A Appendix

A.1 Clustering models and hyperparameters

We run all our experiments on Ubuntu 18.04 on hardware with RAM 32GB, AMD Ryzen 7 3700X, and 8 physical cores. One full pass over all models takes approximately 23 hours. For each model, we apply 300 trials using random search. We now present all hyperparameter bounds for each clustering model used.

Affinity Propagation uses the default sklearn implementation:

```
[{
        "name": "damping",
        "type": "range",
        "bounds": [0.5, 0.999]
    },
    {
        "name": "preference",
        "type": "range",
        "bounds": [-200, 10]
    },
    {
        "name": "max_iter",
        "type": "range",
        "bounds": [100, 500]
}]
```

The best parameter configuration is:

```
{
    'damping': 0.999,
    'preference': -179,
    'max_iter': 480
}
```

Chinese Whispers uses the NetworkX (Hagberg et al., 2006) implementation

```
[{
        "name": "std_multiplier",
        "type": "range",
        "bounds": [-3., 3.],
    },
    {
        "name": "remove_hub_number",
        "type": "range",
        "bounds": [
          0,
          min(
            200,
            int(
              max_samples * 0.6
            )
          )
        ]
    },
    {
        "name": "min_cluster_size",
        "type": "range",
        "bounds": [1, 50]
}]
```

The best parameter configuration is:

```
{
    'std_multiplier': 1.3972,
    'remove_hub_number': 0,
    'min_cluster_size': 31
}
```

DBScan uses the default sklearn implementation:

```
[{
        "name": "min_samples",
        "type": "choice",
        "values": [
          (2 ** x)
          for x in range(1, 5)
        ],
    },
    {
        "name": "eps",
        "type": "range",
        "bounds": [0.01, 5.]
    },
    {
        "name": "metric",
        "type": "choice",
        "values": [
            'cosine', 'braycurtis',
            'canberra', 'chebyshev',
            'correlation'
        ]
}]
```

The best parameter configuration is:

```
{
    'eps': 4.624876515865326,
    'min_samples': 4,
    'metric': 'chebyshev'
}
```

HDBScan uses the standard hdbscan [6] package implementation:

```
[{
        "name": "min_samples",
        "type": "choice",
        "values": [
          (2 ** x) for x in range(1, 5)
        ],
    },
    {
        "name": "eps",
        "type": "range",
        "bounds": [0.01, 5.]
    },
    {
        "name": "metric",
        "type": "choice",
        "values": [
            'cosine', 'braycurtis',
            'canberra', 'chebyshev',
            'correlation'
        ]
}]
```

The best parameter configuration is:

```
{
    'cluster_selection_epsilon': 4.66328,
    'alpha': 0.5626970995217562,
    'min_samples': 2,
    'min_cluster_size': 16
}
```

MeanShift uses the default sklearn implementation:

[6] hdbscan.readthedocs.io

```
[{
        "name": "bandwidth",
        "type": "choice",
        "values": [
          (x**2) for x in range(1, 5)
        ]
    },
    {
        "name": "min_bin_freq",
        "type": "choice",
        "values": [x for x in range(1, 10)]
    },
    {
        "name": "max_iter",
        "type": "choice",
        "values": [x for x in
        range(400, 1000, 500)]
}]
```

The best parameter configuration is:

```
{
    'bandwidth': 16,
    'min_bin_freq': 8,
    'max_iter': 750
}
```

A.2 Experiment: Linear separability

k	% Variance	accuracy (mean / std)
2	0.10	0.55 / 0.10
3	0.14	0.51 / 0.05
10	0.34	0.59 / 0.08
20	0.50	0.76 / 0.03
30	0.62	0.77 / 0.02
50	0.76	0.83 / 0.06
75	0.87	0.87 / 0.05
100	0.94	0.87 / 0.05

Table 4: Mean and standard deviation of the accuracy of a linear classifier trained on the 2 most common classes of WordNet meanings for the word *one*.

k	% Variance	accuracy (mean / std)
10	0.27	0.82 / 0.03
20	0.41	0.81 / 0.04
30	0.50	0.85 / 0.03
50	0.63	0.92 / 0.03
75	0.73	0.94 / 0.02
100	0.81	0.95 / 0.02

Table 5: Mean and standard deviation of the accuracy of a linear classifier trained on the 2 most common classes of WordNet meanings for the word *was*.

k	% Variance	accuracy (mean / std)
10	0.09	0.57 / 0.02
10	0.29	0.82 / 0.03
20	0.42	0.82 / 0.04
30	0.51	0.83 / 0.03
50	0.72	0.85 / 0.04
75	0.78	0.84 / 0.04
100	0.79	0.85 / 0.05

Table 6: Mean and standard deviation of the accuracy of a linear classifier trained on the 2 most common classes of WordNet meanings for the word *is*.

k	% variance	accuracy (mean / std)
2	0.08	0.38 / 0.03
3	0.11	0.38 / 0.04
10	0.28	0.65 / 0.03
20	0.43	0.76 / 0.04
30	0.53	0.83 / 0.03
50	0.67	0.93 / 0.01
75	0.77	0.95 / 0.01
100	0.83	0.95 / 0.01

Table 7: Mean and standard deviation of the accuracy of a linear classifier trained on the 4 most common semantic classes for the word `was`.

