Friedrich-Alexander-Universität
Technische Fakultät

Pattern Recognition Lab

# The hippocampus and language: Word to word prediction in terms of the successor representation

Philipp Rost

Friedrich-Alexander-Universität Erlangen-Nürnberg
Department Informatik

July 14, 2022

# Agenda

1. Introduction
2. Theoretical Background
3. Framework
4. Results
5. Conclusion

# Introduction

# Introduction

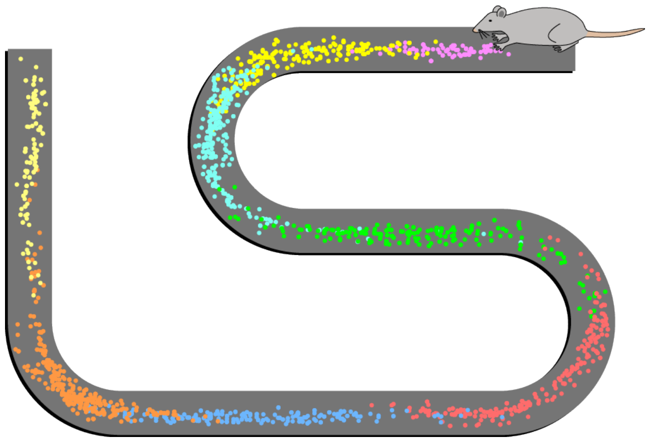
- Understanding the human brain is a challenge as old as science itself
- Currently available technology as a metaphor (from abacus to computer)
- Projects exist researching the brain as a whole but also for distinct parts, e.g. the hippocampus
- Goal: Expanding the application of the Successor Representation (SR) to language
  - supposedly used by the hippocampus to predict following states/positions
- To reach it, a neural network is trained with samples extracted from two books

# Theoretical Background

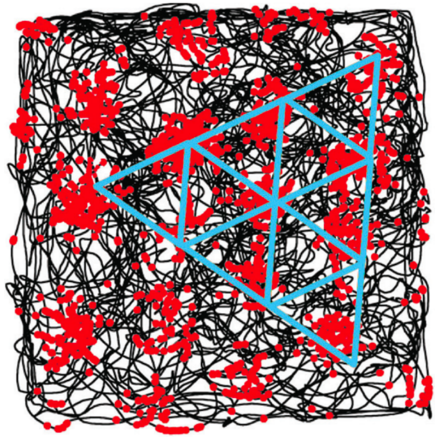
# Hippocampus

- Has the shape of a seahorse
- Key factor in forming memories (not preserving them!) [1]
- Related to emotions [2]
- Responsible for any kind of navigation (e.g. ranking stuff like danger of animals)
  - Crafts a cognitive room of the “surroundings” by using place cells and grid cells
  - Place cell: irregular arranged, fires at specific positions in space (“states”)
  - Grid cell: lattice-like arranged, fires continuously

Hippocampus and seahorse [3]

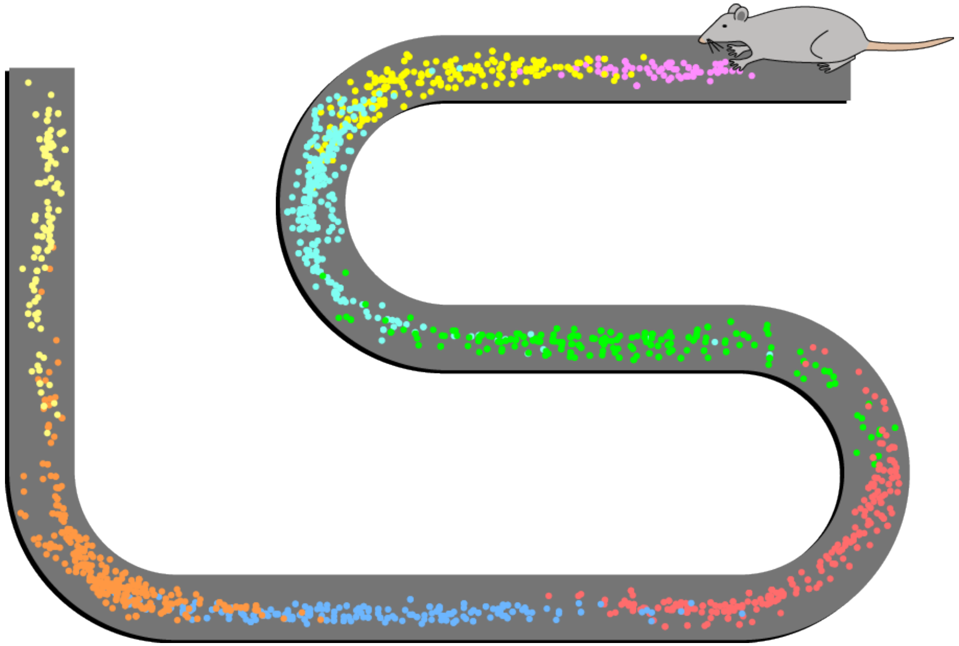


Color coded place cell activity (= states) [4]

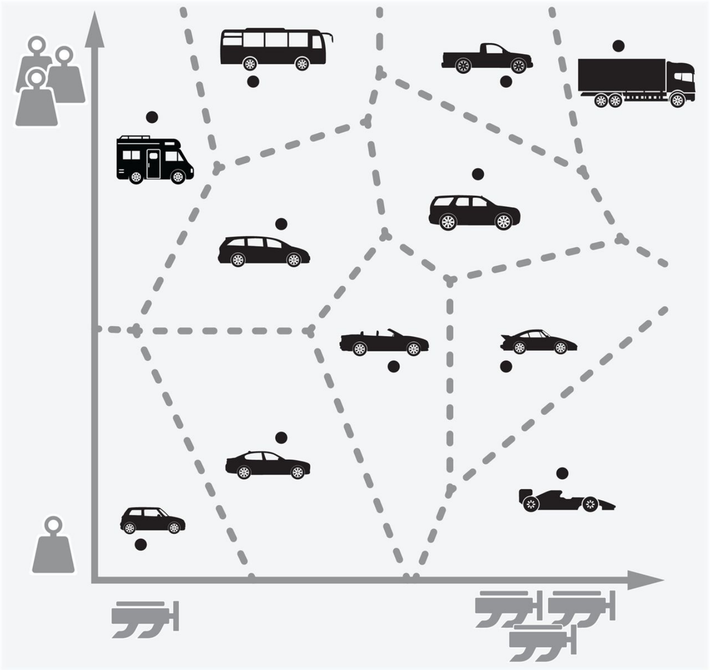


Grid cells form a triangulation [5]

