UNIVERSITÄT LEIPZIG

European Research Council

Behaviour Profiles and hierarchical cluster analysis

Natalia Levshina ©2017

Summer School of Linguistics

Litomyšl, August 2017

Outline

1. Introduction to Behaviour Profiles and hierarchical cluster analysis

2. Univariate BP: verbs of communication

3. Multivariate BP: polysemy of SPEAK

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

Manhattan distances

- The simplest method: take the absolute differences between *a* and *b* and sum them up.
- E.g. *think* and *walk*:

```
> abs(0.01 - 0.05) + abs(0.99 - 0.5) + abs(0 - 0.45)
[1] 0.98
```

Manhattan distances with R

```
> dist(test, method = "manhattan")
         walk think
think    0.98
believe 1.58  0.60
```

Euclidean distances

- In real life, represents the distance between objects ‘as the crow flies’.
- Euclidean distance is computed as the square root of the sum of the squared differences between *a* and *b* in each column.
- Cf. Pythagoras’ law!
- E.g. *think* and *walk*:

```
> sqrt((0.01 - 0.05)^2 + (0.99 - 0.50)^2 + (0 - 0.45)^2)
[1] 0.6664833
```

Euclidean distances with R

```
> dist(test) #the default option
             walk       think
think     0.6664833
believe   0.9788769  0.3937004
```

Canberra distances

- sum of all $|a - b|/|a + b|$ for each column (i.e. absolute differences and absolute sums)
- 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: $0.67 + 0.33 + 1 = 1.99$

Canberra distances with R

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

Steps of BP analysis

1. Create the distributional profiles of your units.

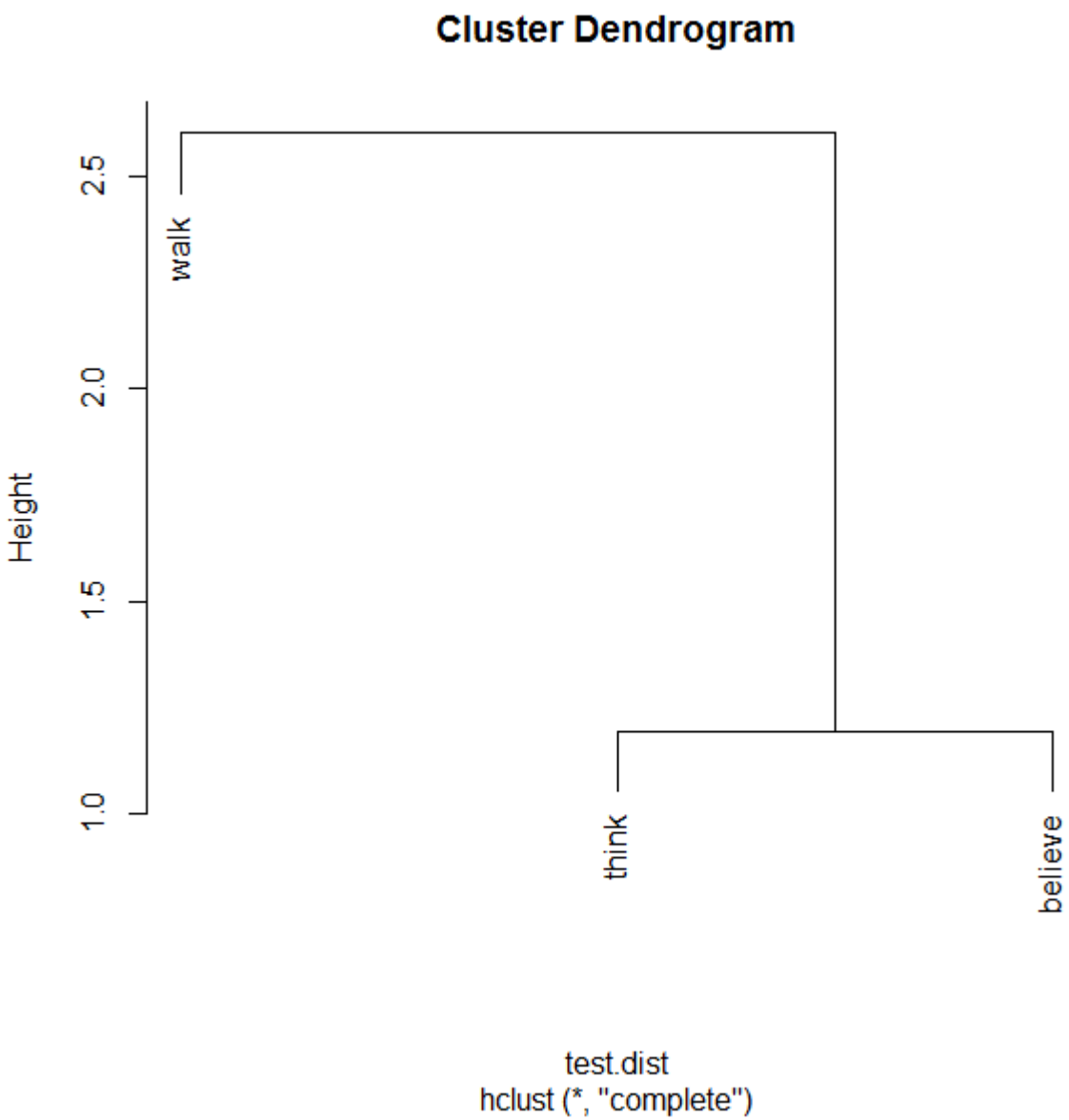
2. Compute distances between them.