A.3 Experiment: Clusterability

Figure 5 shows that there is a correlation between semantics and syntax. The dominance refers to the percentage homogeneity of the majority tag inside the sampled part-of-speech annotations.

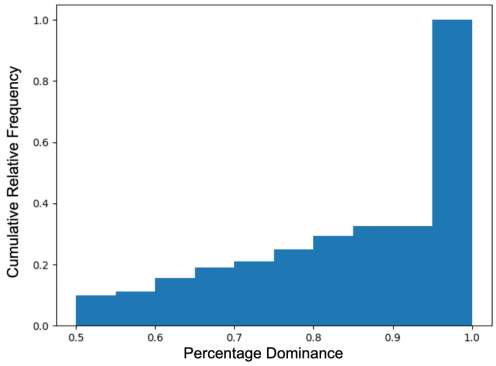

Figure 5: Cumulative dominance of the most occurring cluster. Dominance of a part of speech tag is measured by the percentage cover that the majority class intakes.

Looking at individual (Figure 6) examples, we observe that when different semantic regions also differ in part-of-speech, such a clear distinction becomes more likely, albeit still not the rule.

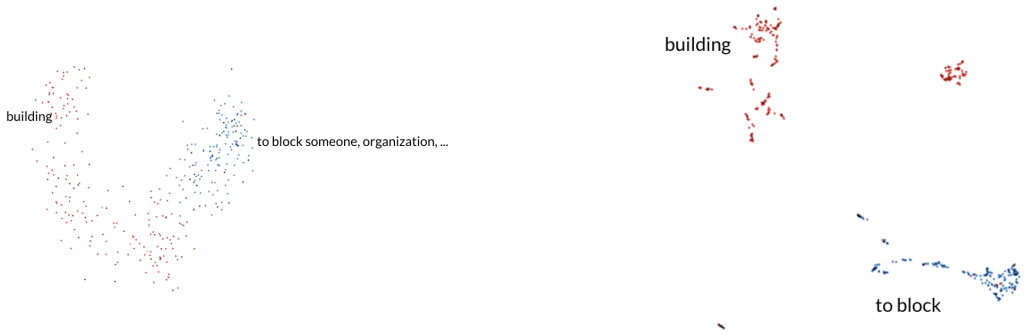

Figure 6: PCA (left) and UMAP (right) visualizations for contextual word embeddings sampled for the word `block`. Red points denote *nouns*, blue points denote *verbs*.

Partition	Sample
1	Ms. Gotbaum tried to slide her handcuffed **arms** from her back to her front . . . That magisterial use of the upper body and **arms** is her physical signature entered the stage in pairs and forcefully stretched their **arms** and legs ripped from patients' **arms** as they were carried away
2	She swooped him up into her **arms** and kissed him madly . . . I place my son in her **arms** and I pray that it somehow comforts her perfect babies . . . into the loving **arms** of middle class+ Americans which he falls back into her **arms** like a baby Sometimes I took it into my **arms** and felt its surprising heft
3	. . . and shuttle robotic **arms** of a solar array and truss . . . a contingent of young **arms** that will allow us to win now By and large, those **arms** remained as fictional as those in "The War" . . . and extensive use of robotic **arms** operating at their limits staff of strong young **arms** that might have tamed the National League East
4	The classic years of the **arms** race, the 1950s and '60s before . . . that concerns over nuclear **arms** proliferation in the Middle East Russian adherence to another **arms** control treaty he will press for peace and an eventual **arms** cut for the states payments to the companies that supplied **arms** to Iraq were often delayed Mr. Safar denies any wrongdoing, including any **arms** dealings
5	leaned back in his chair and, with **arms** crossed If you feel yourself falling, spread your **arms** Russian adherence to another **arms** control treaty Agamemnon, **arms** raised ... barely contained violence Mr. James sat with his **arms** folded, his head lowered she felt tears in her eyes and held her **arms** out in simple joy
6	this country is so polarized that people spring to **arms** against any proposal At least they are carrying **arms** to protect themselves His organization issued a call to **arms** he shoves it in the faces of his comrades in `arms` people who had taken up `arms` against the United States Mostly non-Arab rebels took up **arms** in early 2003

Table 8: Representative samples for the partitions found by the Chinese Whispers clustering model for the word `arms`. Partitions 1-3 consider a person's arms, whereas partition 4 considers `arms` as a synonym to `weaponry`. Partitions 1, 2 and 3 strongly contrast in sentiment (scared, loving, and confident respectively).