# Projective map theory & cognitive room



Place cell fires if rat is about to enter the state, e.g. turquoise cell before in front of the first arch [4]



Cognitive rooms help to locate unknown objects and put them into relationship, e.g. unknown cars [6]

- Claim: the hippocampus applies the projective map theory and encodes each state within a cognitive room [7]

# Successor Representation (SR)

- Claim: the hippocampus applies the projective map theory and encodes each state within a cognitive room
  - Where does the claim originate? The proposed technique works fine for spatial navigation [7]
  - Mathematification of the concepts: Successor Representation
- Roots lay in reinforcement learning (and transition probability matrices) and can be computed like

$$M_a = \sum_{t=0}^{a} \gamma^t T^t,$$

with discount factor $\gamma \in (0, 1)$, $a = 1, ..., \infty$ and transition probability matrix $T$

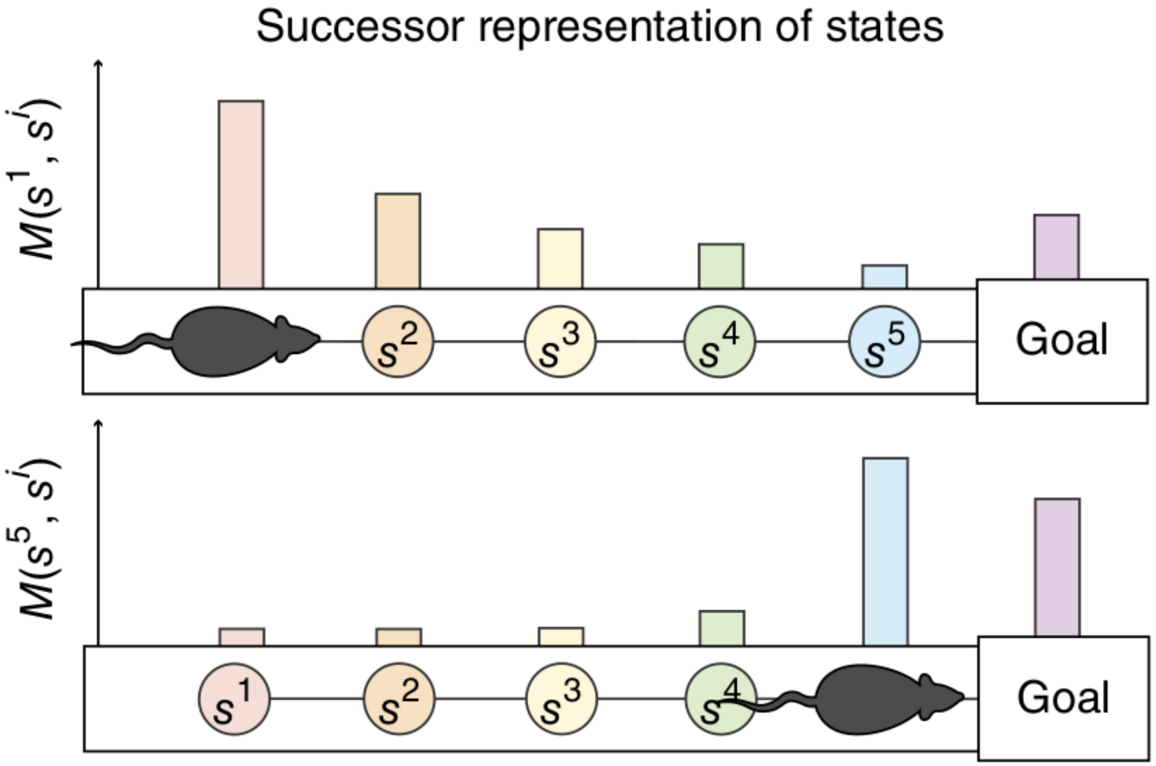
- The policy/structure of the language is encoded in matrix $T \implies$ the SR is policy-dependent (it is based on RL)
- By inspecting row $k$ it is possible to follow all paths starting at state $k$

  $\implies$ $M_a$ reveals all successor states

# Successor Representation

Example 1/2 of interpreting a SR matrix $M$

- Row $i$: resembles (all) successor states of state $i$ (the higher the value the higher the chance to be in this state after the next step)