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

Hierarchical cluster analysis

- Take a distance matrix.
- Pick the smallest distance between two objects and merge them in one cluster.
- Then pick the next smallest distance between two objects and/or clusters and merge 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")

Clustering methods

- OK, but how to compute distances between clusters?
  - Single (the minimally possible distances between the clusters are compared, and the smallest is taken)
  - Complete (the maximally possible distances between clusters are compared, and the smallest is taken)
  - Average (the average distances between the clusters are computed, and the smallest is taken)
  - Ward (based on variance minimization)

A simple analogy

- Imagine you have a choice between two clubs. How do you decide which one to join?
- Single: you choose the club your best friend has joined.
- Complete: you find out which club your biggest enemy is a member of. You choose the other one.
- Average: you choose the club which has on average more likable members.

What is your own social strategy?

Outline

1. Introduction to Behaviour Profiles and hierarchical cluster analysis

2. Univariate BP: verbs of communication

3. Multivariate BP: polysemy of SPEAK

Verbs of communication revisited

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

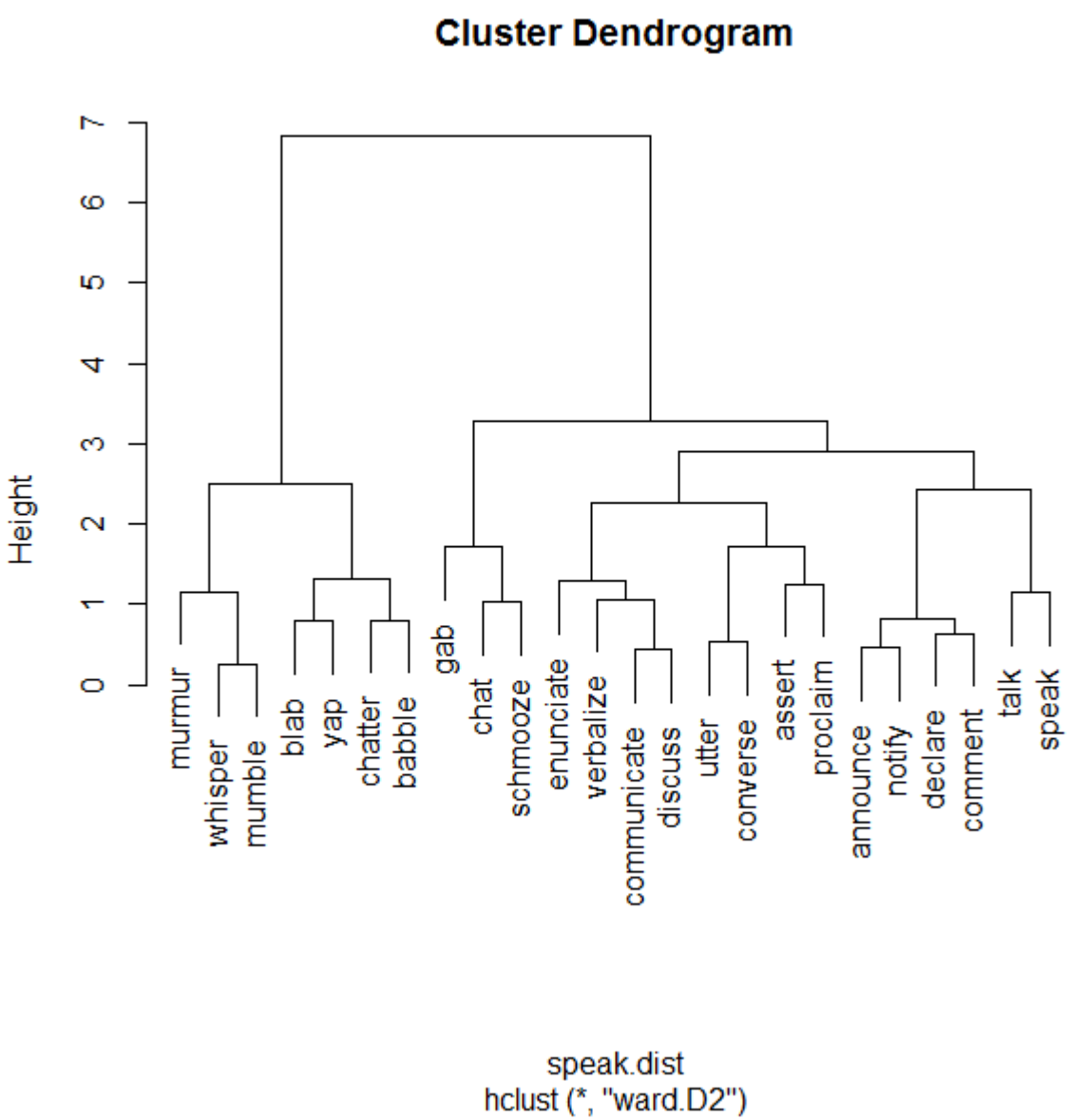
```
> speak.bp <- prop.table(as.matrix(speak), 1)
```

