A Reflection on the Clustering in Corpus Linguistics

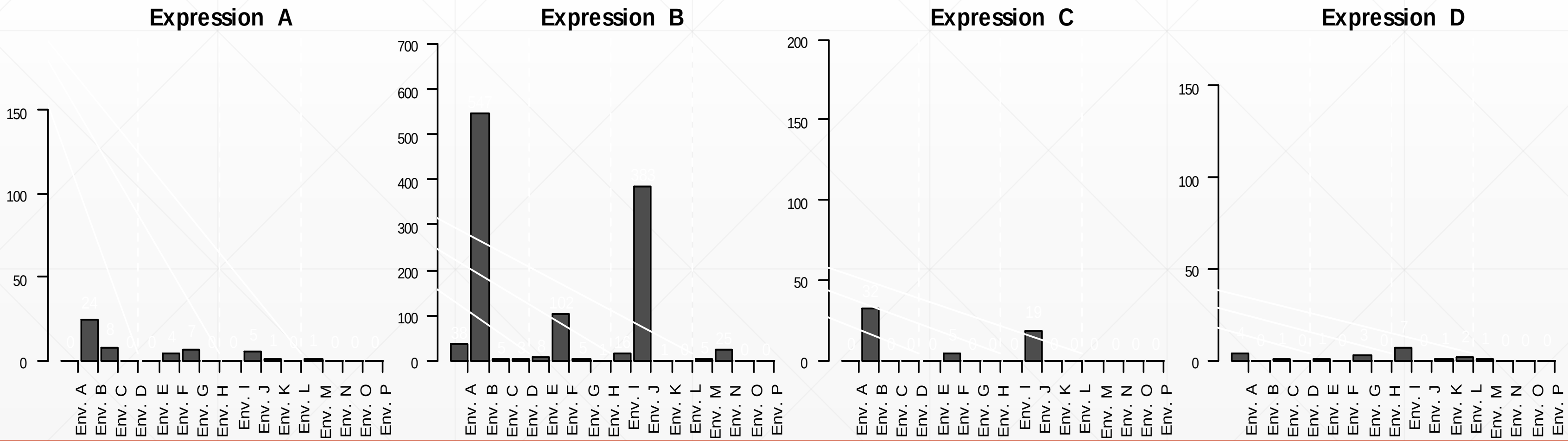
Akitaka Yamada,
Georgetown University
ay314@georgetown.edu

1 Introduction

Topic: to discuss the metric selection in corpus linguistics.

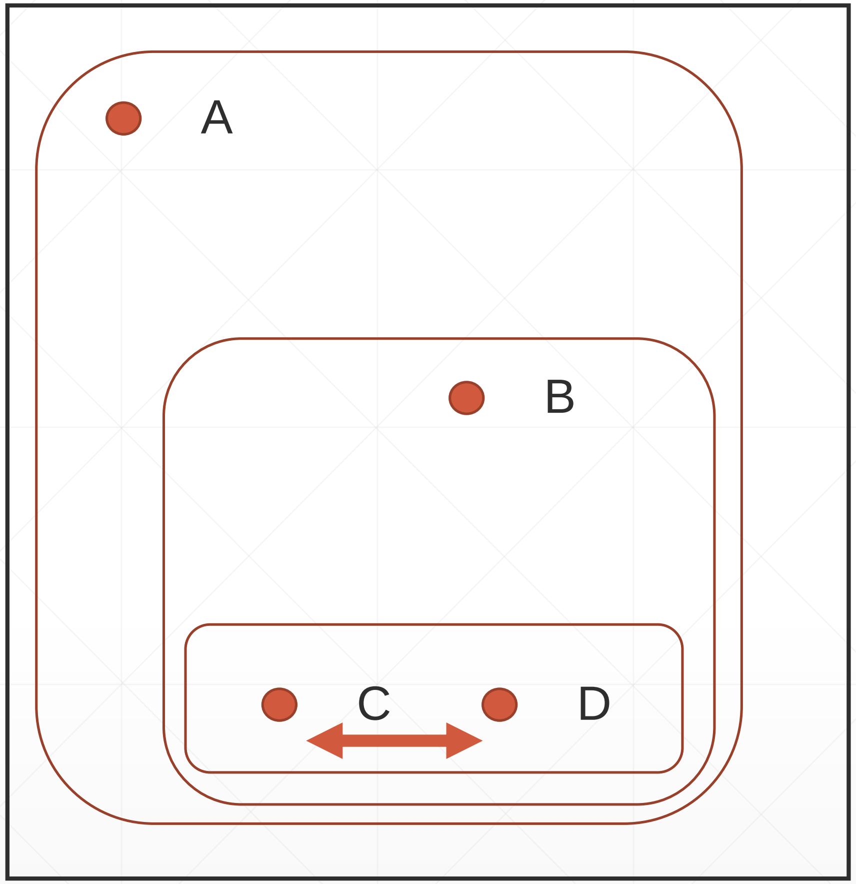
In corpus linguistics, we often classify competing expressions.