*Superscript means index not power; image taken from [7]*

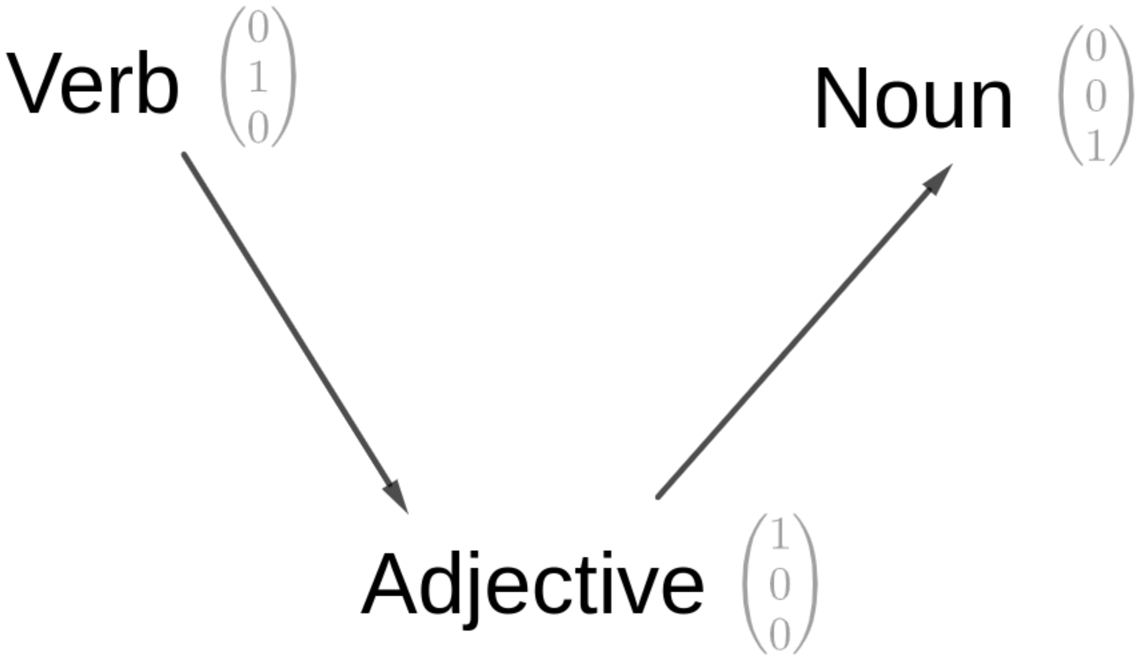
# Framework

# Overview

- Shallow dense neural network & supervised learning
- Goal: Learning the SR i.e., a transition probability matrix
- Two configurations were tested
  1. Artificial rules with manufactured data set (“First model”)
  2. Self derived rules and data set (“Word to word model”)
- Rule: word pair consisting of a predecessor and successor word serving as input and output
- In case of word to word models: Data was collected from two books (german & english)
- The quality of the learned rules determines the SR

# First model

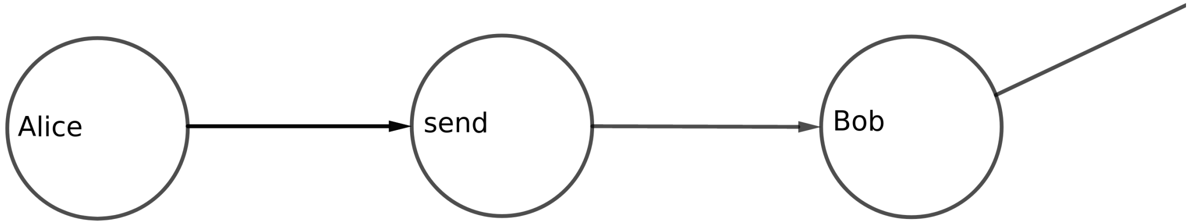
- Cognitive room consists of all words used for training
- Data was generated by made up rules like `Verb → Adjective` using 1-hot-encoded vectors
- The single predictions after training describe the transition probability matrix
- This type of model is tailored and clear



# Word to word model

- In principle similar to the first model approach
- But rules and data derived from real language examples
- Books were parsed using techniques from Natural Language Processing (via spacy)
  - In german and english, because the former's word order is more variable $\rightarrow$ may cause troubles

**Alice sends Bob** a message. $\Longrightarrow$

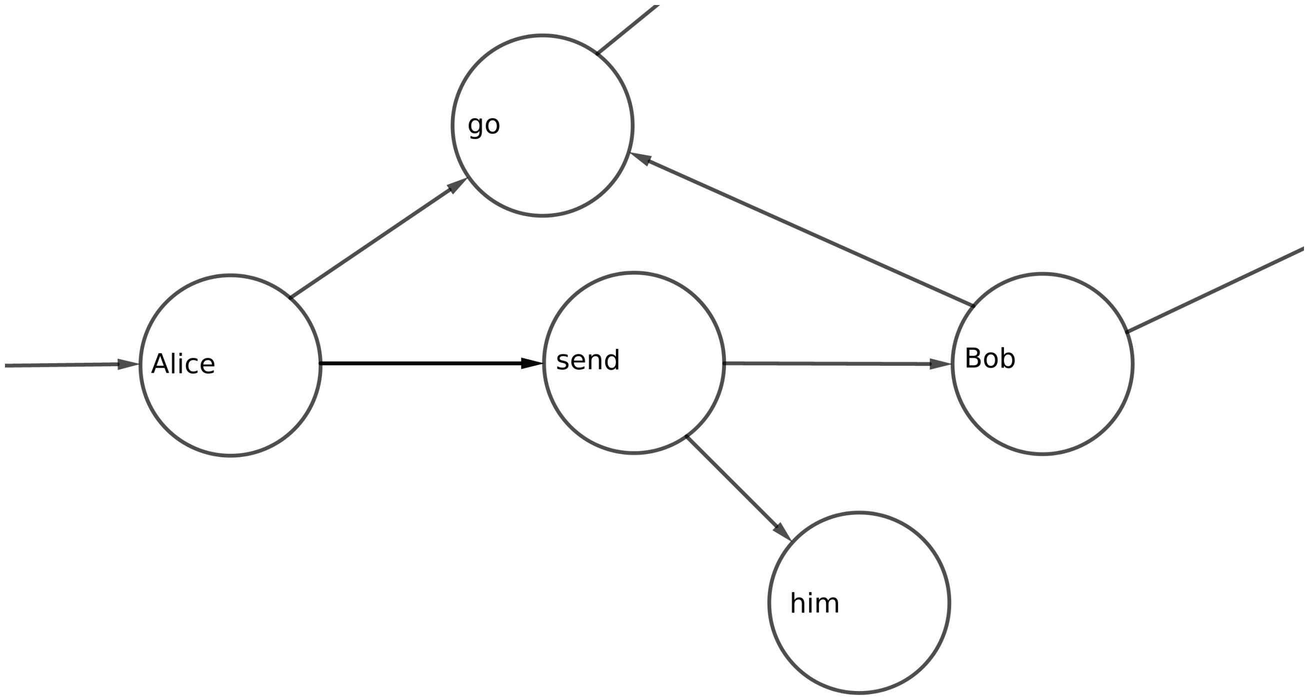


# Word to word model

- In principle similar to the first model approach
- But rules and data derived from real language examples
- Books were parsed using techniques from Natural Language Processing (via spacy)
  - In german and english, because the former's word order is more variable $\rightarrow$ may cause troubles

**Alice sends Bob** a message. [...] **Alice goes** to the grocery store. [...]. Peter **sent him** a letter. [...] **Bob went** to his friend.

