

# Towards an omnilingual model for solving morphological analogies

Safa AlSaidi, Amandine Decker, Stephanie Monteiro

M2 — Software Project



- 1 Reminder
- 2 Our approach
- 3 Regression task
- 4 Results & discussion
- 5 Software
- 6 Future work

# 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

# First trial results

|                    | hungarian, german | turkish, finnish | hungarian, finnish |
|--------------------|-------------------|------------------|--------------------|
| Cosine similarity  | <b>58.9</b>       | <b>39.5</b>      | <b>18.9</b>        |
| Euclidean distance | 57.7              | 39.1             | 16.8               |

Table: Accuracy for the regression task on the three (source, target) language pairs.

- 1 Reminder
- 2 Our approach**
- 3 Regression task
- 4 Results & discussion
- 5 Software
- 6 Future work

# 3 approaches

- ① Comparable data
- ② Omnilingual model
- ③ Sigmorphon 2019

- 1 Reminder
- 2 Our approach
- 3 Regression task**
- 4 Results & discussion
- 5 Software
- 6 Future work

# New results

Table: Accuracy (in %) of 3 runs of the regression model.

| Language  | ANNr (previous)<br>(mean $\pm$ std.) | actual                  |
|-----------|--------------------------------------|-------------------------|
| Arabic    | <b>77.97</b> $\pm$ 16.03             | <b>61.13</b> $\pm$ 0.83 |
| Finnish   | <b>37.78</b> $\pm$ 9.28              | <b>77.56</b> $\pm$ 1.78 |
| Georgian  | <b>94.66</b> $\pm$ 1.13              | <b>86.40</b> $\pm$ 0.62 |
| German    | <b>86.38</b> $\pm$ 0.45              | <b>86.93</b> $\pm$ 0.78 |
| Hungarian | <b>53.83</b> $\pm$ 3.12              | <b>78.98</b> $\pm$ 0.50 |
| Maltese   | <b>75.00</b> $\pm$ 5.08              | <b>79.66</b> $\pm$ 1.11 |
| Navajo    | <b>31.74</b> $\pm$ 0.90              | <b>45.88</b> $\pm$ 0.24 |
| Russian   | <b>75.15</b> $\pm$ 0.44              | <b>70.53</b> $\pm$ 0.37 |
| Spanish   | <b>86.27</b> $\pm$ 0.71              | <b>91.12</b> $\pm$ 1.06 |
| Turkish   | <b>61.95</b> $\pm$ 10.86             | <b>80.34</b> $\pm$ 0.79 |
| Japanese  | <b>61.60</b> $\pm$ 1.33              | <b>79.92</b> $\pm$ 0.02 |



- 1 Reminder
- 2 Our approach
- 3 Regression task
- 4 Results & discussion**
- 5 Software
- 6 Future work

# Bilingual analogies

In our dataset:  $\text{WORD}_1 \text{ FEATURES } \text{WORD}_2$

An analogy:  $\text{WORD}_{1,A}:\text{WORD}_{2,A}::\text{WORD}_{1,B}:\text{WORD}_{1,B}$   
where  $\text{FEATURES}_A = \text{FEATURES}_B$

Bilingual analogies:  $\text{LANGUAGE}_A \neq \text{LANGUAGE}_B$

→ keep only the subset of *shared features*

# Shared features

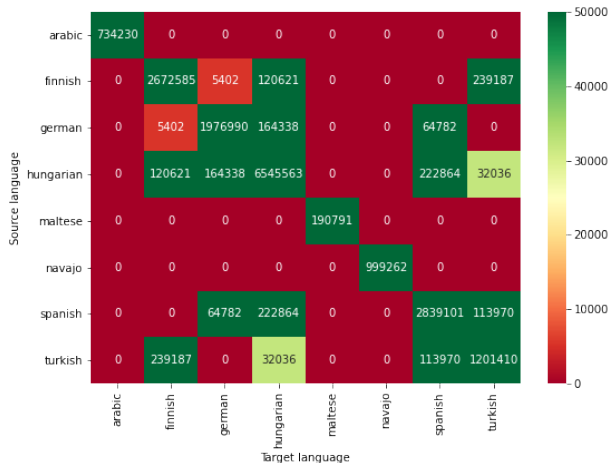


Figure: Number of possible analogies for each pair of languages

# Comparison between monolingual and bilingual results

|           | Finnish          | German           | Hungarian        | Spanish          | Turkish          |
|-----------|------------------|------------------|------------------|------------------|------------------|
| Finnish   | /                | 43.96 $\pm$ 1.48 | 80.93 $\pm$ 1.94 | /                | 82.00 $\pm$ 1.90 |
| German    | 92.63 $\pm$ 0.10 | /                | 68.17 $\pm$ 3.12 | 68.17 $\pm$ 3.12 | /                |
| Hungarian | 43.07 $\pm$ 0.48 | 85.92 $\pm$ 0.83 | /                | 85.92 $\pm$ 0.83 | 40.92 $\pm$ 2.46 |
| Spanish   | /                | 93.97 $\pm$ 0.25 | 93.97 $\pm$ 0.25 | /                | 94.05 $\pm$ 0.31 |
| Turkish   | 65.89 $\pm$ 1.59 | /                | 71.76 $\pm$ 0.92 | 93.18 $\pm$ 1.90 | /                |