Given the following barplots, for example, we sometimes ask which expression is **<u>the closest</u>** to the expression A.



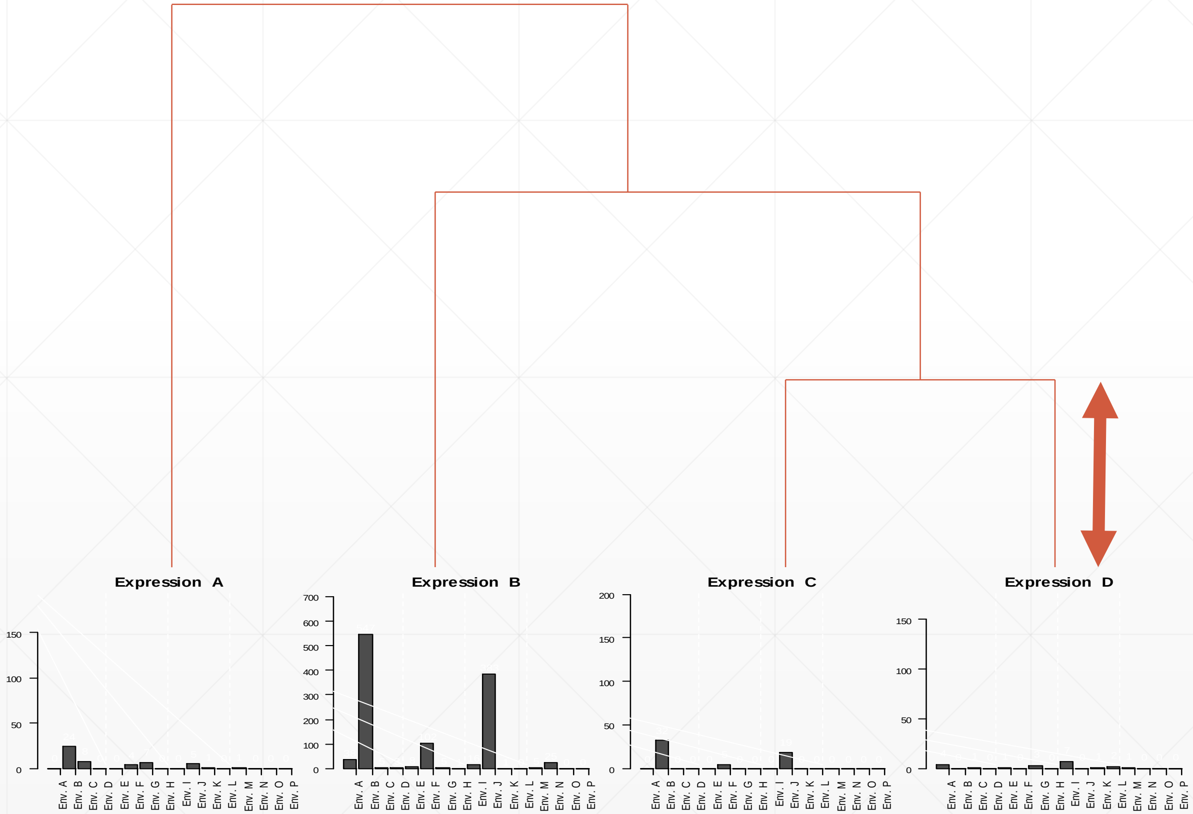
2 Hierarchical Clustering

2 Hierarchical agglomerative clustering analysis



Hierarchical agglomerative clustering is a frequently used explorative statistical method in corpus linguistics (Baayen 2008; Gries 2013; etc.).