$\Longrightarrow$



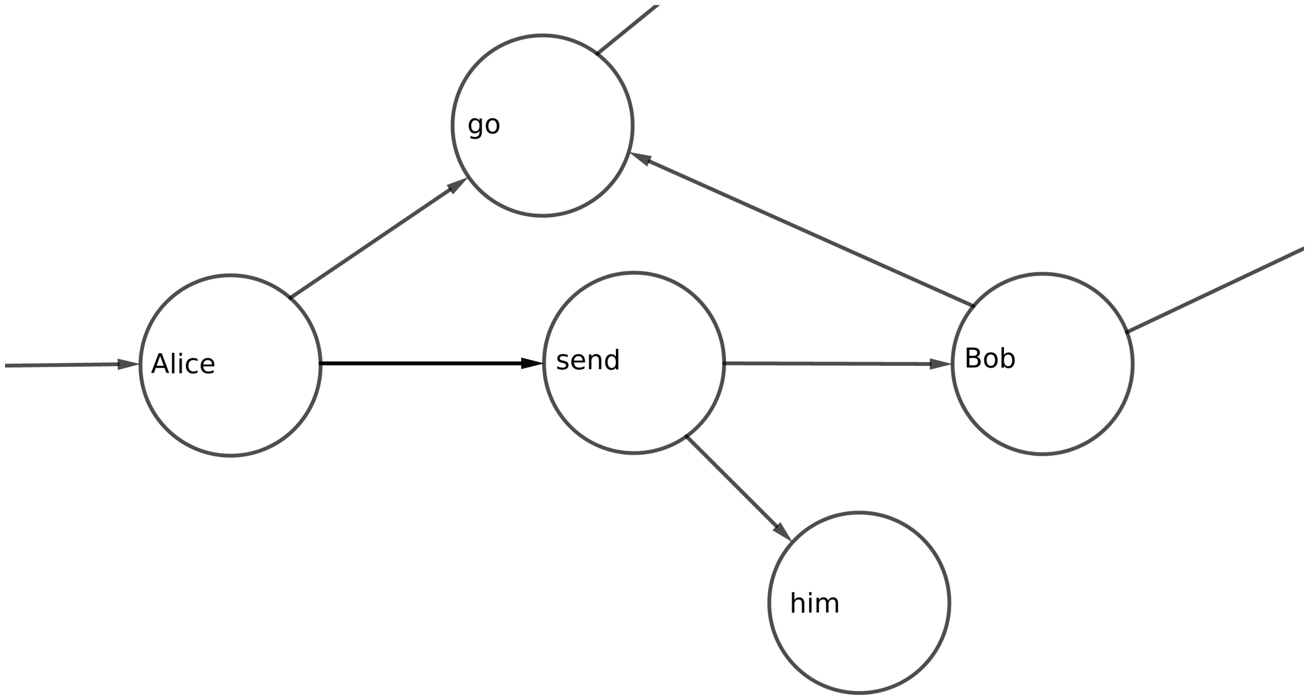
# Word to word flavors
## 1-hot-encoded vectors & Word vectors

- Word to word models come in two flavors
  - 1-hot-encoded vectors and
  - Word vectors

- Are 300$d$ real valued vectors
- Potentially incorporate more information about a word

$\Rightarrow$ Better learning possible?
- Probably the hippocampus receives multiple signals which are in total more related to word vector than to 1-hot-encoded vector

$\Rightarrow$ Closer to reality



# Word to word Models

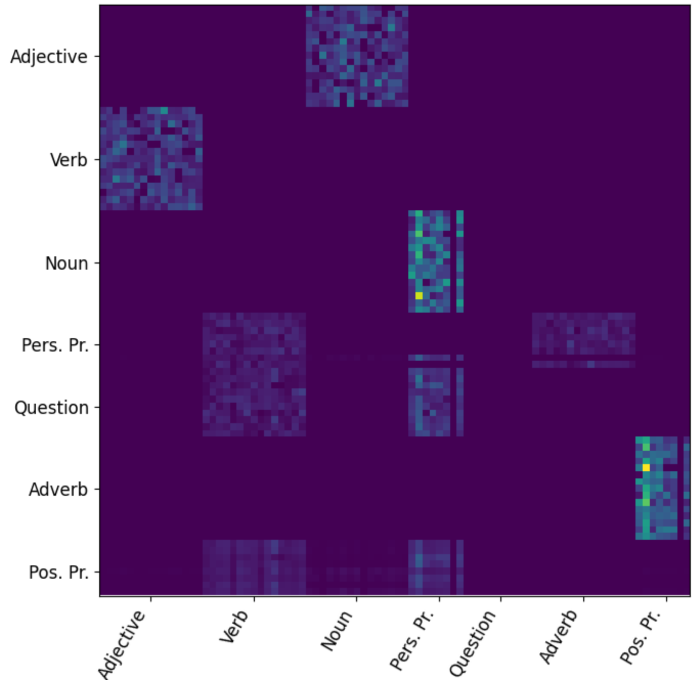
## Average approach

- Predicting all instances of a word class at once and average the result (into one vector)
- Word classes are inferred by `spacy`, 10 in total are used
- Idea: Meta word pairs like `Pronoun → Verb` appear more frequently than `he → plays`
- Averaging is done with both vector types: 1-hot-encoded vectors and word vectors