2. Compute the distances between the verbs:

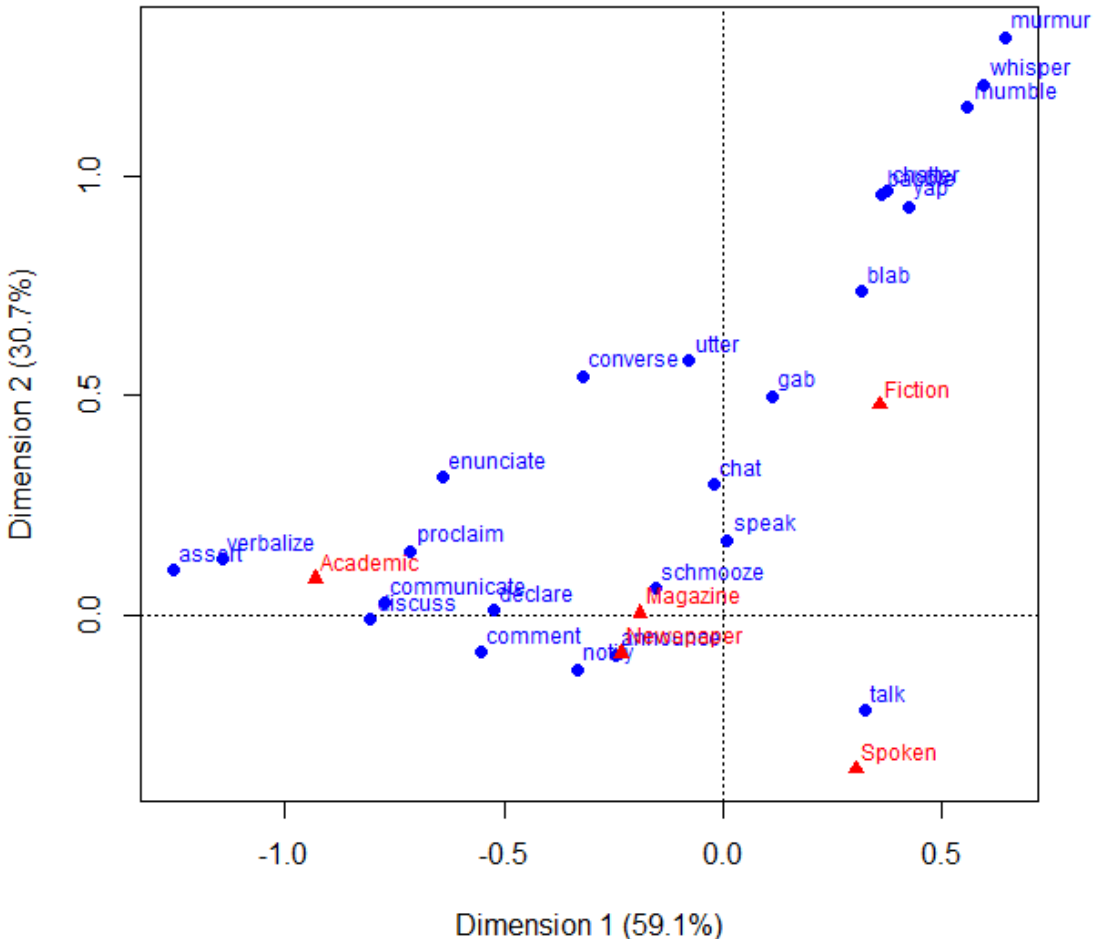
```
> speak.dist <- dist(speak.bp, method =
"canberra")
```

3. Perform a hierarchical cluster analysis:

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

Data: SPEAK from COCA

```
> str(speak_poly)
'data.frame':  120 obs. of  9 variables:
 $ Sense    : Factor w/ 12 levels "act_spokesperson",..: 1 1 1
1 1 1 1 1 1 1 ...
 $ SubjSem  : Factor w/ 2 levels "Hum","NonHum": 1 1 1 1 1 1 1
1 1 1 ...
 $ Subcorpus: Factor w/ 3 levels "Acad","Fiction",..: 1 1 1 2 2
3 3 3 3 3 ...
 $ Form     : Factor w/ 5 levels "speak","speaking",..: 2 4 4 1
1 3 1 3 1 4 ...
```

…

12 senses of SPEAK

1. Produce words

    Cats and dogs can't speak.

2. Express thoughts or feelings about something

    He spoke at one point about whether or not this crisis could be resolved between Arabs or between Iraq and the rest of the world.

3. Touch upon a topic, mention, bring up

    Have I spoken of her looks?

4. Regarding, as for

    Speaking of lust, let's hit the trail.

12 senses of SPEAK (cont.)

5. Characterize, define smth/smb as smth/smb

   He frequently spoke of the Roman communion as "the only true Church…"

6. Act as a spokesperson

   Ladies and gentlemen, I, your captain, speak for the crew of the S.S. Ariel, bidding you farewell.

7. Engage in a conversation with smb.

   I spoke with you this morning.

8. Speak publicly

   Speaking in Colombia, the president added new priorities.

12 senses of SPEAK (cont.)

9. Tell

   They'll see me as a man who speaks the truth and has suffered on.

10. Pronounce

    I'm not sure he spoke my name.

11. Have command of a language

    Jakov had lived in America for thirty years but he did not speak English.

12. Express non-verbally

    …the pseudosmile around the corners of his mouth spoke volumes.

