UNIVERSITÄT LEIPZIG

erc European Research Council

Behaviour Profiles and hierarchical cluster analysis

Natalia Levshina ©2017

Corpus Linguistics Summer School

Birmingham, July 2017

Outline

1. Introduction to Behaviour Profiles and hierarchical cluster analysis

2. Univariate BP: verbs of communication

3. Multivariate BP: polysemy of SEE

4. MDS vs. MCA vs. cluster analysis: when to use which method?

Behavioural Profiles

- BP are a popular method of comparing the corpus-based distributional properties of several near synonyms or word senses.
- Based on ideas of Atkins (1987), Hanks (1996)
- Developed by Divjak (2003) and Gries (2006)

Steps of BP analysis

1. Create the distributional profiles of your units.

2. Compute distances between them.

3. Represent them visually, e.g. with the help of a cluster analysis.

Distributional profiles

Verb	Transitive	Intransitive	Clause
walk	1	99	0
think	5	50	45
believe	30	20	50

Verb	Transitive	Intransitive	Clause
walk	0.01	0.99	0
think	0.05	0.5	0.45
Believe	0.3	0.2	0.5

Steps of BP analysis

1. Create the distributional profiles of your units

2. Compute distances between them

3. Represent them visually, e.g. with the help of a cluster analysis

Canberra distances

- sum of all $|x - y|/|x + y|$ for each column
- E.g. *think* and *walk*:
    - Transitive: $|0.05 - 0.01|/|0.05 + 0.01| = 0.67$
    - Intransitive: $|0.5 - 0.99|/|0.5 + 0.99| = 0.33$
    - Clause: $|0.45 - 0|/|0.45 + 0| = 1$
    - Total: $0.67 + 0.33 + 1 = 1.99$

```
> dist(rbind(walk, think, believe), method =
"canberra")
            walk    think
think   1.995526
believe 2.599349 1.195489
```

Steps of BP analysis

1. Create the distributional profiles of your units.

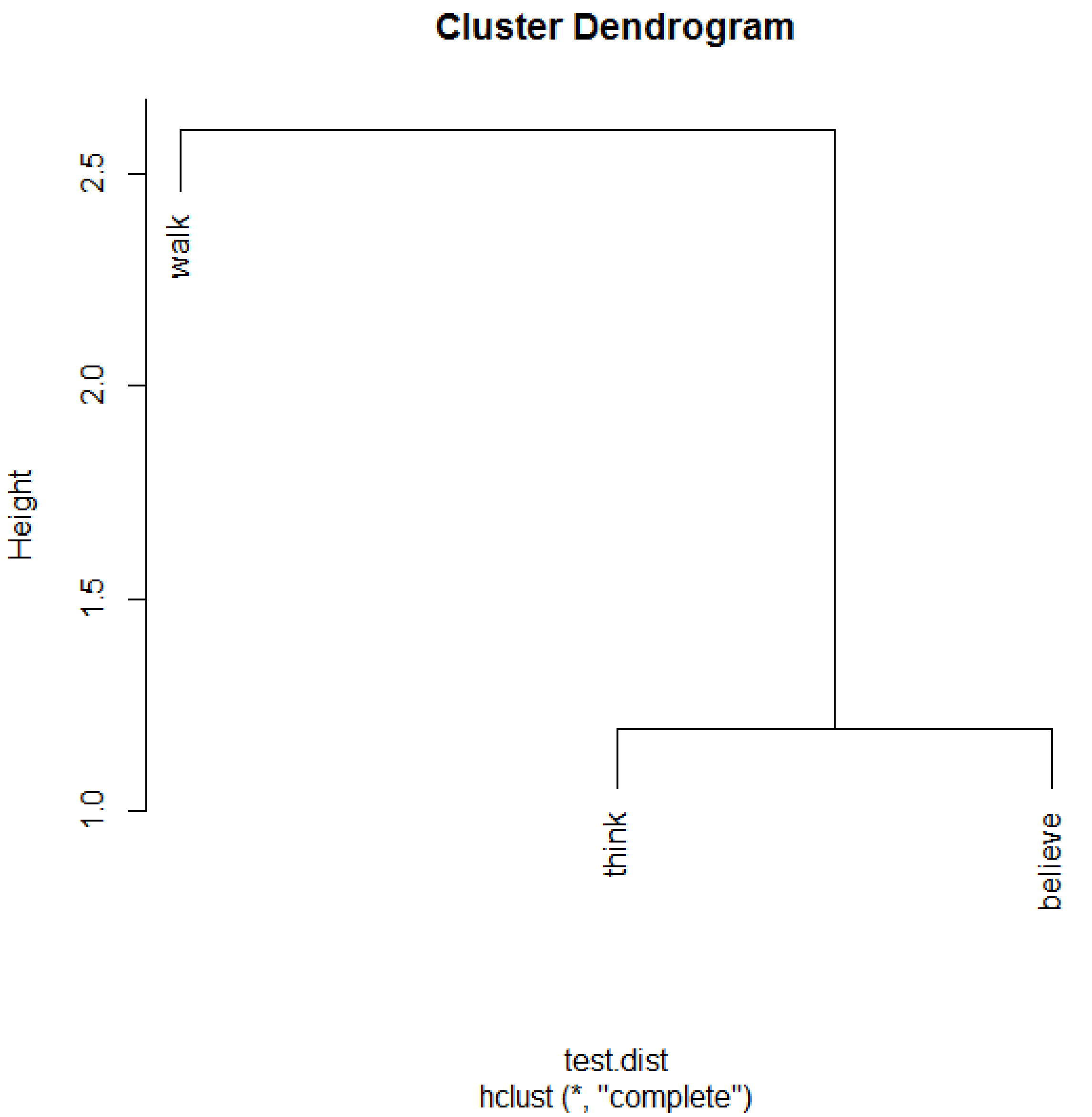
2. Compute distances between them.