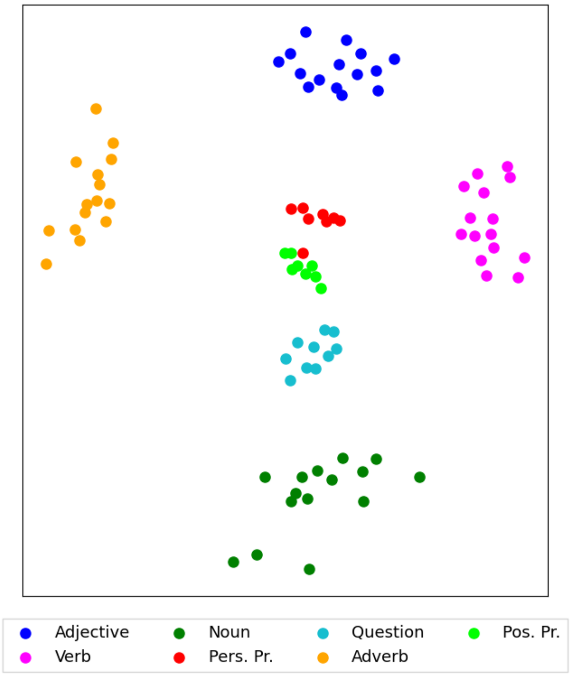
# Results

# Results – First model

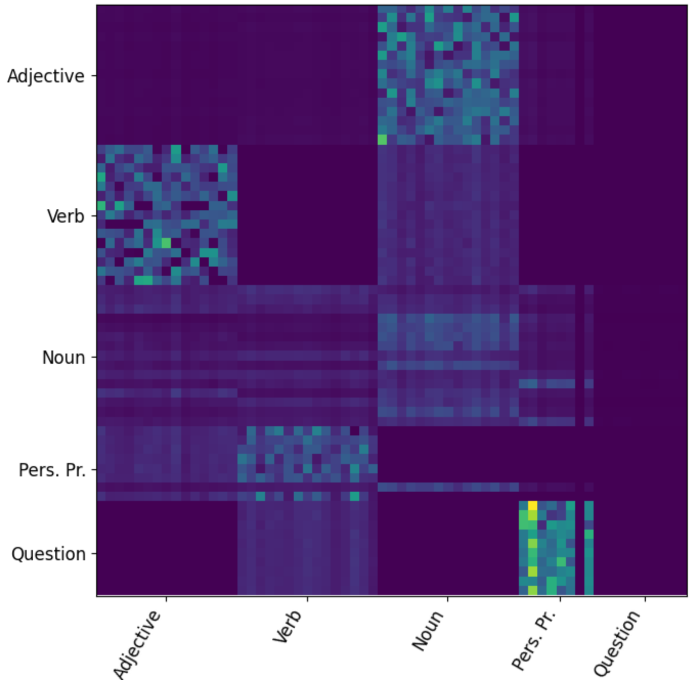
- Prediction works quite well i.e., the rules are recognizable e.g., `Adjective → Noun`
- MDS plot shows clustered word classes




Learned SR



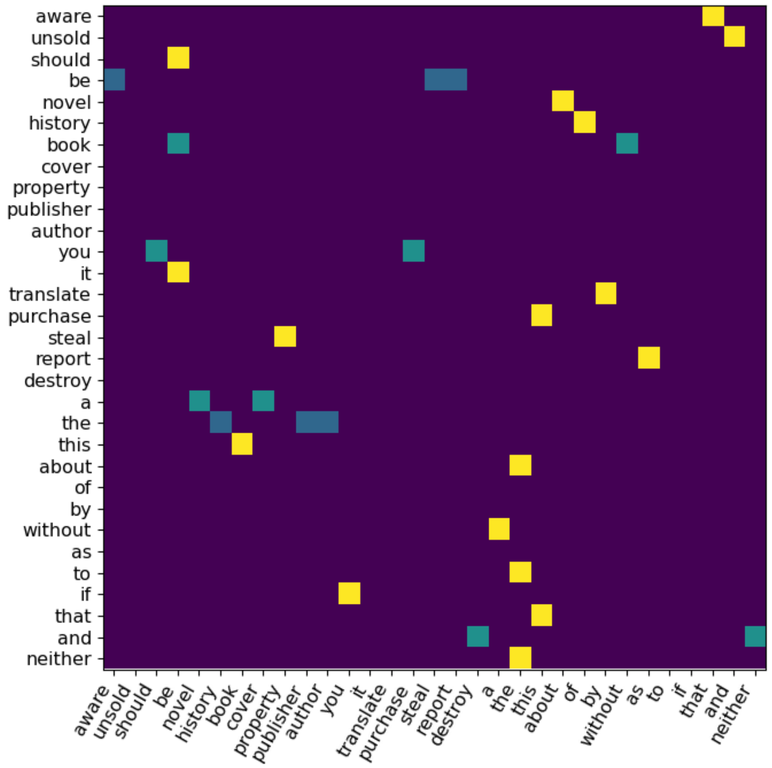

MDS plot



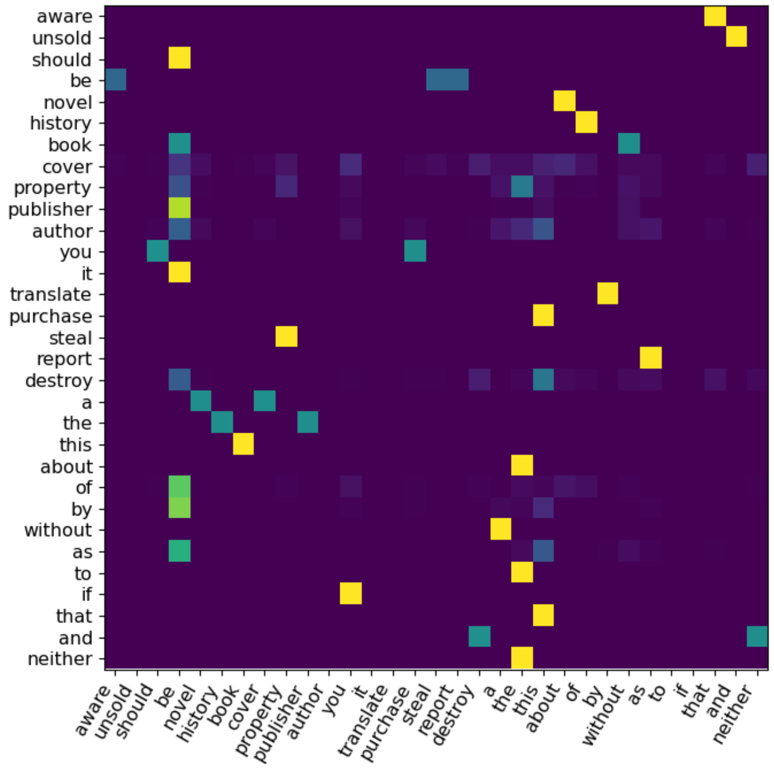

SR for $t = 2$, less word classes

# Results – Word to word models

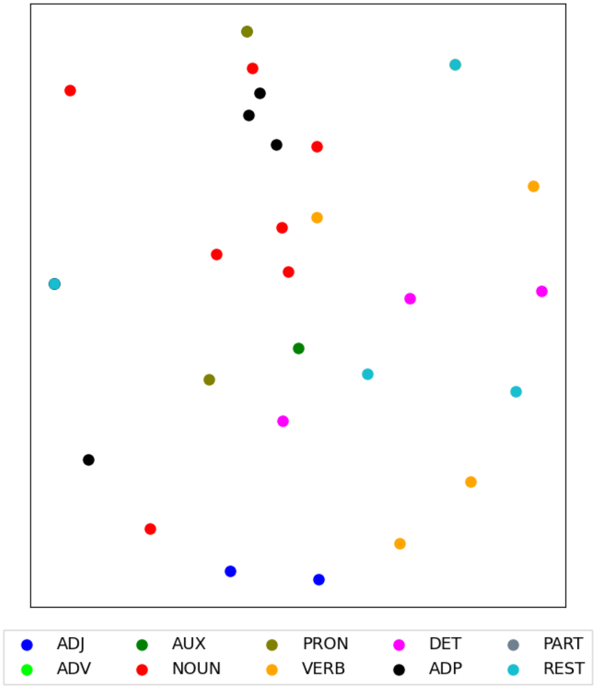
- Comparison with a ground truth/statistical assessment possible by a metric