How would you cluster the senses?

Creating BP for many variables

First, split the data frame into a list with data for individual senses:

```
> speak.split <- split(speak_poly,
speak_poly$Sense)
```

Remove the first column with senses (no longer necessary):

```
> speak.split <- lapply(speak.split, function(x) x
= x[, -1])
```

Create the BP vectors for each of the senses:

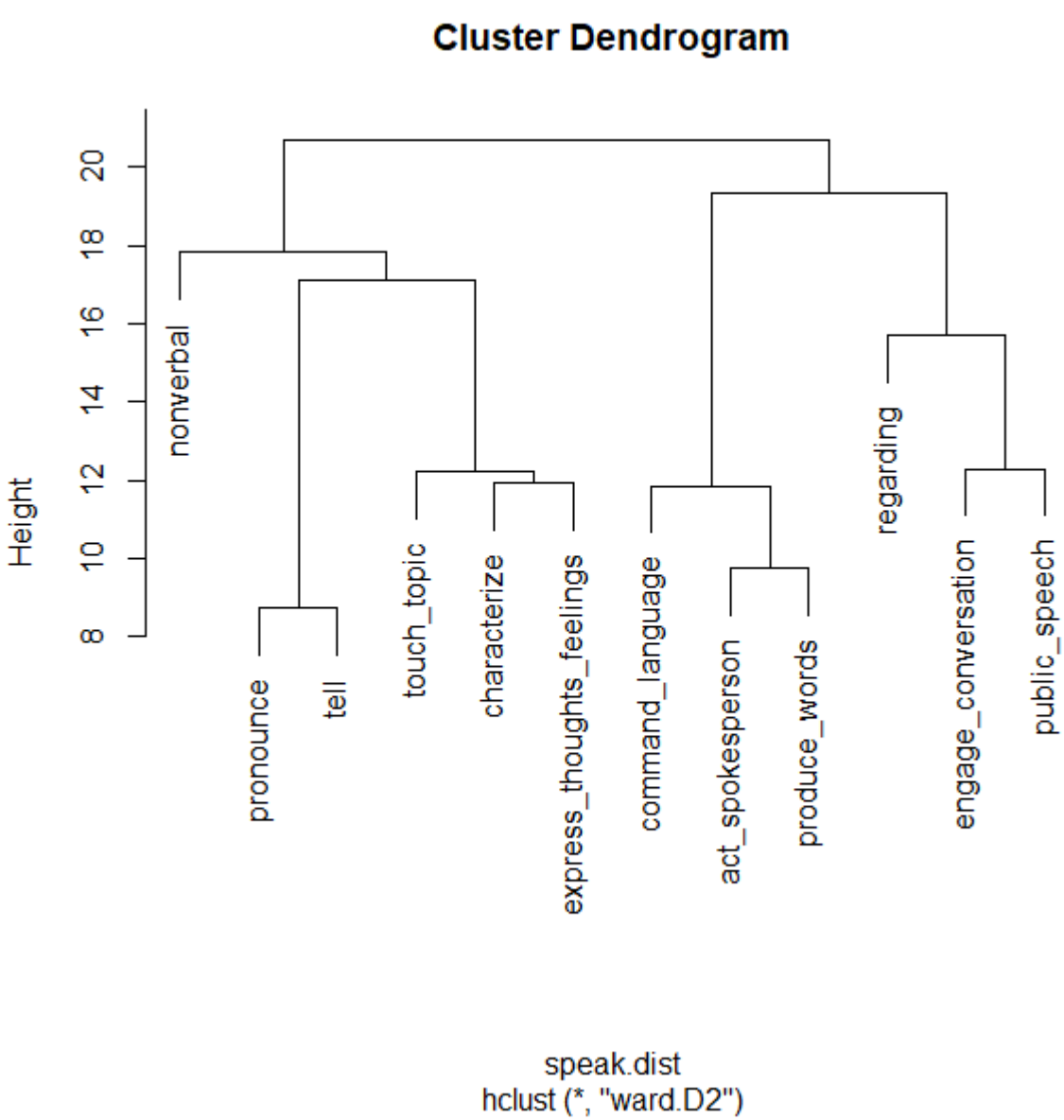
```
> speak.split.bp <- lapply(speak.split, bp)
```

Put the vectors in one matrix:

```
> speak.bp <- do.call(rbind, speak.split.bp)
```

