Filling the Gap: Decoding of Word Embeddings for Generation of Coherent New Words

Safa AlSaidi, Amandine Decker, Stephanie Monteiro

M2 – Software Project

1 Reminder

2 State of the project

3 Results

4 Software

Reminder of our aim

- Regression task based on transfer

$$A : B :: C : X \xrightarrow{X=?} A : B :: C : D$$
$$e.g.\ \ dog : dogs :: chat : X \rightarrow chats$$

- Input: A and B in language 1, C in language 2
- Output: D in language 2
- Same transformation for A, B and C, D

1 Reminder

2 State of the project

3 Results

4 Software

Since the last time ...

1. Run the experiments again
   - More runs
   - Typo in our code: two regression models now
2. Improve our software
3. Write the report

Analogy solver model

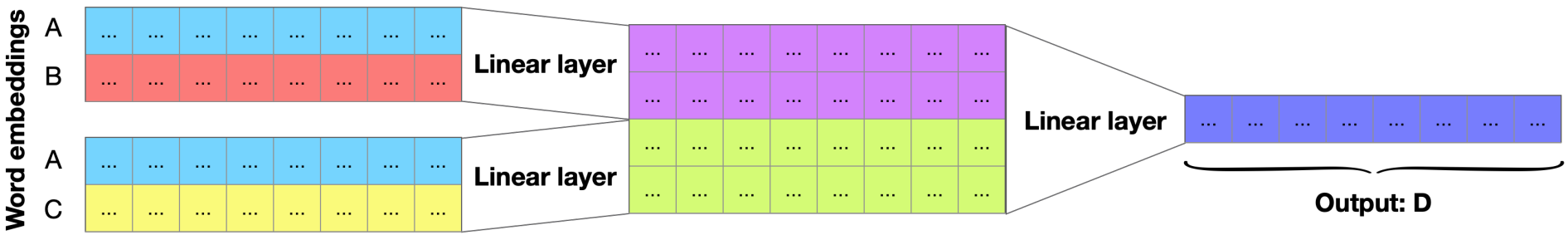

Figure: Analogy solver model

1 Reminder

2 State of the project

3 Results

4 Software

Regression model

Language	**ANNr (previous)** (mean ± std.)	**old model** (mean ± std.)	**new model** (mean ± std.)
Arabic	77.97 ± 16.03	59.14 ± 1.76	61.13 ± 0.83
Finnish	37.78 ± 9.28	76.61 ± 1.15	76.46 ± 1.58
Georgian	94.66 ± 1.13	85.51 ± 2.00	84.67 ± 2.78
German	86.38 ± 0.45	89.26 ± 0.51	88.70 ± 0.58
Hungarian	53.83 ± 3.12	78.49 ± 0.65	78.72 ± 0.53
Maltese	75.00 ± 5.08	77.37 ± 2.27	78.04 ± 1.44
Navajo	31.74 ± 0.90	46.14 ± 0.54	45.74 ± 0.99
Russian	75.15 ± 0.44	72.51 ± 0.46	72.23 ± 0.44
Spanish	86.27 ± 0.71	91.18 ± 0.51	91.72 ± 0.43
Turkish	61.95 ± 10.86	80.37 ± 0.82	80.37 ± 1.00
Japanese	61.60 ± 1.33	74.75 ± 1.33	72.58 ± 2.47

Table: Accuracy (in %) of 10 runs of the regression models (3 runs for previous results).

Omnilingual model

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	61.85±1.79	3.22±1.73	28.58±2.69	/	55.23±3.85
German	3.15±1.47	64.38±1.27	**66.91±4.49**	62.07±3.58	/
Hungarian	36.26±4.52	55.25±1.47	**73.33±1.31**	**78.36±1.33**	**32.00±3.03**
Spanish	/	61.16±2.54	**74.05±1.77**	69.38±1.65	70.67±4.03
Turkish	54.12±1.48	/	25.38±3.94	65.72±6.09	**52.23±1.09**

