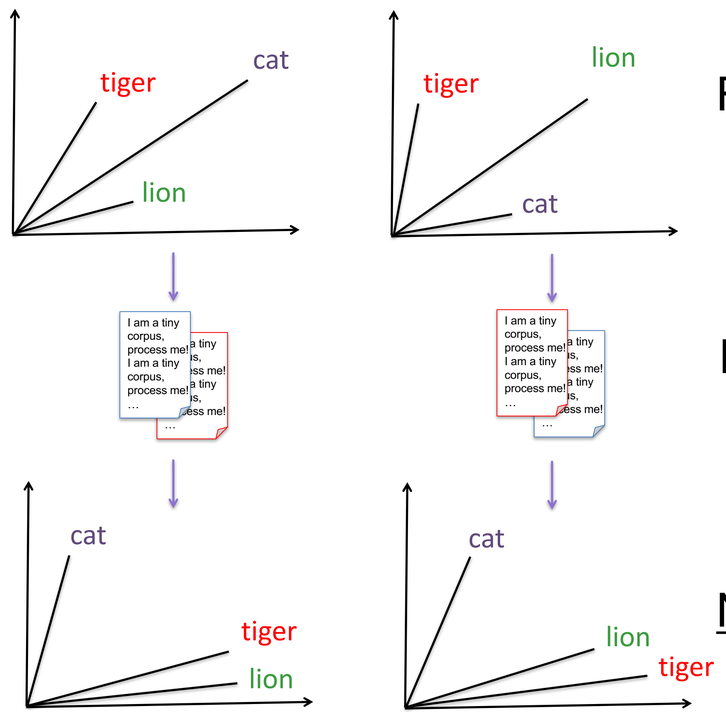
# Bad Company—
# Neighborhoods in Neural Embedding Spaces Considered Harmful


**Johannes Hellrich[1,2] & Udo Hahn[2]**

1 Graduate School "The Romantic Model. Variation – Scope – Relevance"
2 Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena, Germany


## Randomness and Neural Word Embeddings




Random initialization

Random text processing

Non-identical embeddings

**Unreliable, problem for qualitative interpretation!**

| Word | Disputed Closest Neighbor |
|---|---|
| mouse | mice, rat, cat |
| cock | cocks, arty, hen |
| ass | atheist, fool, fool |
| toilet | ironing, dressing, dressing |

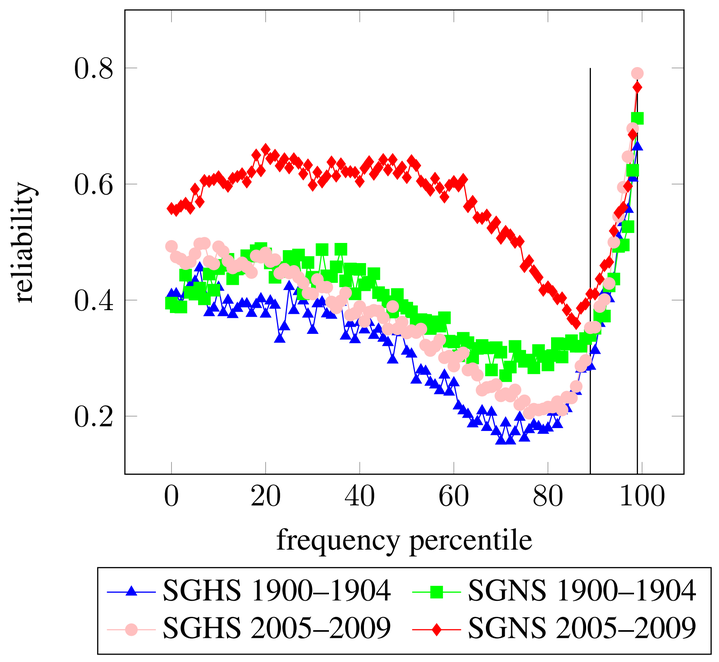
- Used in multiple papers on diachronic semantic change (e.g., Hellrich & Hahn, *Digital Humanities 2016*, pp. 545–547, 2016).
- Also used for investigating geographic variation (Kulkarni et al., *ICWSM-16*, pp. 615–618, 2016).

## Quantifying Reliability by Comparing Models

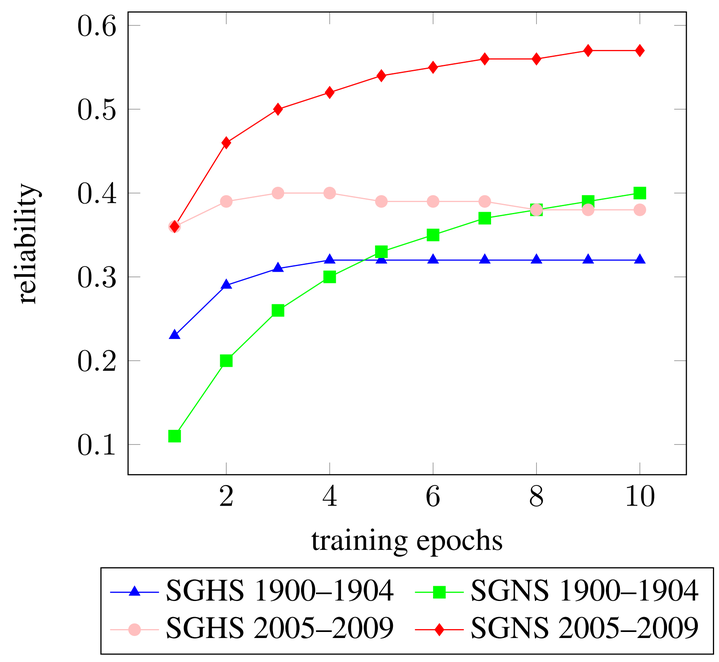
- Trained three models each on Google Books 5-gram sub-corpora.
- Models are skip-gram with negative sampling (**SGNS**) and hierarchical softmax (**SGHS**).
- Reliability = percentage of identical neighbors between models, averaged over all words.