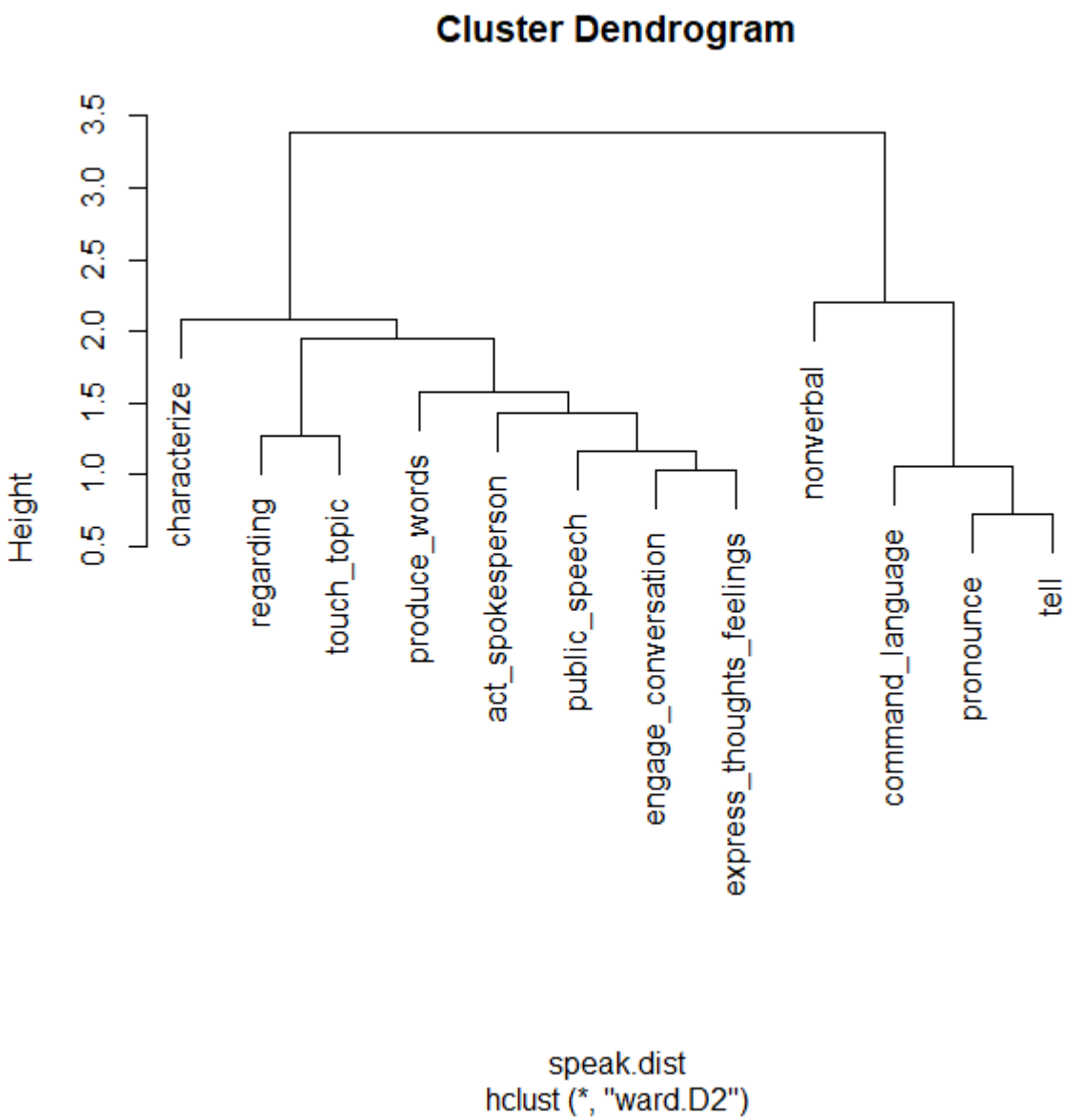
Clustering the senses: Canberra distances, Ward clustering

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


Cluster Dendrogram
Height
8
10
12
14
16
18
20
nonverbal
pronounce
tell
touch_topic
characterize
express_thoughts_feelings
command_language
act_spokesperson
produce_words
regarding
engage_conversation
public_speech
speak.dist
hclust (*, "ward.D2")

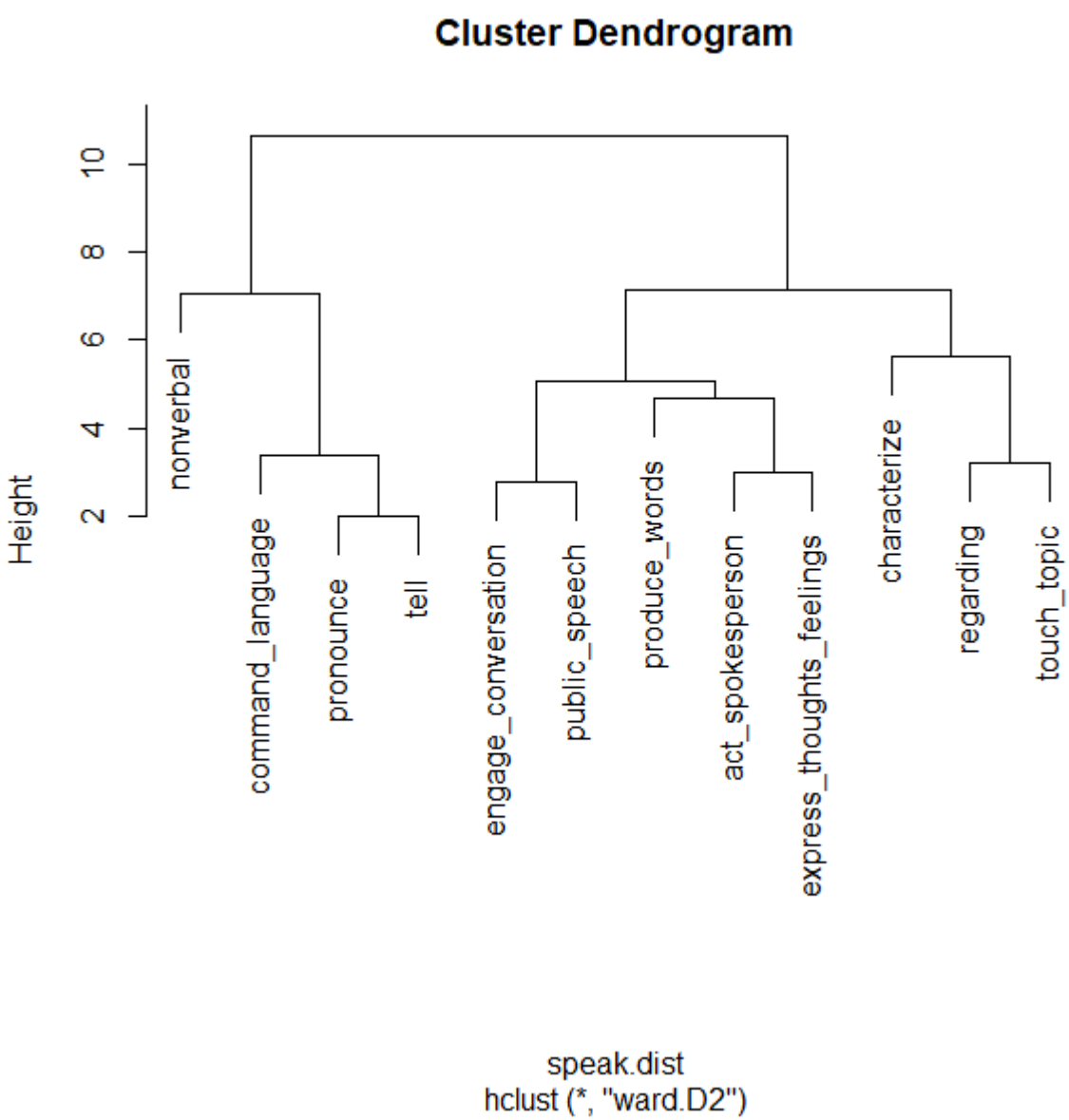
Trying other options: Euclidean distance, Ward clustering