Metric selection plays a pivotal role in this analysis.



2 Hierarchical agglomerative clustering analysis

Question:

How do we measure the distance or the similarity among barplots?

An important caveat:

1. No **absolute** answer.
2. A choice of one measure over the others reflects the researcher's **subjective** attitude/perspective toward the data and the analysis.

Nevertheless:

Considering the nature of the corpus data, we can, at least, say the following statements:

Main claims: (i) our familiar **Euclidean distance** is not the only choice; and, in most cases, not the best choice.

(ii) The **Hellinger distance** is an underdiscussed but promising alternative.

(iii) The information lost in clustering can be recovered by a good visualization.

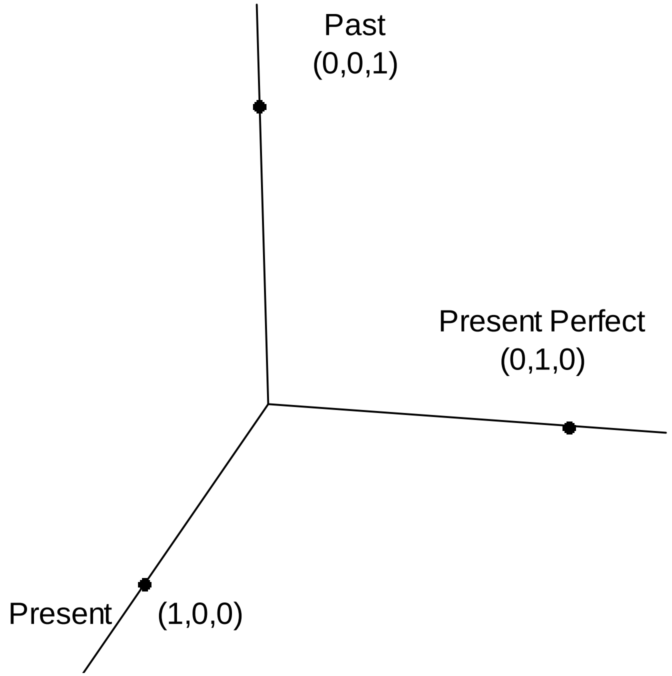
Why?

3 Information Geometry

3 Information geometry

Distribution of verbs

As a warm-up discussion, let us consider the distributional property of the prob. distributions!



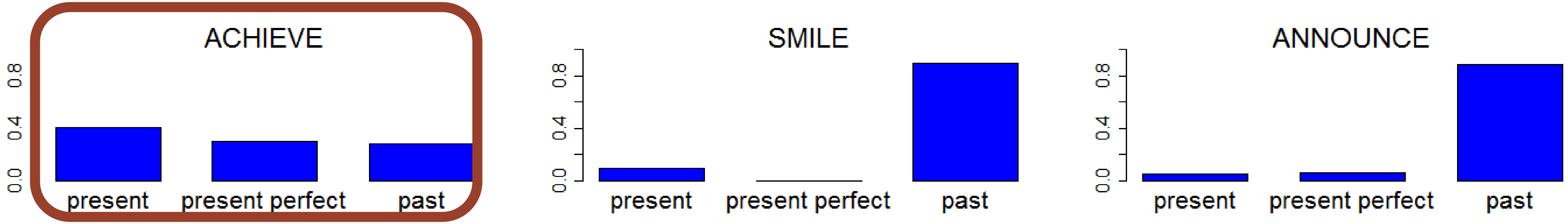
Example:

1. We are interested in the use of **Present Perfect**.

2. How is it different from **the Past** and **the Present**?

Suppose you have searched for these three forms, using COCA.

