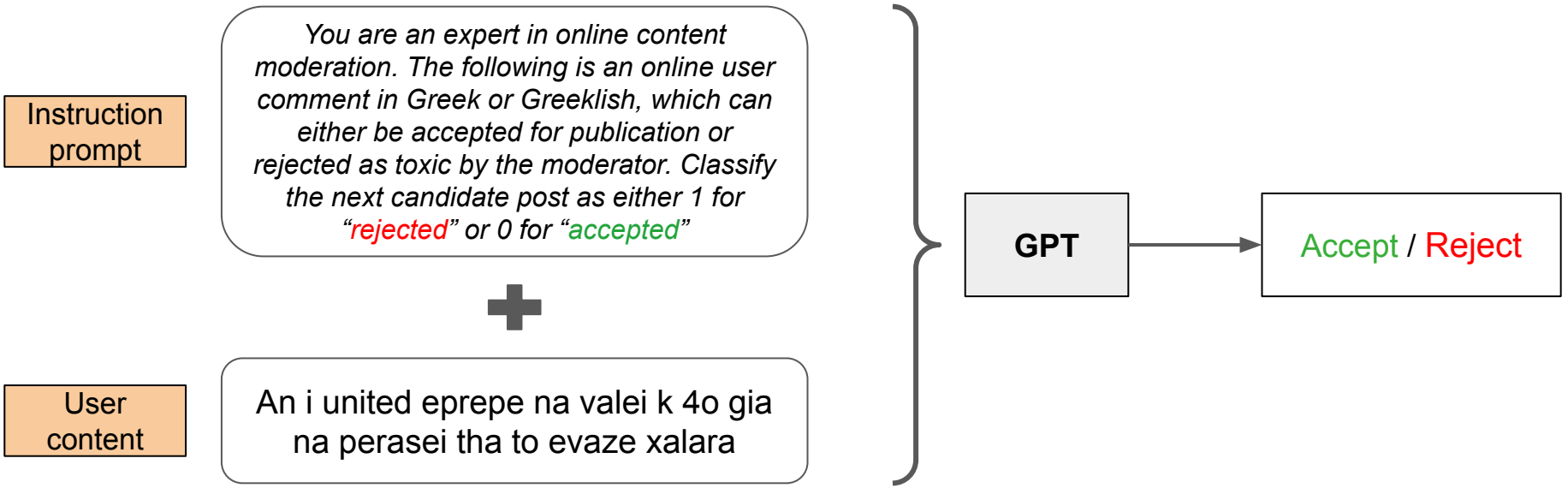
mT5's multilingual sub-word tokens might not be optimal for Greeklish.

Terms altered by GPT-based

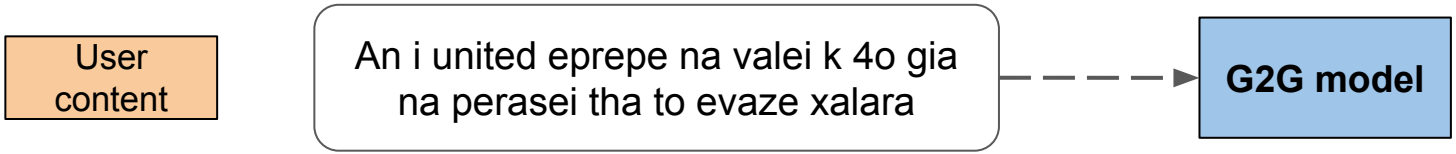
Greeklish (SOURCE)	**Greek** (TRANSLITERATED)
An i united eprepe na valei k 4o gia na perasei tha to evaze xalara	Αν η United έπρεπε να βάλει **και** 4ο για να περάσει, θα το έβαζε άνετα
Τα topika einai ola xera kai giafto den paei kanena paidaki na spasi ta podia tou	Τα τοπικά είναι όλα ξερά και γι'αυτό δεν πάει κανένα παιδάκι να σπάσει τα πόδια του
Na afairethoun t vraveia apo ton berg ton anastasiou kai ton risvani na t dwsoun ston mitroglou p einai stin agglia ston mitsel pou pire prwtathlima apo ton au-gousto p perase kai kalutero neo paixth na to dwsoun ston vergo	Να αφαιρεθούν τα βραβεία από τον Βεργκ τον Αναστασίου και τον Ρισβάνη να τα δώσουν στον Μήτρογλου που είναι στην Αγγλία στον Μίτσελ που πήρε πρωτάθλημα από τον Αύγουστο που πέρασε και για καλύτερο νέο παίχτη να το δώσουν στον Βέργο

- GPT-based models try to clarify abbreviations, short forms, and named entities.
- Usually leads to less noisy & better transliteration.