3. Represent them visually, e.g. with the help of a cluster analysis.

Hierarchical cluster analysis

- Begins with a distance matrix, like MDS
- Picks the small distance between two objects and merges them in one cluster.
- Then picks the next smallest distance between two objects and/or clusters and merges them.
- Stops when all objects are merged in one cluster tree.

```
> plot(hclust(test.dist))
```


Cluster Dendrogram
Height
2.5
2.0
1.5
1.0
walk
think
believe
test.dist
hclust (*, "complete")

Outline

1. Introduction to Behaviour Profiles and hierarchical cluster analysis

2. Univariate BP: verbs of communication

3. Multivariate BP: polysemy of SEE

4. MDS vs. MCA vs. cluster analysis: when to use which method?

Verbs of communication revisited

1. Transform the frequencies into proportions (row sums = 1):

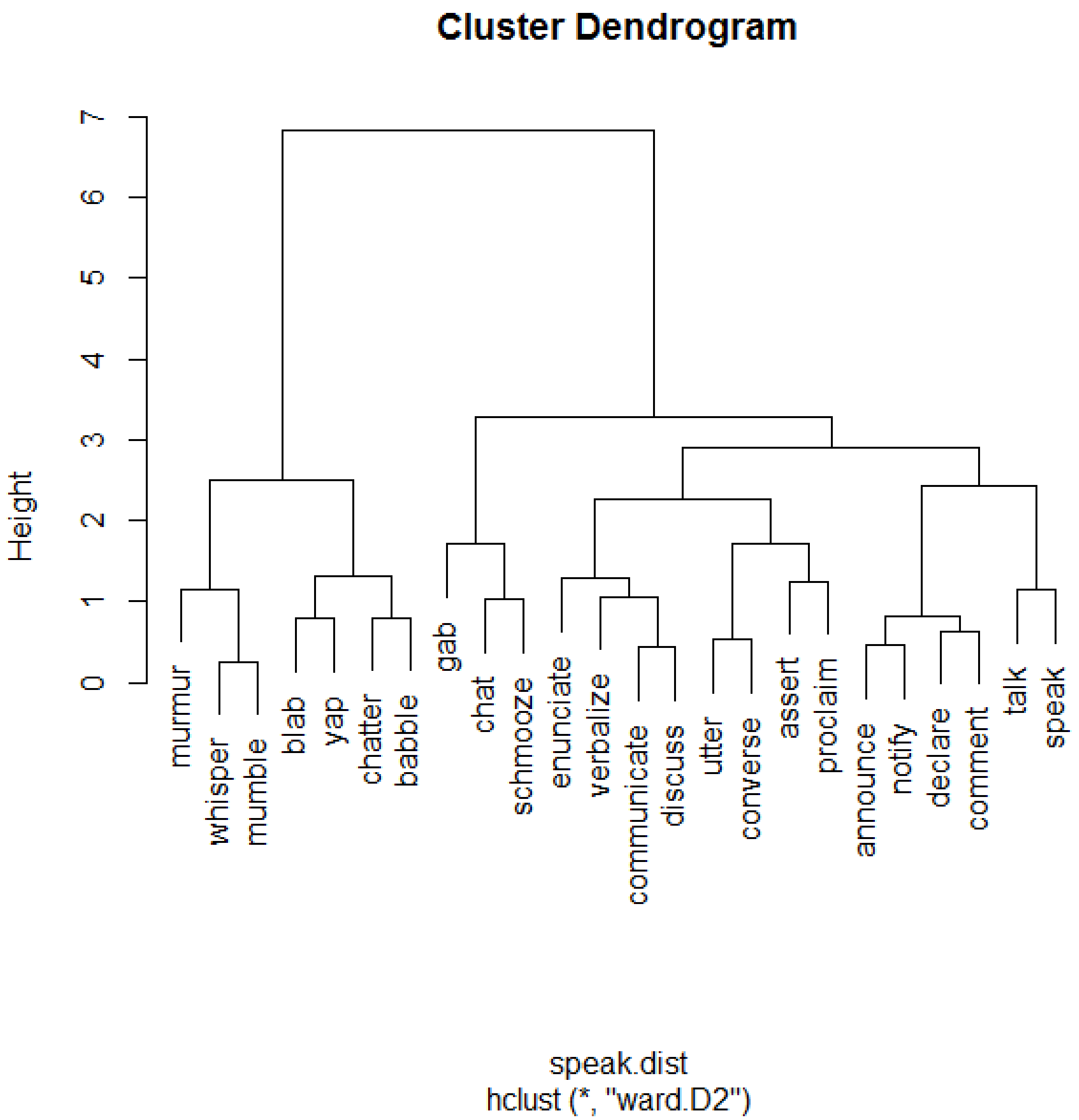
```
> speak.bp <- prop.table(as.matrix(speak), 1)
```