Table: Accuracy (in %) of 10 runs of the new omnilingual regression model

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	60.95±1.75	3.34±1.16	30.34±2.02	/	53.82±2.74
German	3.05±1.43	63.78±0.88	60.99±4.86	62.65±3.47	/
Hungarian	35.06±4.49	56.35±1.96	71.69±1.84	70.61±7.90	27.10±3.17
Spanish	/	61.90±2.42	67.32±5.05	67.90±1.97	66.12±6.86
Turkish	54.00±2.41	/	24.19±2.21	64.94±5.61	50.39±1.64

Table: Accuracy (in %) of 10 runs of the old omnilingual regression model

1 Reminder

2 State of the project

3 Results

4 Software

Last time

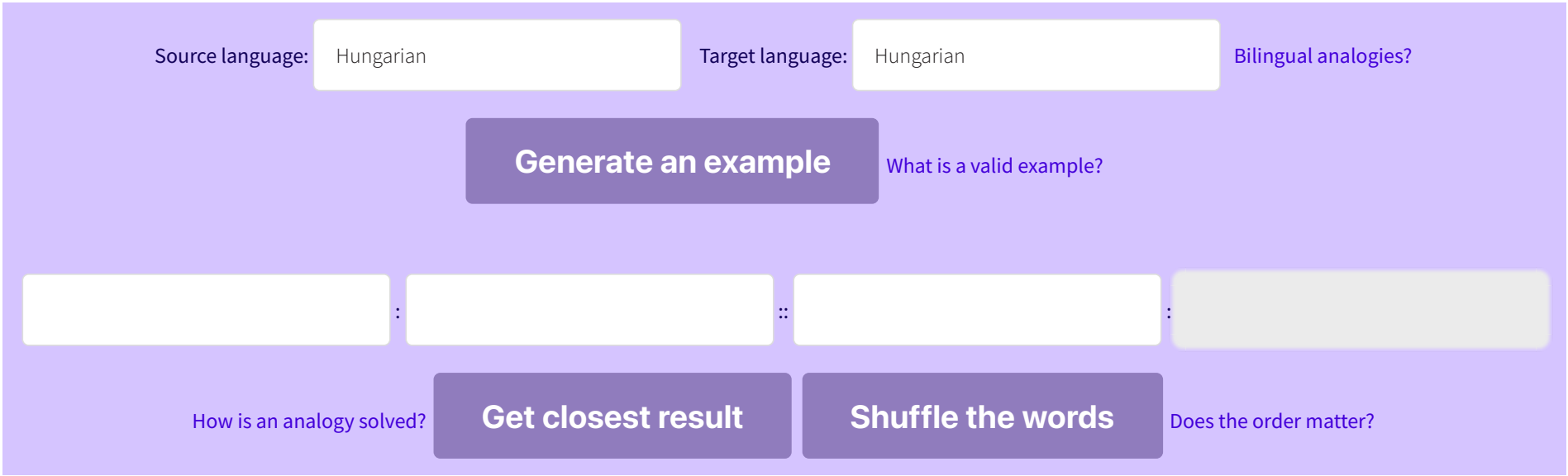

Figure: Preview of our previous software

No answer given

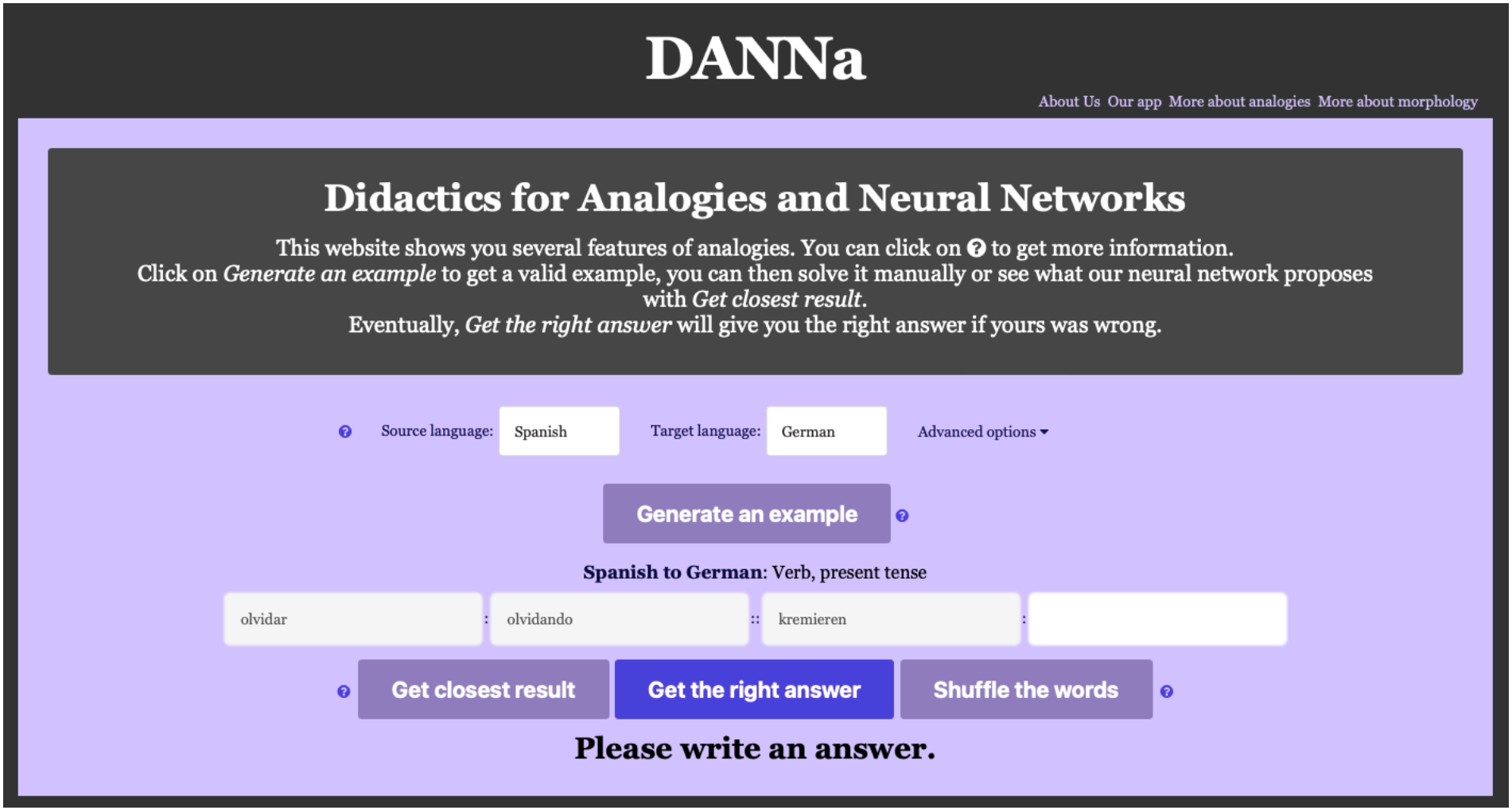

Figure: Preview of our software

Valid analogy

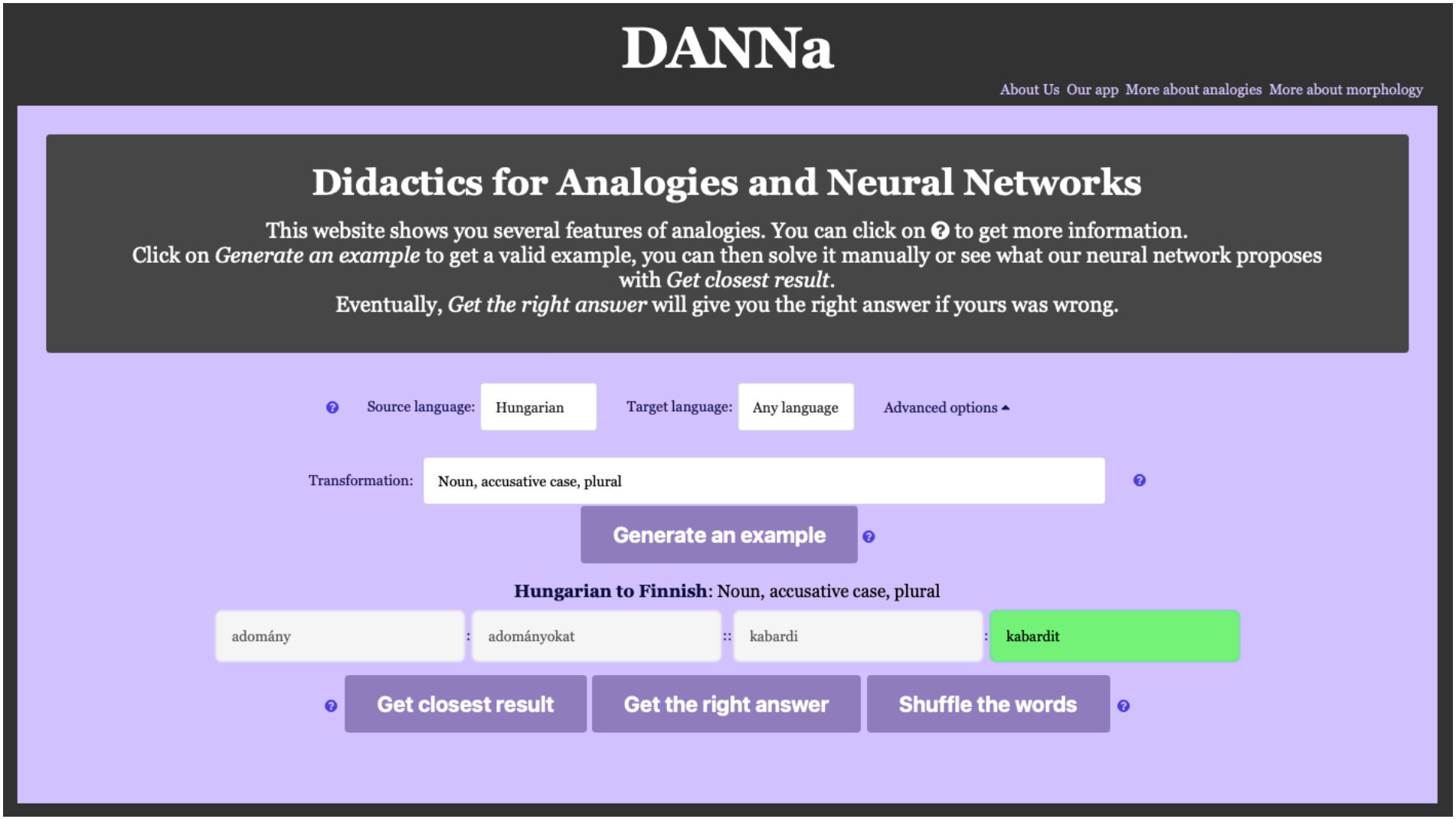

Figure: Preview of our software

Invalid analogy

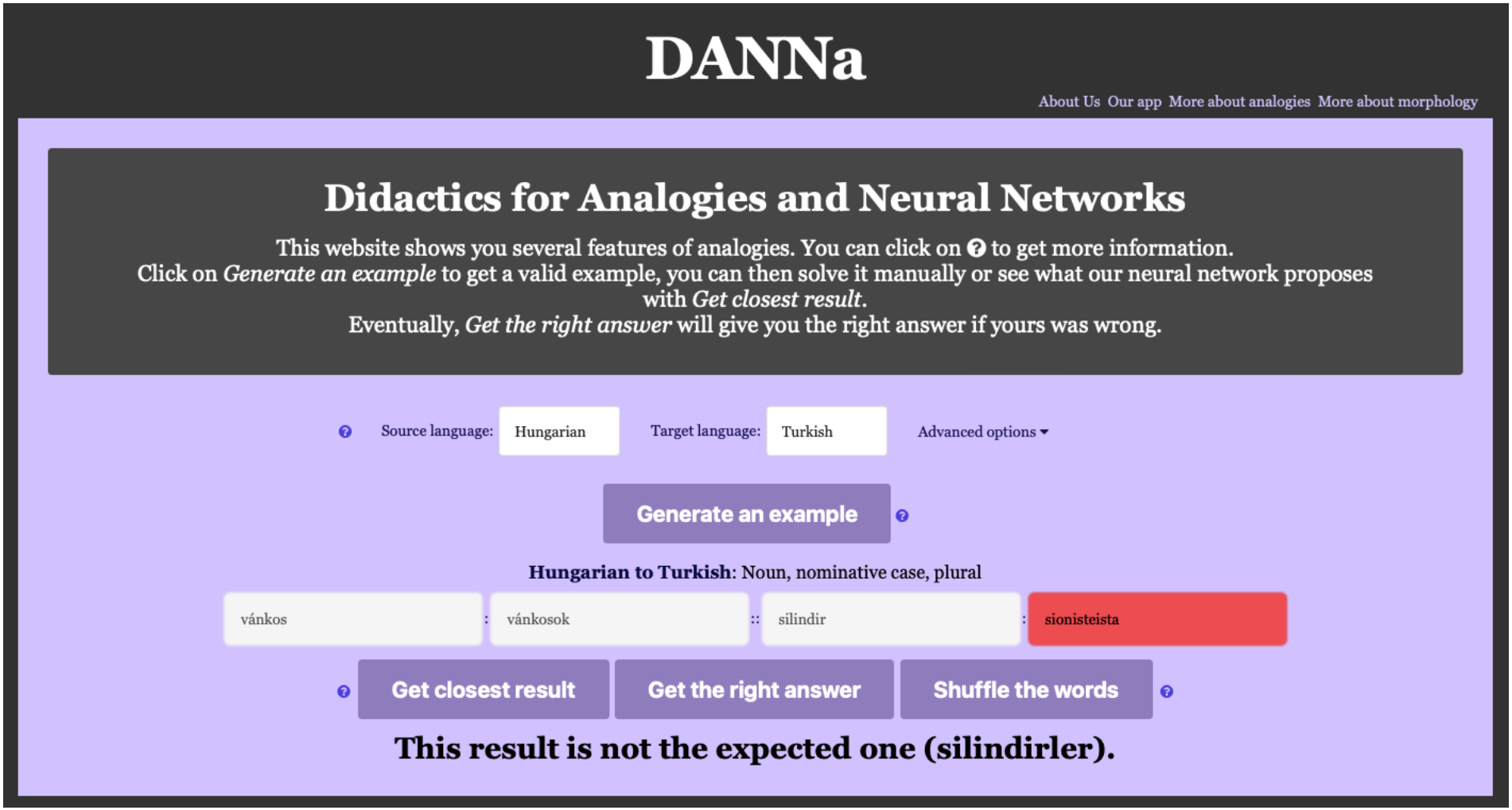

Figure: Preview of our software

شكرا جزيلا

Thank you

Merci

អរគុណ

Obrigado

Monolingual VS Bilingual (old)