Table: Monolingual analogies: Accuracy ( $\pm$ std) on 3 runs

|           | Finnish          | German           | Hungarian | Spanish      | Turkish |
|-----------|------------------|------------------|-----------|--------------|---------|
| Finnish   | /                | <b>81.88</b>     | 35.88     | /            | 30.19   |
| German    | 80.31            | /                | 30.41     | 35.10        | /       |
| Hungarian | 48.83 $\pm$ 3.19 | 78.41 $\pm$ 1.59 | /         | <b>91.62</b> | 33.93   |
| Spanish   | /                | 17.63            | 83.26     | /            | 40.63   |
| Turkish   | 45.81 $\pm$ 0.17 | /                | 16.17     | 70.27        | /       |

Table: Bilingual analogies: Accuracy ( $\pm$ std) on 3 runs

# Omnilingual model

Languages which share features with at least one other language: Finnish, German, Hungarian, Turkish, Spanish

|           | Finnish          | German           | Hungarian        | Spanish          | Turkish          |
|-----------|------------------|------------------|------------------|------------------|------------------|
| Finnish   | $60.30 \pm 1.26$ | $3.08 \pm 0.86$  | $31.78 \pm 1.79$ | /                | $52.62 \pm 1.84$ |
| German    | $3.08 \pm 0.86$  | $63.27 \pm 0.68$ | $57.71 \pm 0.48$ | $62.47 \pm 2.41$ | /                |
| Hungarian | $31.78 \pm 1.79$ | $57.71 \pm 0.48$ | $71.12 \pm 1.04$ | $62.89 \pm 1.86$ | $24.73 \pm 1.26$ |
| Spanish   | /                | $62.47 \pm 2.41$ | $62.89 \pm 1.86$ | $66.82 \pm 1.34$ | $62.20 \pm 6.57$ |
| Turkish   | $52.62 \pm 1.84$ | /                | $24.73 \pm 1.26$ | $62.20 \pm 6.57$ | $49.73 \pm 0.82$ |

Table: Accuracy ( $\pm$ std) on 5 runs

# Next time: Sigmorphon 2019 [McCarthy et al., 2019]

88 languages: 8/10 from Sigmorphon 2016 [Cotterell et al., 2016]

→ Arabic, Finnish, German, Hungarian, Russian, Spanish, Turkish, Maltese  
(Georgian and Navajo missing)

Aim: apply trained models to the new dataset

- 1 Reminder
- 2 Our approach
- 3 Regression task
- 4 Results & discussion
- 5 Software**
- 6 Future work

# What we want

- Solving analogies: monolingual and bilingual
- Use the omnilingual model



# What it looks like

The interface is a web-based tool for generating bilingual analogies. It features a light purple background with white text and buttons. At the top, there are two input fields for 'Source language' and 'Target language', both containing the word 'Hungarian'. To the right of these fields is the text 'Bilingual analogies?'. Below these fields is a large purple button with the text 'Generate an example'. To the right of this button is the text 'What is a valid example?'. Below the button, there are four input fields arranged horizontally. The first three fields are white and contain the words 'apple', 'orange', and 'pear' respectively. The fourth field is a light gray placeholder. Between the first and second fields is a colon ':', and between the second and third fields is a double colon '::'. Below the input fields, there are two purple buttons: 'Get closest result' and 'Shuffle the words'. To the left of the 'Get closest result' button is the text 'How is an analogy solved?', and to the right of the 'Shuffle the words' button is the text 'Does the order matter?'.

Source language: Hungarian Target language: Hungarian Bilingual analogies?

Generate an example What is a valid example?

apple : orange :: pear :

How is an analogy solved? Get closest result Shuffle the words Does the order matter?

Figure: Preview of our software

- 1 Reminder
- 2 Our approach
- 3 Regression task
- 4 Results & discussion
- 5 Software
- 6 Future work**

# Future work

- Run final experiments
- Improve and adapt our software
- Continue writing the report

شكرا جزيلا

Thank you

Merci

អរគុណ

Obrigado

# References I



Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., and Hulden, M. (2016).

The sigmorphon 2016 shared task—morphological reinflection.  
In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany.  
Association for Computational Linguistics.



McCarthy, A. D., Vylomova, E., Wu, S., Malaviya, C., Wolf-Sonkin, L., Nicolai, G., Kirov, C., Silfverberg, M., Mielke, S. J., Heinz, J., Cotterell, R., and Hulden, M. (2019).

The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection.  
In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy.  
Association for Computational Linguistics.