2. Compute the distances between the verbs:

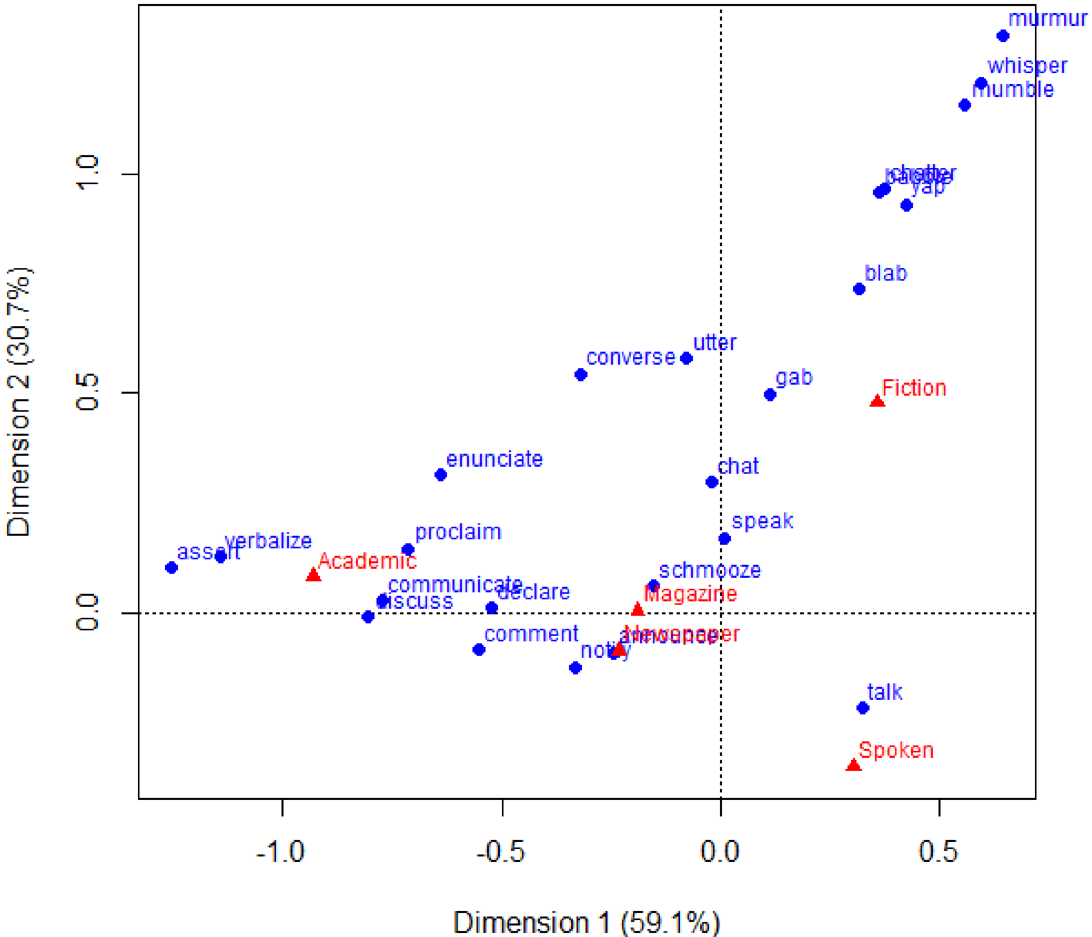
```
> speak.dist <- dist(speak.bp, method = "canberra")
```