3. As a result, you have got the following **relative frequencies**:



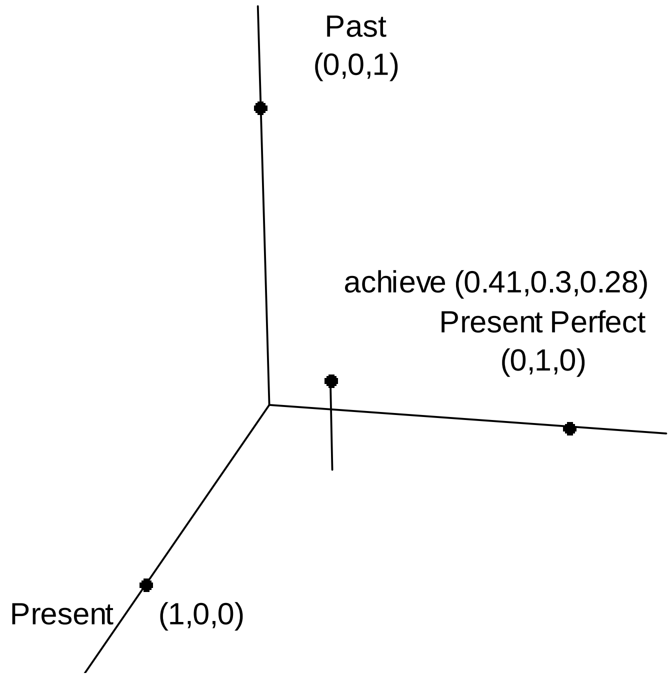
4. In order to understand the nature of the Euclidean distance, let us put these verbs in the **three** dimensional space!

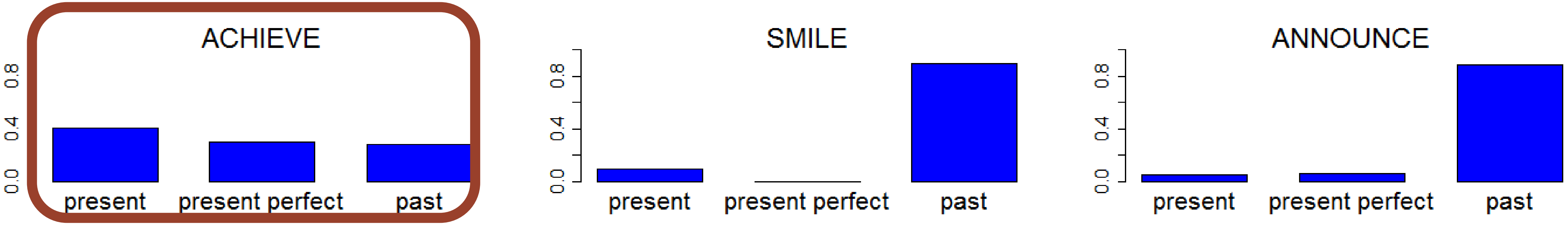
3 Information geometry

Distribution of verbs

Where are those verbs found in COCA corpus?