Ground truth, Transition probability matrix




Learned transition probability matrix



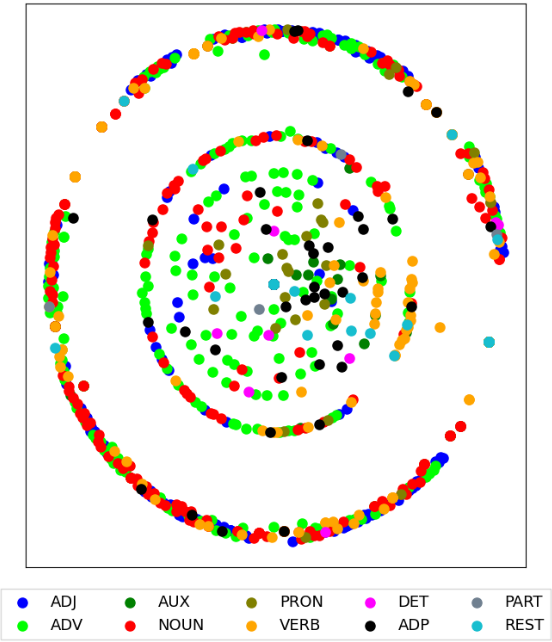

Learned MDS

# Results – Word to word models

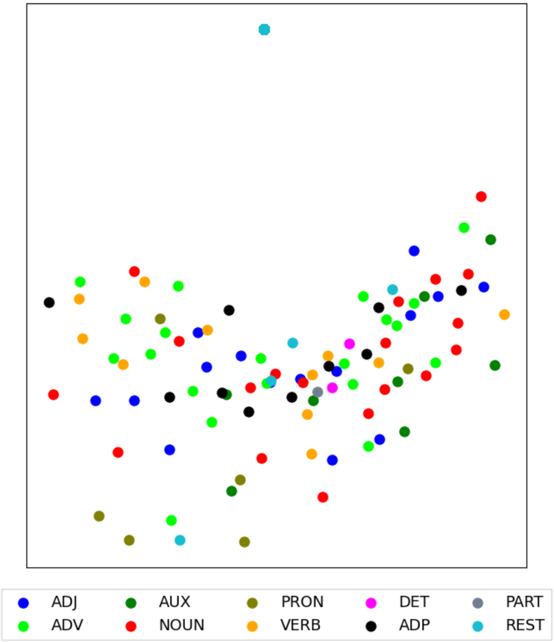
- It is possible to compare the results to a ground truth/statistical assessment $\Longrightarrow$ Metric $d_A$
- Surprisingly 1-hot-encoded vectors outperform word vectors i.e., word vectors are just bad
- German or english doesn't make that much of a difference

| Version | Metric |
|---|---|
| german, 1-hot-encoded vector | 0.08 |
| german, word vector | 0.74 |
| english, 1-hot-encoded vector | 0.10 |
| english, word vector | 0.78 |