3. Perform a hierarchical cluster analysis:

```
> see.clust <- hclust(speak.dist, method = "ward.D2")
> plot(see.clust))
```


Cluster Dendrogram
Height
0
1
2
3
4
5
6
7
murmur
whisper
mumble
blab
yap
chatter
babble
gab
chat
schmooze
enunciate
verbalize
communicate
discuss
utter
converse
assert
proclaim
announce
notify
declare
comment
talk
speak
speak.dist
hclust (*, "ward.D2")

Compare: simple CA (Dimensions 1 & 2)



Outline

1. Introduction to Behaviour Profiles and hierarchical cluster analysis

2. Univariate BP: verbs of communication

3. Multivariate BP: polysemy of SEE

4. MDS vs. MCA vs. cluster analysis: when to use which method?

Data: some senses of SEE

```
> str(see)
'data.frame':  88 obs. of  7 variables:
 $ Sense  : Factor w/ 6 levels "be_characterized",..: 1 1 1 1 1
1 1 1 1 1 ...
 $ SubjSem: Factor w/ 2 levels "Abstr","Hum": 1 1 1 1 1 1 1 1 1
1 ...
 $ Val    : Factor w/ 4 levels "DO","DO_Ved",..: 1 1 1 1 1 1 2
1 1 1 ...
 $ ObjSem : Factor w/ 3 levels "Abstr","Hum",..: 1 1 1 1 1 1 1
1 1 2 ...
 $ Morph  : Factor w/ 5 levels "Gerund","Imper",..: 4 4 4 4 4 4
4 4 4 4 ...
 $ ObjDef : Factor w/ 4 levels "Def","Indef",..: 3 3 3 4 3 3 4
4 3 3 ...
 $ Voice  : Factor w/ 2 levels "Act","Pass": 1 1 1 1 1 1 1 1 1
1 ...
```