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	/	42.06±2.82	80.13±1.37	/	81.53±1.57
German	94.51±0.60	/	71.94±2.84	71.94±2.84	/
Hungarian	42.99±4.69	84.22±3.33	/	84.22±3.33	40.75±1.89
Spanish	/	93.75±0.38	93.75±0.38	/	96.40±0.42
Turkish	67.43±1.06	/	**73.61±0.77**	95.17±1.12	/

Table: Accuracy (in %) of 10 runs of the old regression model

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	/	83.24±0.35	35.58±0.74	/	28.14±1.14
German	80.13±0.57	/	30.14±0.29	12.92±6.03	/
Hungarian	51.21±3.09	78.09±0.74	/	94.05±0.12	34.55±0.60
Spanish	/	36.79±11.74	79.11±0.63	/	41.83±0.85
Turkish	47.23±0.91	/	15.29±0.85	70.79±0.06	/

Table: Accuracy (in %) of 5 runs of the old bilingual regression models

Monolingual VS Bilingual (new)

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	/	39.75±2.67	80.33±1.95	/	79.89±2.43
German	94.26±0.63	/	70.55±2.61	70.55±2.61	/
Hungarian	43.93±3.44	85.39±1.76	/	85.39±1.76	40.98±3.42
Spanish	/	94.26±0.53	94.26±0.53	/	95.83±0.24
Turkish	64.98±2.76	/	70.74±2.23	94.03±3.70	/