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


Cluster Dendrogram
Height
0.5
1.0
1.5
2.0
2.5
3.0
3.5
characterize
regarding
touch_topic
produce_words
act_spokesperson
public_speech
engage_conversation
express_thoughts_feelings
nonverbal
command_language
pronounce
tell
speak.dist
hclust (*, "ward.D2")

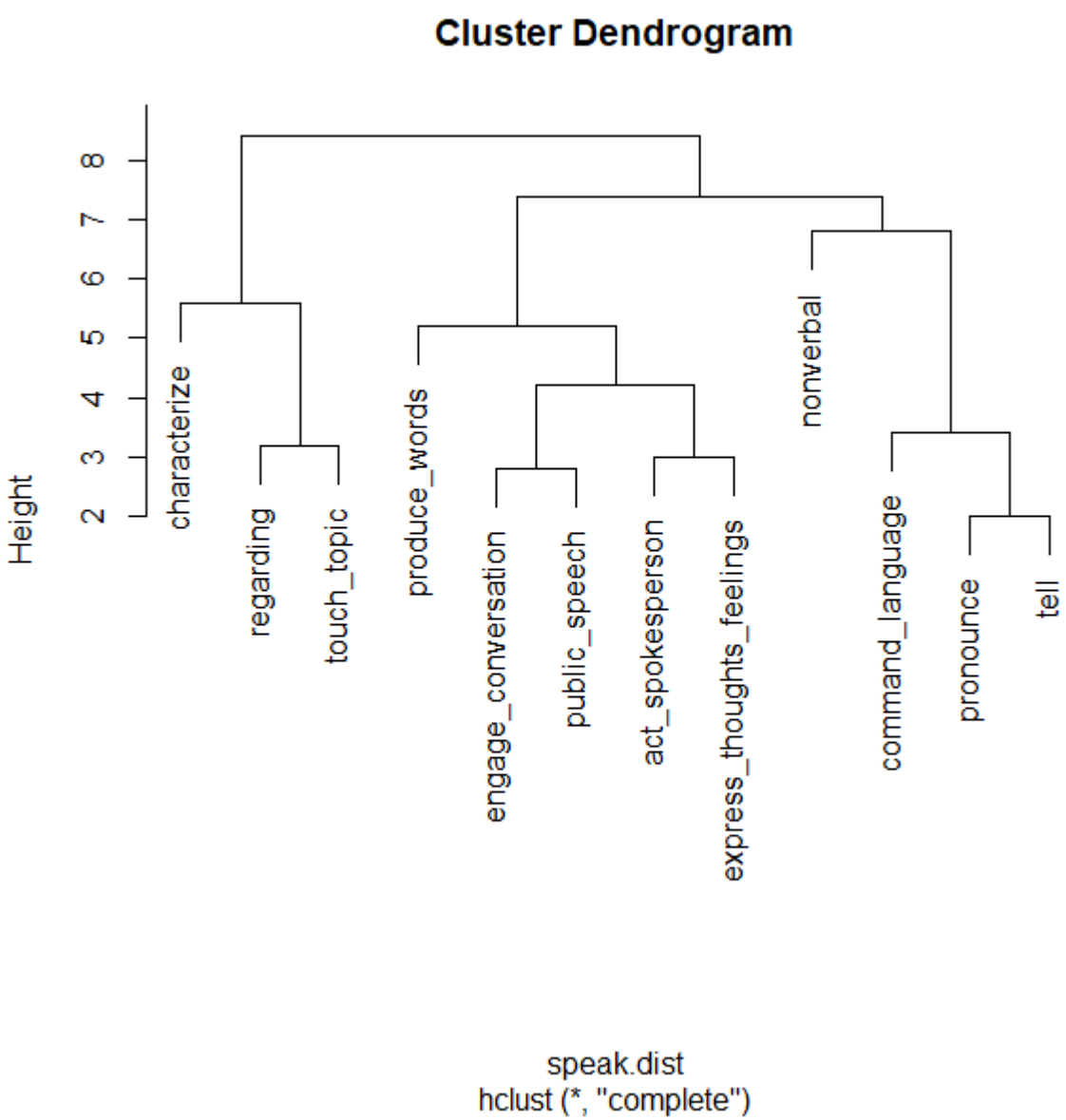
Trying other options: Manhattan distance, Ward clustering