Configurations & metric w.r.t. ground truth




MDS of german, 1-hot-encoded vector

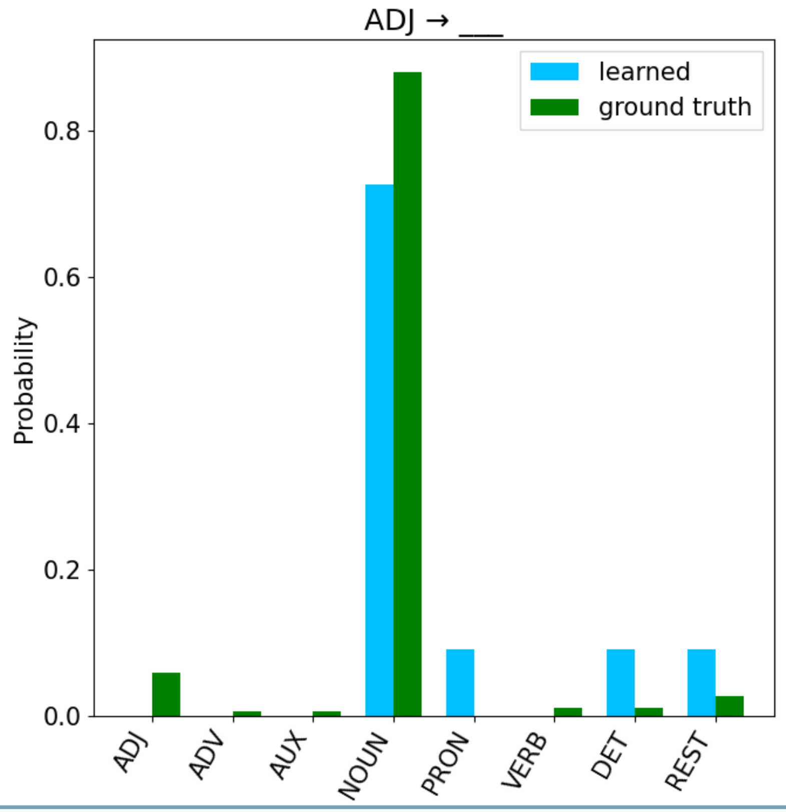


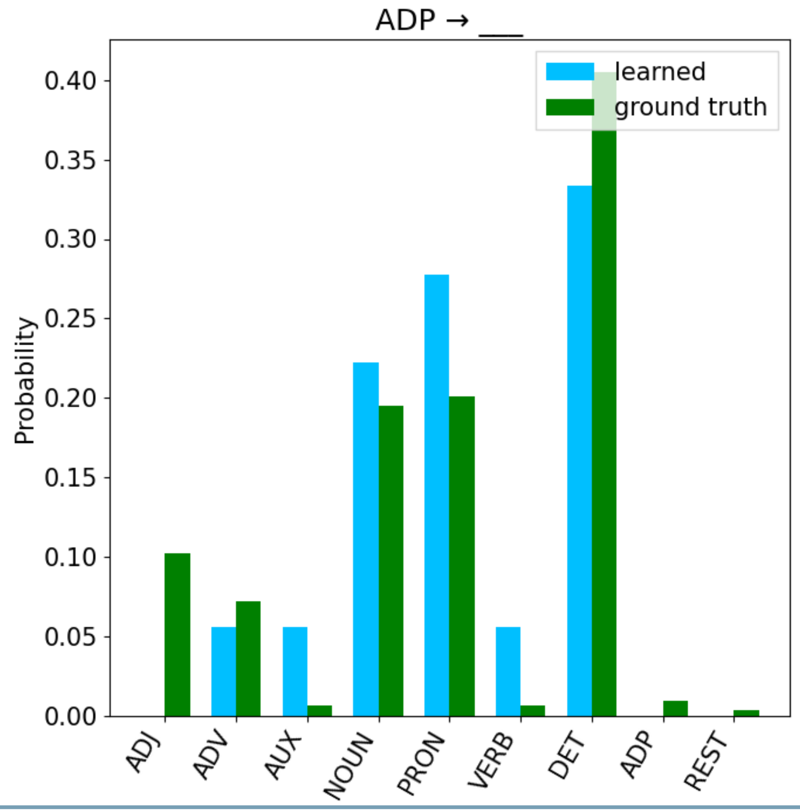

MDS of german, word vectors

*If you want to know more about the metric, you can ask after the talk*

# Results – Averaging models

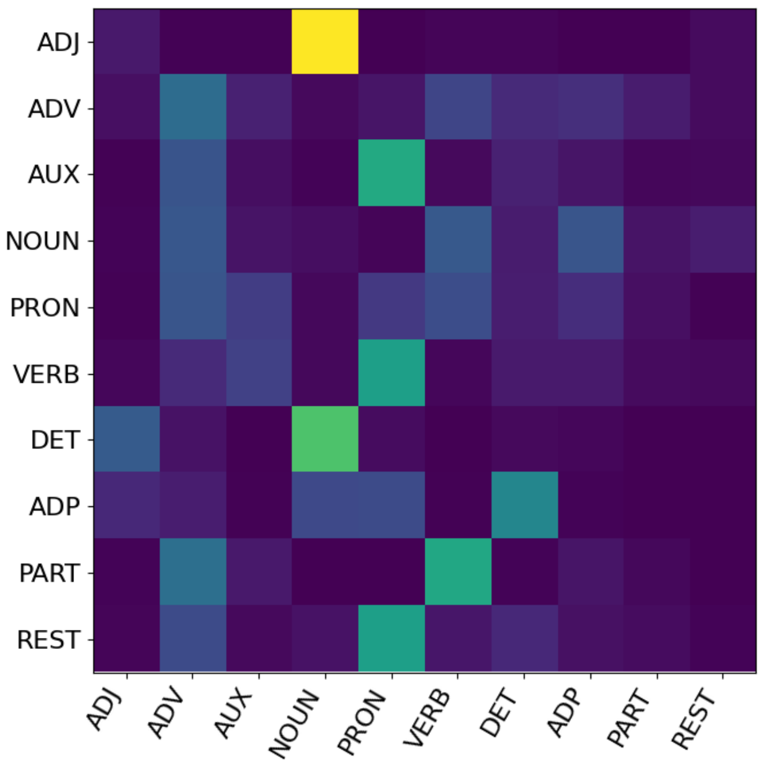
- Outcome of the plain vector models wasn't satisfying (as seen in the MDS plots), so averaging was established
- Results were indeed exploitable i.e., word class transition probabilities are partially reflected very accurately