Table: Accuracy (in %) of 10 runs of the new regression model

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	/	66.01±33.01	36.28±0.76	/	28.34±0.48
German	64.43±32.23	/	30.53±0.57	11.92±3.30	/
Hungarian	50.61±2.26	77.98±1.24	/	94.02±0.29	33.89±0.77
Spanish	/	32.42±17.01	78.99±0.13	/	40.45±1.52
Turkish	46.43±1.24	/	16.07±0.90	70.86±0.04	/

Table: Accuracy (in %) of 5 runs of the new bilingual regression models

Bilingual

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	/	66.01±33.01	36.28±0.76	/	28.34±0.48
German	64.43±32.23	/	30.53±0.57	11.92±3.30	/
Hungarian	50.61±2.26	77.98±1.24	/	94.02±0.29	33.89±0.77
Spanish	/	32.42±17.01	78.99±0.13	/	40.45±1.52
Turkish	46.43±1.24	/	16.07±0.90	70.86±0.04	/

Table: Accuracy (in %) of 5 runs of the new bilingual regression models

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	/	83.24±0.35	35.58±0.74	/	28.14±1.14
German	80.13±0.57	/	30.14±0.29	12.92±6.03	/
Hungarian	51.21±3.09	78.09±0.74	/	94.05±0.12	34.55±0.60
Spanish	/	36.79±11.74	79.11±0.63	/	41.83±0.85
Turkish	47.23±0.91	/	15.29±0.85	70.79±0.06	/

Table: Accuracy (in %) of 5 runs of the old bilingual regression models

Omnilingual model: Full dataset vs Bilingual features only

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	61.85±1.79	3.22±1.73	28.58±2.69	/	55.23±3.85
German	3.15±1.47	64.38±1.27	**66.91±4.49**	62.07±3.58	/
Hungarian	36.26±4.52	55.25±1.47	**73.33±1.31**	**78.36±1.33**	**32.00±3.03**
Spanish	/	61.16±2.54	**74.05±1.77**	69.38±1.65	70.67±4.03
Turkish	54.12±1.48	/	25.38±3.94	65.72±6.09	**52.23±1.09**

Table: Accuracy (in %) of 10 runs of the new omnilingual regression model trained on the full dataset

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	**52.41±1.35**	24.79±1.89	38.22±2.26	/	48.81±1.33
German	18.40±4.44	31.24±1.25	**79.08±1.85**	55.48±1.09	/
Hungarian	**64.11±2.06**	68.52±1.10	**10.56±0.37**	**79.29±1.81**	**41.26±2.96**
Spanish	/	51.27±1.50	**80.04±1.64**	**28.91±0.65**	74.57±5.75
Turkish	**59.89±1.54**	/	26.52±2.69	83.49±2.41	21.86±1.04

Table: Accuracy (in %) of 10 runs of the new omnilingual regression model trained only on the bilingual features

Analogy solver model

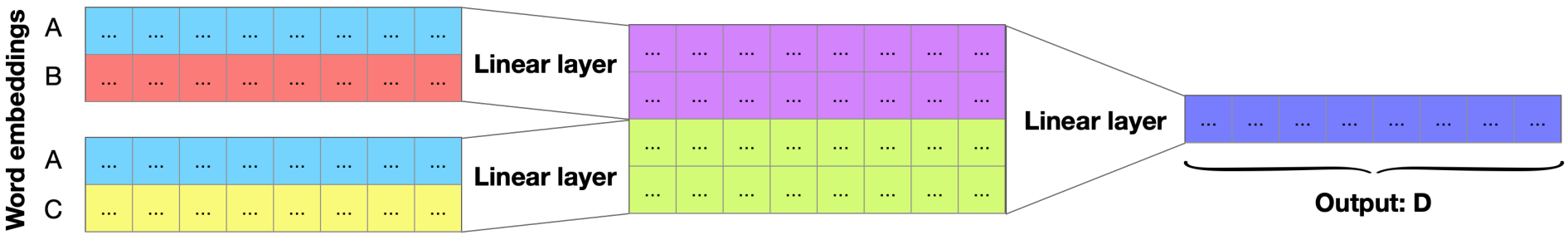

Figure: Analogy solver model