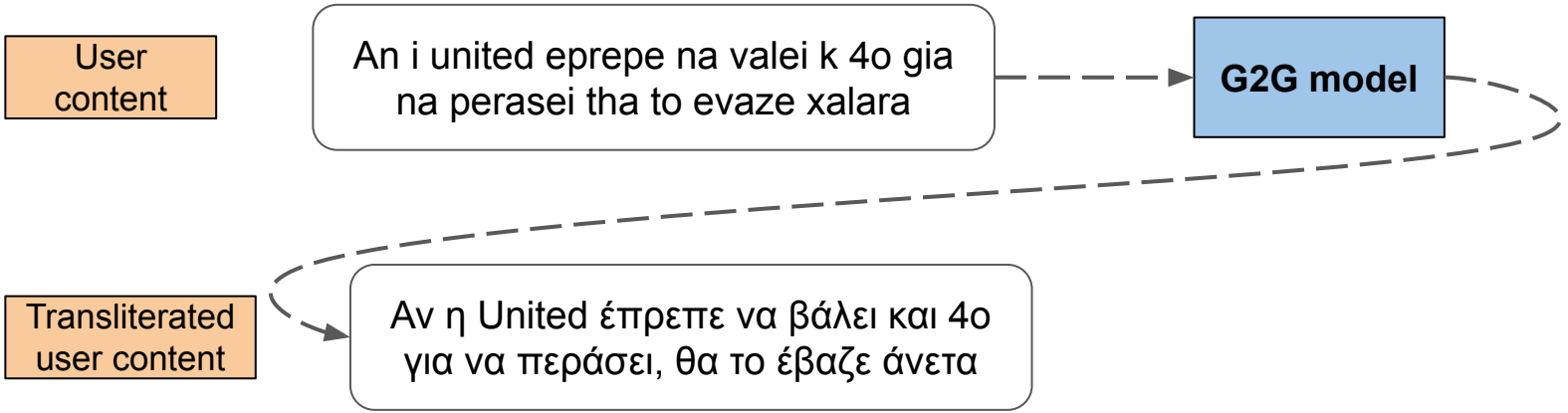
Downstream Task: Content Moderation



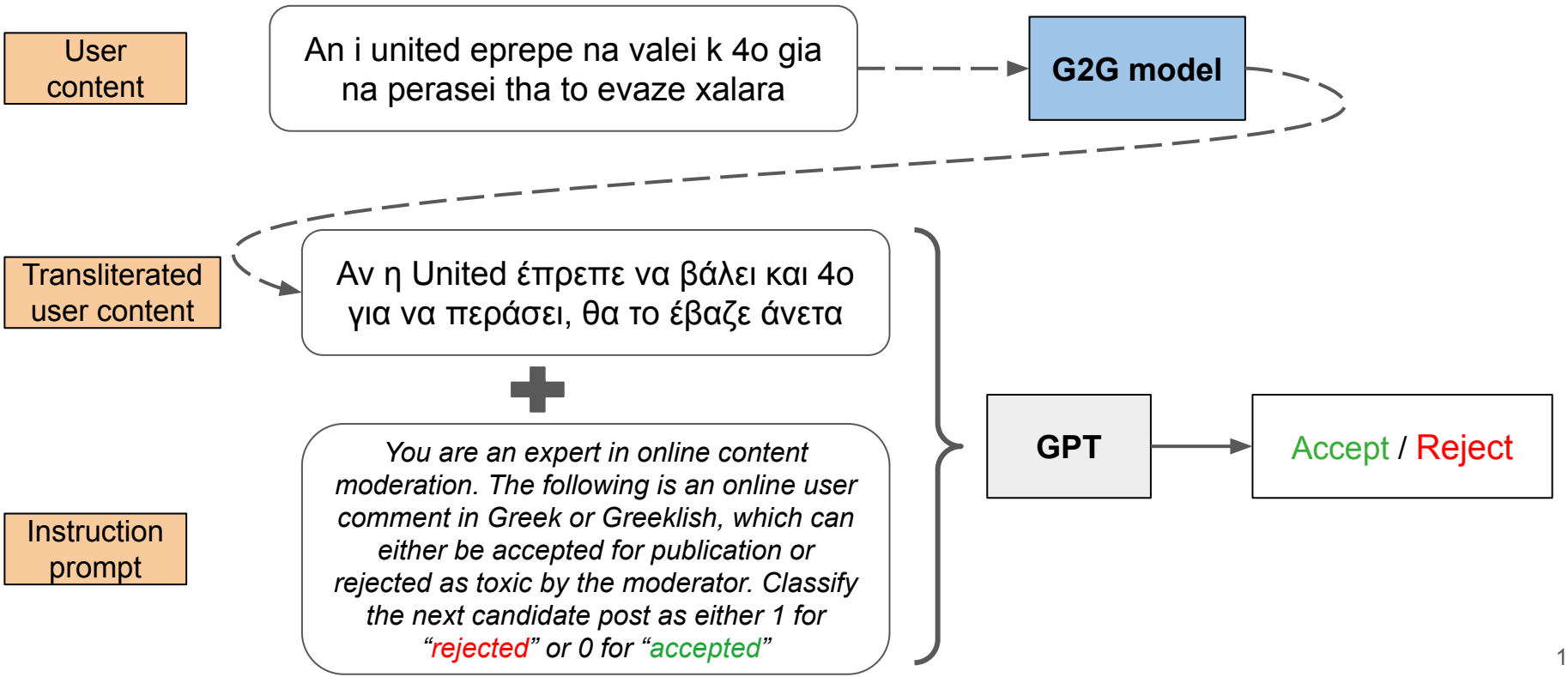
Downstream Task: Content Moderation (with transliteration)



Downstream Task: Content Moderation (with transliteration)



Downstream Task: Content Moderation (with transliteration)



Downstream Task: Content Moderation

	Greeklish		Transliterated	
Accuracy	0.49		**0.53**	
F_2 ($P_{rejected}$, $P_{accepted}$)	0.54		**0.69**	
	Precision	**Recall**	**Precision**	**Recall**
Accepted	0.47	0.71	**0.50**	**0.89**
Rejected	0.54	**0.30**	**0.69**	0.21
Average	0.51	0.51	**0.59**	**0.55**

- On average, adding a transliteration step helps (even if it's not perfect).
- No change needed in the model of the downstream task!
- We use F_2 as defined by Pavlopoulos et al. (2017) (https://aclanthology.org/D17-1117.pdf)

Conclusions

- We evaluated various approaches for G2G transliteration, featuring approaches based on:
  - Transliteration Rules + {Statistical / Deep Learning} LM
  - Fine-tuned multilingual transformers
  - A previous system based on statistical machine learning
  - Zero/Zero/Few-shot prompting LLMs (GPT-3.4, GPT-4)

Conclusions

- We evaluated various approaches for G2G transliteration, featuring approaches based on:
  - Transliteration Rules + {Statistical / Deep Learning} LM
  - Fine-tuned multilingual transformers
  - A previous system based on statistical machine learning
  - Zero/Zero/Few-shot prompting LLMs (GPT-3.4, GPT-4)

- **We conclude that:**

Conclusions

- We evaluated various approaches for G2G transliteration, featuring approaches based on:
  - Transliteration Rules + {Statistical / Deep Learning} LM
  - Fine-tuned multilingual transformers
  - A previous system based on statistical machine learning
  - Zero/Few-shot prompting LLMs (GPT-3.4, GPT-4)

- **We conclude that:**
- Zero/Few-shot prompting LLMs yields the best results.
  ..but..
- An old system based of statistical LM can be competitive, often outperforming newer approaches and techniques.

- Greeklish-to-Greek transliteration is far from solved.
  ..but..
- Even non-perfect transliteration can help significantly in downstream tasks (experimental evidence from content moderation).

Thank you!

https://github.com/nlpaueb/greeklish

@AUEBNLPGroup

{toumazatos,ion,annis}@aueb.gr