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


Cluster Dendrogram
Height
2
4
6
8
10
nonverbal
command_language
pronounce
tell
engage_conversation
public_speech
produce_words
act_spokesperson
express_thoughts_feelings
characterize
regarding
touch_topic
speak.dist
hclust (*, "ward.D2")

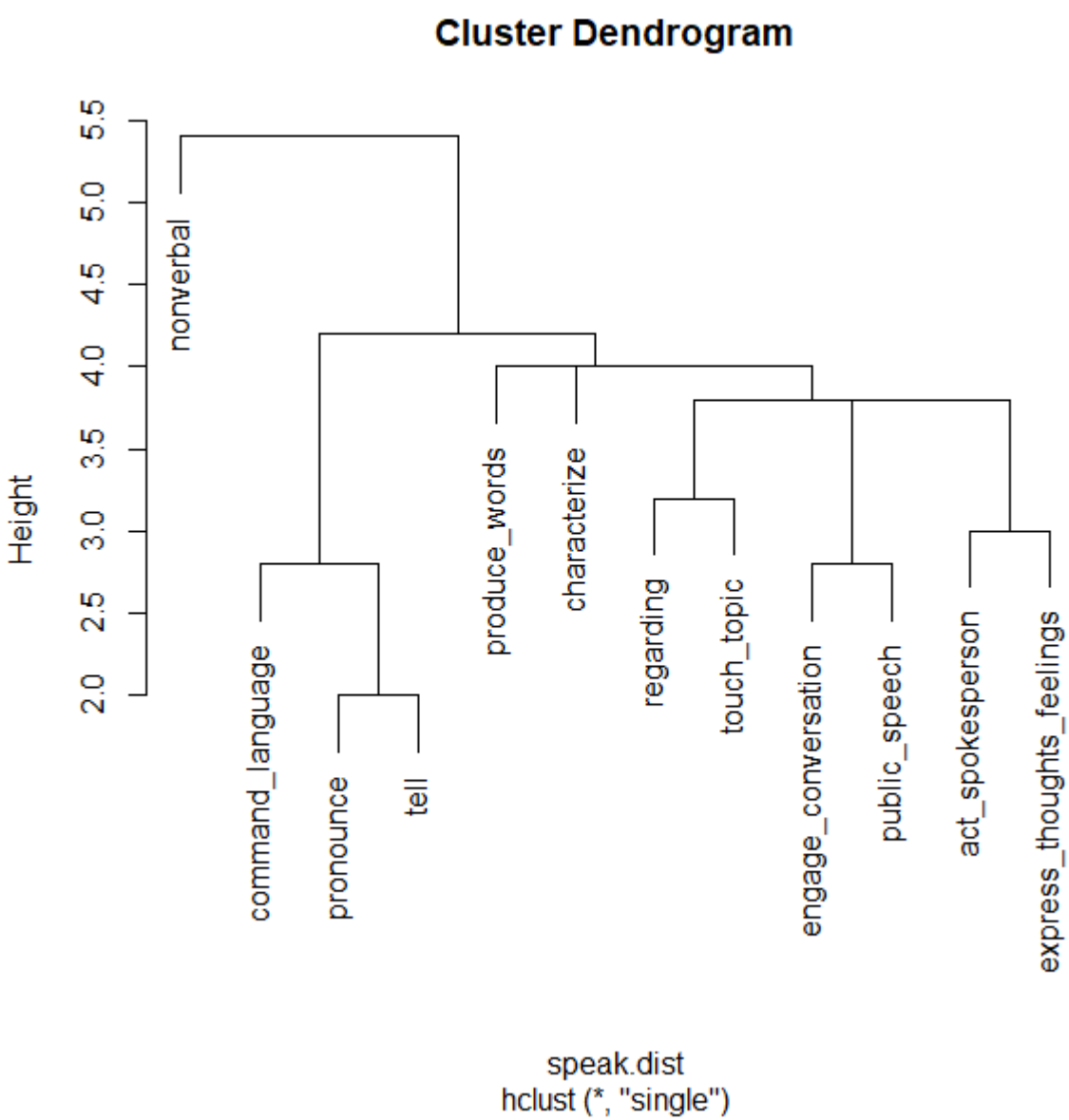
Trying other options: Manhattan distances, complete clustering

```
> speak.dist <- dist(speak.bp, method =
"manhattan")
> speak.clust <- hclust(speak.dist, method =
"complete")
> plot(speak.clust)
```