$$r@n := \frac{1}{t * n} \sum_{j=1}^{t} || \bigcap_{k=1}^{3} \{W_{1 \leq i \leq n, j, k}\} ||$$

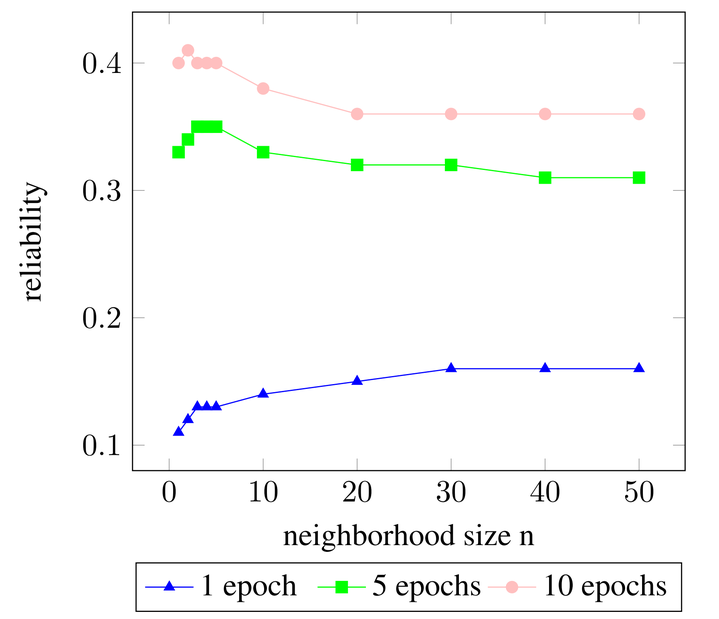
| Training Scenario | | | Top-1 Reliability | | | Similarity Accuracy | | |
|---|---|---|---|---|---|---|---|---|
| Language | Time Span | Embeddings | 1 Epoch | 5 Epochs | 10 Epochs | 1 Epoch | 5 Epochs | 10 Epochs |
| English Fiction | 1900–1904 | SGNS | 0.11 | 0.33 | 0.40 | 0.45 | 0.51 | 0.51 |
| | | SGHS | 0.23 | 0.33 | 0.33 | 0.46 | 0.45 | 0.45 |
| | 2005–2009 | SGNS | 0.36 | 0.54 | 0.57 | 0.58 | 0.58 | 0.57 |
| | | SGHS | 0.36 | 0.39 | 0.38 | 0.55 | 0.52 | 0.52 |
| German | 1900–1904 | SGNS | 0.20 | 0.47 | 0.54 | 0.45 | 0.56 | 0.56 |
| | | SGHS | 0.34 | 0.43 | 0.42 | 0.48 | 0.49 | 0.47 |
| | 2005–2009 | SGNS | 0.31 | 0.50 | 0.53 | 0.51 | 0.54 | 0.54 |
| | | SGHS | 0.34 | 0.38 | 0.36 | 0.49 | 0.48 | 0.47 |
| Normalized German | 1900–1904 | SGNS | 0.19 | 0.45 | 0.52 | 0.47 | 0.55 | 0.57 |
| | | SGHS | 0.32 | 0.42 | 0.42 | 0.47 | 0.48 | 0.48 |
| | 2005–2009 | SGNS | 0.30 | 0.48 | 0.52 | 0.54 | 0.59 | 0.60 |
| | | SGHS | 0.33 | 0.37 | 0.36 | 0.51 | 0.52 | 0.52 |




Influence of frequency on reliability. Lines mark words known to have changed semantically. English Fiction data.




Influence of the number of training epochs on reliability. Models trained on English Fiction.




Influence of neighborhood size on reliability. SGNS models trained on 1900-1904 English Fiction.

## Recommendations

- Be skeptical when confronted with qualitative interpretations / illustrations based on neighbors in neural embedding spaces.
- SGNS with 4-6 epochs is the best compromise if neural embeddings need to be used, SGHS are beneficial if only a single epoch of training is possible.
- Avoid subsampling, process complete corpora (see Hellrich & Hahn, *LaTeCH @ ACL 2016, pp. 111–117, 2016).*
- Seriously consider using a modified SVD approach (Levy et al., TACL 3:211–225, 2015) instead of neural embeddings—it was shown to be viable for diachronic analysis (Hamilton et al., ACL 2016, pp. 1489–1501, 2016) and seems to be unaffected by reliability problems.


This research was conducted within the Graduate School "The Romantic Model. Variation – Scope – Relevance" supported by grant GRK 2041/1 from the Deutsche Forschungsgemeinschaft (DFG). http://www.modellromantik.uni-jena.de/?lang=en