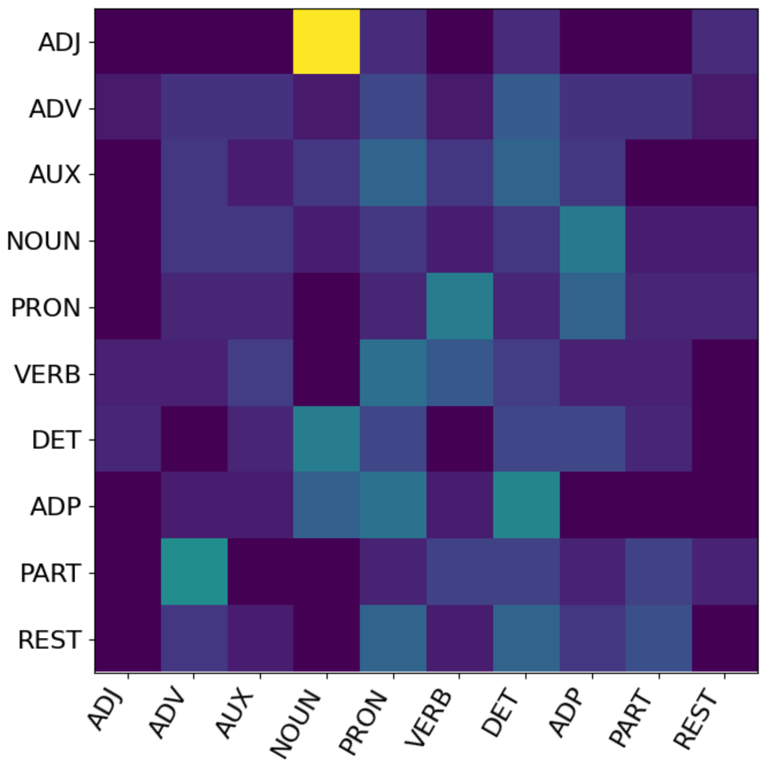



# Results – Averaging models

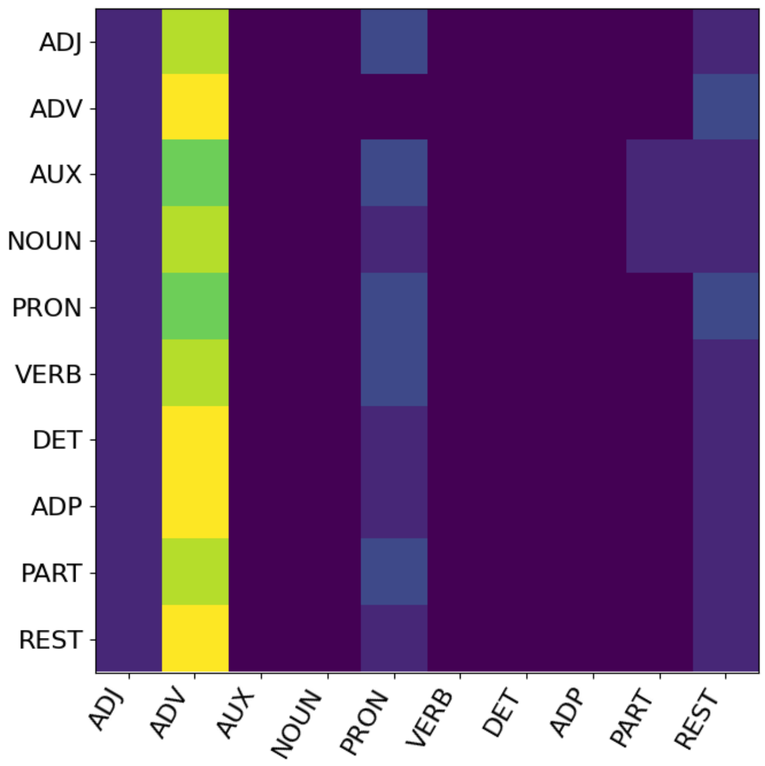
- Matrices are $10 \times 10$, so we display them




ground truth (german)




german, 1-hot-encoded vector




german, word vector

# Results – Averaging models

- Accuracy of these models is measured by mean and standard deviation:

| Version | Mean $\mu$ | Standard deviation $\sigma$ |
|---|---|---|
| german, 1-hot-encoded vector | 7.3 | 2.0 |
| german, word vector | 14.0 | 2.1 |
| english, 1-hot-encoded vector | 8.1 | 3.3 |
| english, word vector | 10.2 | 3.6 |

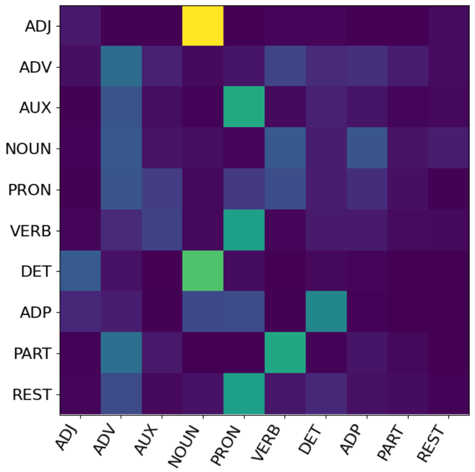
*Mean and standard deviation in* $10^{-2}$

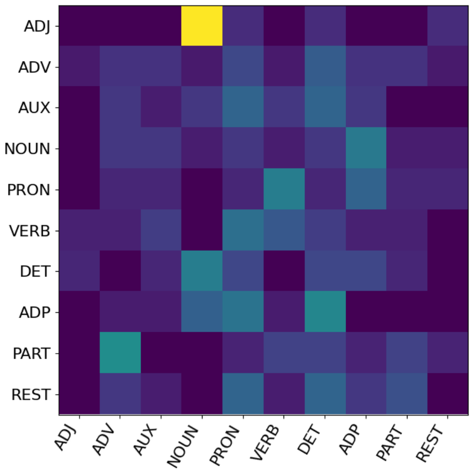
- Sadly, the outcome of word vector models is quite bad again
- But the 1-hot-encoded vectors seem to grasp the grammatical structure (bar and matrix plot)

# Conclusion

# Conclusion

- By far most of the time was consumed by finding proper values, sadly with bad luck
- Plenty of configurations didn't improve the results or were worse. Two of them were
  - Multiple hidden layers
  - Predicting only most frequent words
- Due to the lack of valid data from real experiments interpretation regarding our daily life is difficult
- Performance of word vectors disappointing, which is a drawback because they might be closer to actual signals
- Some learning does happen (Average approach)



Friedrich-Alexander-Universität
Technische Fakultät

Pattern
Recognition
Lab

# References

# References I


[1] M. Trepel, *Neuroanatomie*. München: Elsevier, 2017, ISBN: 9783437412882.

[2] N. Garzorz-Stark, *Basics Neuroanatomie*, 2nd ed. Urban and Fischer/Elsevier, 2018, ISBN: 9783437424588.

[3] L. Seress, "Hippocampus and seahorse," Online, accessed on May 27th 2022, License: https://creativecommons.org/licenses/by-sa/3.0/. (2010), [Online]. Available: https://commons.wikimedia.org/wiki/File:Hippocampus_and_seahorse_cropped.JPG.

[4] Stuartlayton, Online, accessed on May 14th 2022, License: https://creativecommons.org/licenses/by-sa/3.0/, User on https://en.wikipedia.org/wiki/ (english Wikipedia). (Jan. 2013), [Online]. Available: https://commons.wikimedia.org/wiki/File:Place_Cell_Spiking_Activity_Example.png.

[5] M.-B. Moser, D. Rowland, and E. Moser, "Place cells, grid cells, and memory," *Cold Spring Harbor perspectives in medicine*, vol. 5, a021808, Feb. 2015. DOI: 10.1101/cshperspect.a021808.

[6] J. L. S. Bellmund, P. Gärdenfors, E. I. Moser, and C. F. Doeller, "Navigating cognition: Spatial codes for human thinking," *Science*, vol. 362, 6415 Nov. 2018. DOI: 10.1126/science.aat6766. [Online]. Available: https://science.sciencemag.org/content/362/6415/eaat6766.

[7] K. L. Stachenfeld, M. M. Botvinick, and S. J. Gershman, "The hippocampus as a predictive map," *Nature Neuroscience*, Nov. 2017. DOI: 10.1038/nn.4650. [Online]. Available: https://www.nature.com/articles/nn.4650.

# Thank you for your attention!