Cluster Dendrogram
Height
2
3
4
5
6
7
8
characterize
regarding
touch_topic
produce_words
engage_conversation
public_speech
act_spokesperson
express_thoughts_feelings
nonverbal
command_language
pronounce
tell
speak.dist
hclust (*, "complete")

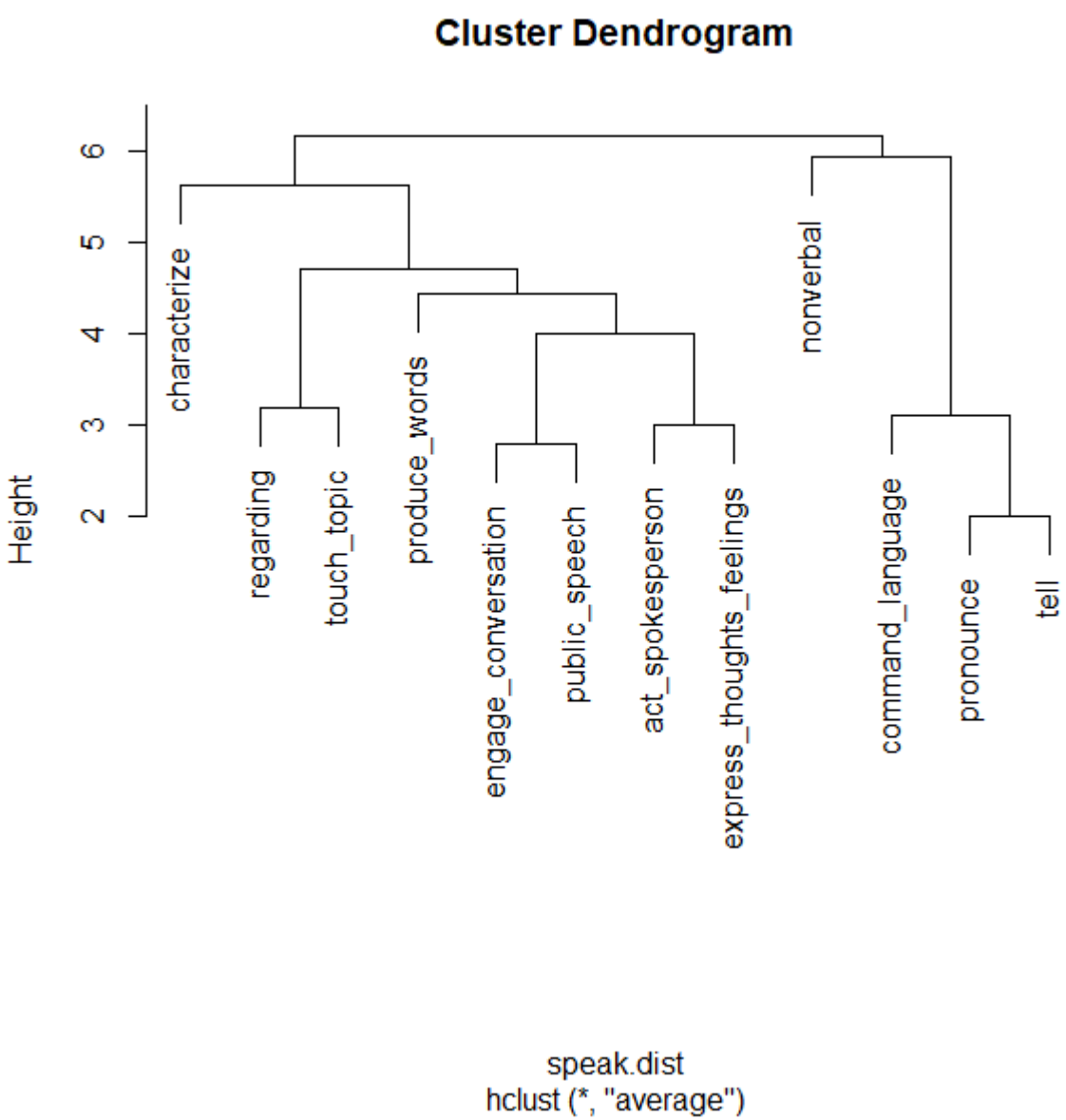
Trying other options: Manhattan distance, single clustering

```
> speak.dist <- dist(speak.bp, method = "manhattan")
> speak.clust <- hclust(speak.dist, method = "complete")
> plot(speak.clust)
```


Cluster Dendrogram
Height
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
nonverbal
command_language
pronounce
tell
produce_words
characterize
regarding
touch_topic
engage_conversation
public_speech
act_spokesperson
express_thoughts_feelings
speak.dist
hclust (*, "single")

Trying other options: Manhattan distance, average clustering

```
> speak.dist <- dist(speak.bp, method = "manhattan")
> speak.clust <- hclust(speak.dist, method = "average")
> plot(speak.clust)
```


Cluster Dendrogram
Height
2
3
4
5
6
characterize
regarding
touch_topic
produce_words
engage_conversation
public_speech
act_spokesperson
express_thoughts_feelings
nonverbal
command_language
pronounce
tell
speak.dist
hclust (*, "average")

Interpretation

- Which of the solutions speaks to your intuitions the most? The least?

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.