Creating BP for many variables

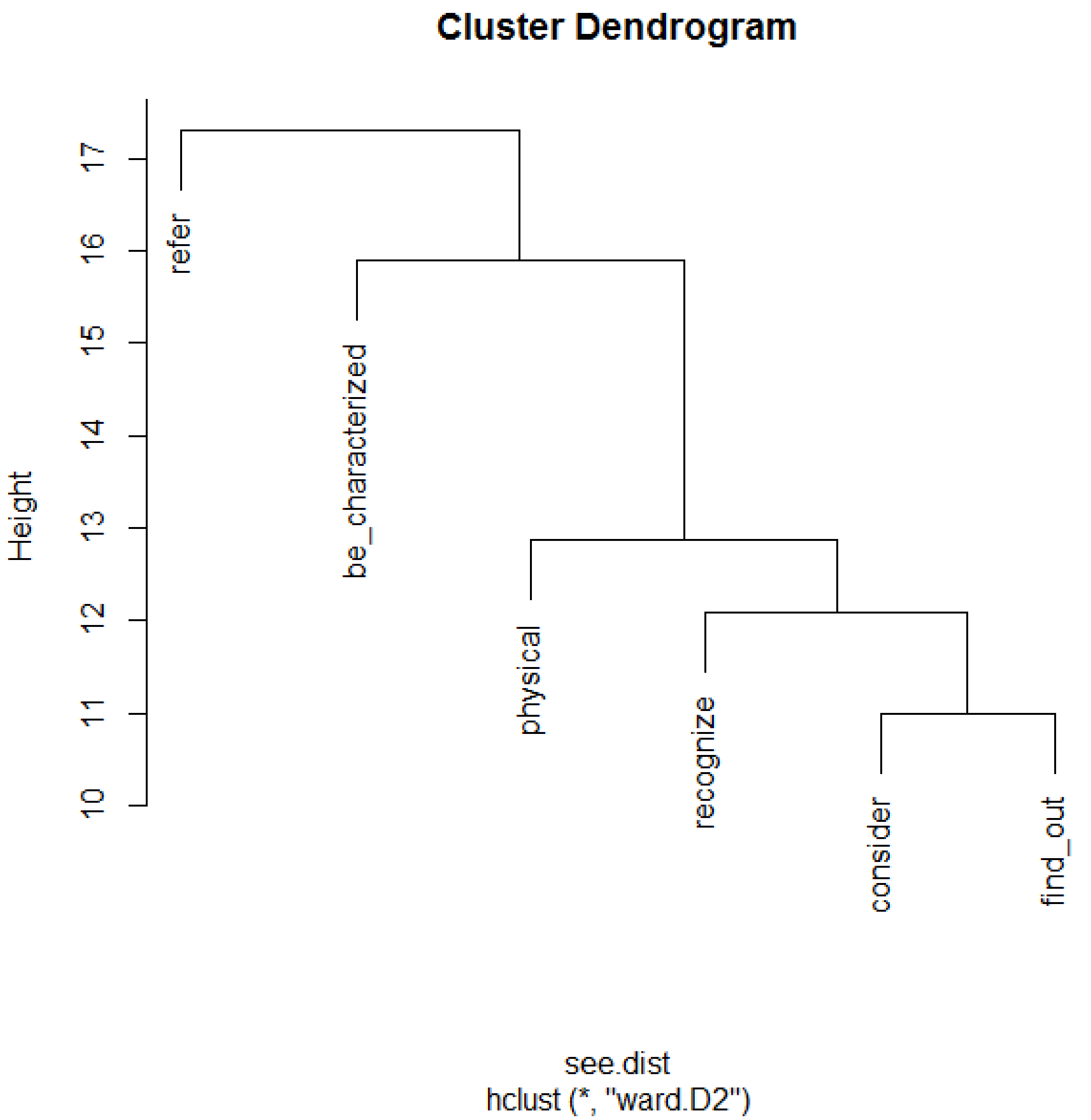
```
> see.split <- split(see, see$Sense)
> see.split <- lapply(see.split, function(x) x =
x[, -1])
> see.split.bp <- lapply(see.split, bp)
> see.bp <- do.call(rbind, see.split.bp)
```

Clustering the senses

```
> see.dist <- dist(see.bp, method = "canberra")

> see.clust <- hclust(see.dist, method =
"ward.D2")

> plot(see.clust)
```


Cluster Dendrogram
Height
10
11
12
13
14
15
16
17
refer
be_characterized
physical
recognize
consider
find_out
see.dist
hclust (*, "ward.D2")

Outline

1. Introduction to Behaviour Profiles and hierarchical cluster analysis

2. Univariate BP: verbs of communication

3. Multivariate BP: polysemy of SEE

4. MDS vs. MCA vs. cluster analysis: when to use which method?

Rules of thumb

- If you expect to find distinct clusters in the data, use cluster analysis.
- If you expect a continuous distribution along some dimensions, use MDS or CA.
    - If you have relatively few variables (e.g. 2 to 5), use CA.
    - If you have many missing values or values with very small counts, use Gower distances and MDS.

References

- Atkins, B.T.S. (1987). Semantic ID tags: Corpus evidence for dictionary senses. The uses of large text databases. *Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary,* 17–36. Waterloo, Canada.
- Divjak, D. (2003). On trying in Russian: A tentative network model for near(er) synonyms. In *Belgian Contributions to the 13th International Congress of Slavicists*, Ljubljana, 15–21 August 2003. Special issue of *Slavica Gandensia,* 25–58.
- Gries, S. Th. (2006). Corpus-based methods and Cognitive Semantics: The many senses of *to run*. In S. Th. Gries & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics. Corpus-based Approaches to Syntax and Lexis,* 57–99. Berlin/New York: Mouton de Gruyter.
- Hanks, P. (1996). Contextual dependency and lexical sets. *International Journal of Corpus Linguistics* 1(1). 75–98.