# Semantic clustering in Dutch


## Tim Van de Cruys

## K.U.Leuven


**Goal:** Automatically clustering nouns by applying machine learning techniques
**Basic approach:** Inducing semantic classes of nouns according to the adjectives those nouns collocate with
**Hypothesis:** Syntactic context is a sufficient cue for semantic clustering

## 1 Calculating semantic similarity

Take a word and its contexts:

| | |
|---|---|
| *verse* 'fresh' | sneup |
| *gezouten* 'salted' | sneup |
| *lekkere* 'tasty' | sneup |
| *zoete* 'sweet' | sneup |
| *pikante* 'spicy' | sneup |

- it can be inferred from the context that *sneup* is some kind of **food**
- in the same way, a computer might be able to discover semantically similar words

**How to determine semantic similarity computationally?**

### 1 Create vectors

| | red | tasty | fast | second-hand |
|---|---|---|---|---|
| **apple** | 2 | 1 | 0 | 0 |
| **wine** | 2 | 2 | 0 | 0 |
| **car** | 1 | 0 | 1 | 2 |
| **truck** | 1 | 0 | 1 | 1 |

### 2 Apply cosine similarity measure

- $\cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$
- Examples:
  cos(car,truck) = 0.94
  cos(apple,truck) = 0.51

## 2 Clustering

### Partitional

'Stand-alone' clusters, not embedded in a structure

### Hierarchical

Complete branching structure, up to the root node

### 1 K-Means clustering

1. Choose *k* cluster centers, which are usually *k* randomly-chosen patterns or *k* randomly defined points inside the vector space.
2. Assign each pattern to the closest cluster center (using the cosine measure).
3. Recompute the cluster centers using the current cluster memberships.
4. If a convergence criterion is met (e.g. no reassignment of patterns to new cluster centers), stop the algorithm. Otherwise, go to step 2.

### 2 Group-average agglomerative clustering

1. Take each individual pattern in the pattern set to form a cluster.
2. The two clusters which are most similar are grouped together. Most similar means: the two clusters with the smallest distance between the averages of the clusters.
3. Step two is repeated until there is only one cluster left. When the algorithm terminates, all clusters are hierarchically connected to the root node.

## 3 Results

- Adjective-noun collocations have been extracted from **Twente News Corpus** (>300M words)
- For the **5.000** most frequent nouns, vectors have been created that contain the frequency of the **20.000** most frequent adjectives

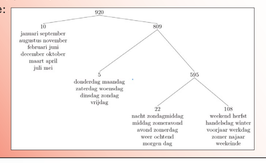
### 1 Partitional clustering

**Examples:**

- mei februari september maart december augustus oktober januari juli april november juni → months
- aanvaller speler middenvelder verdediger linksbuiten international invaller keeper voetballer doelman spits → soccer terms
- guerrillabeweging opstandeling rebellenleider guerrillastrijder guerrilla verzetsbeweging rebel bevrijdingsleger → resistance movement terms
- minuut millimeter seconde cent ton meter centimeter graad kilo kilometer → measure terms

### 2 Hierarchical clustering

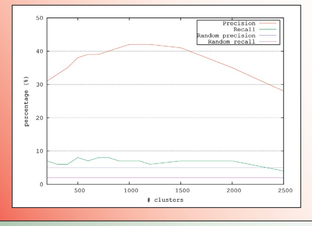
- lower clusters form tight, specific clusters
- cluster higher up in the hierarchy may form more general clusters (such as persons, places, ...)
- Simple example:




## 4 evaluation

### 1 Wordnet comparison evaluation

1. For each cluster, take the word with most semantic relations to other words in Wordnet (=most central word)
2. Get hyponyms, hypernyms, co-hyponyms and synonyms in Wordnet
3. precision: how many words in cluster have equivalent Wordnet-relation
4. recall: how many wordnet-relations have no equivalent in found cluster




### 2 Wu & Palmer evaluation

- Calculate similarity between two words according to distance in hierarchical wordnet
- Instead of having a fixed group of words to compare the clusters to, the cluster quality is calculated according to similarity in WordNet