1. In the case of the verb achieve (0.4, 0.3, 0.3):



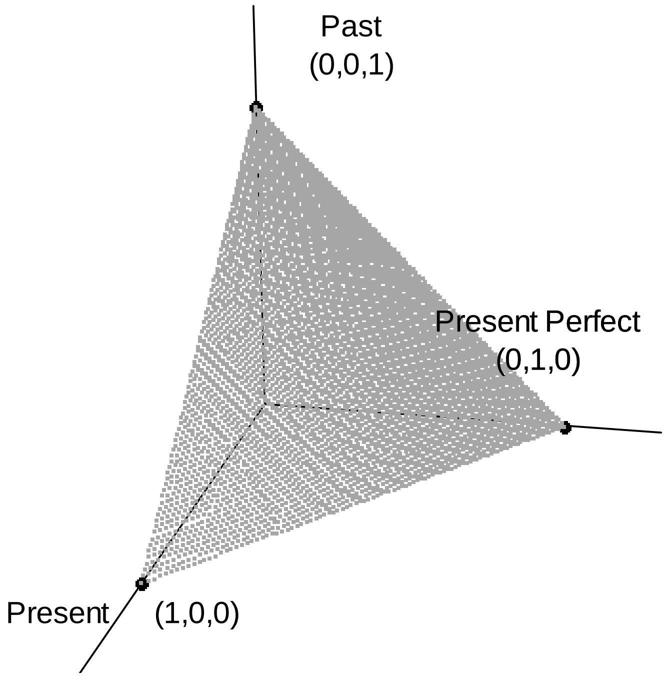


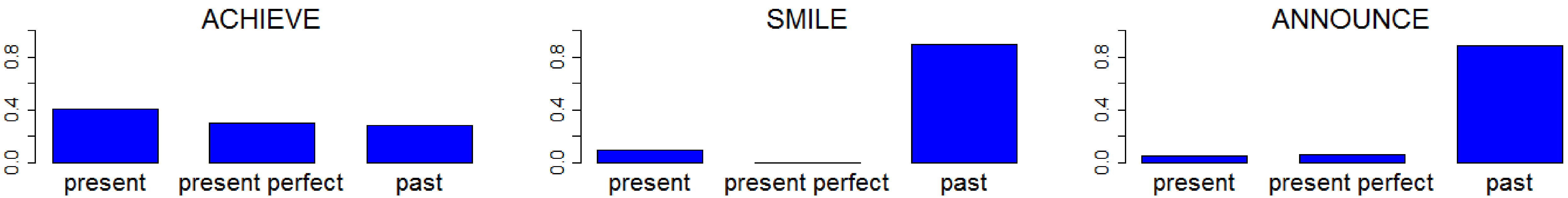
3 Information geometry

Distribution of verbs

Where are those verbs found in COCA corpus?

1. In the case of the verb achieve (0.4, 0.3, 0.3):
2. Can verbs distribute anywhere in this 3D space?
   No, verbs **cannot appear at random!**
   They can only be found within the shaded triangular region
   because of the following constraints: $p_i \geq 0$ $\sum p_i = 1$





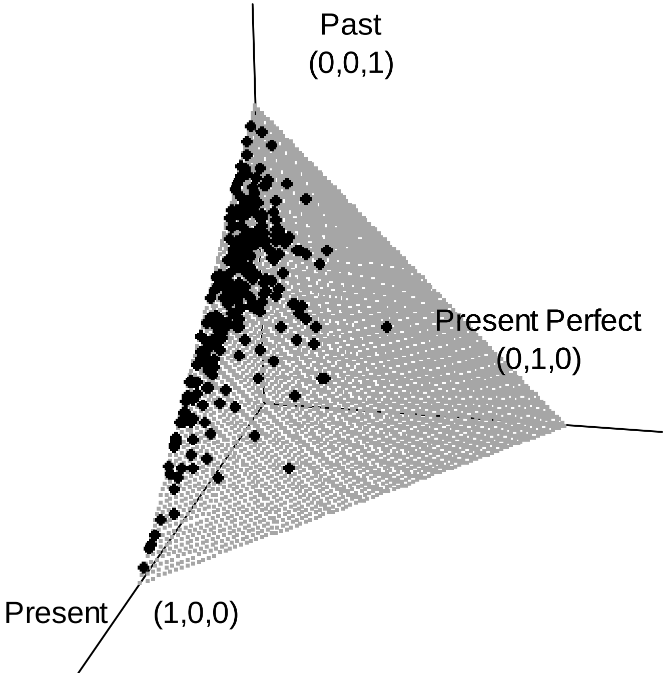
3 Information geometry

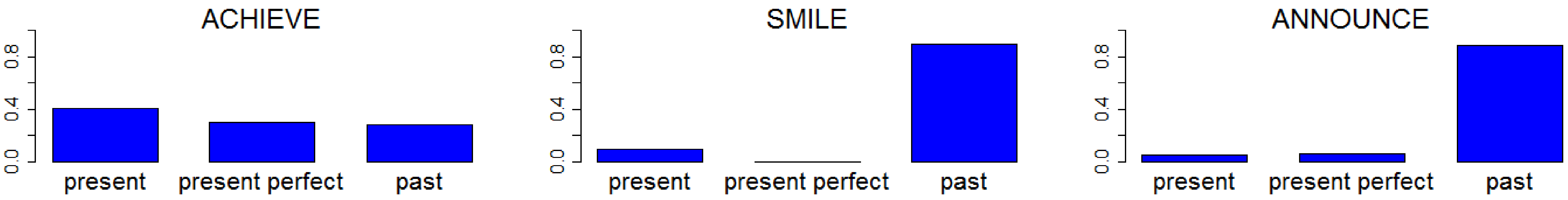
Distribution of verbs

Where are those verbs found in COCA corpus?

1. In the case of the verb achieve (0.4, 0.3, 0.3):
2. Can verbs distribute anywhere in this 3D space?
   No, verbs **cannot appear at random!**
   They can only be found within the shaded triangular region
   because of the following constraints: $p_i \geq 0$ $\sum p_i = 1$

3. **266 most frequently used English verbs** in COCA are plotted in this region:





4 the Euclidean distance and the Hellinger distance

4 Distance between the Euclidean distance and our intuition

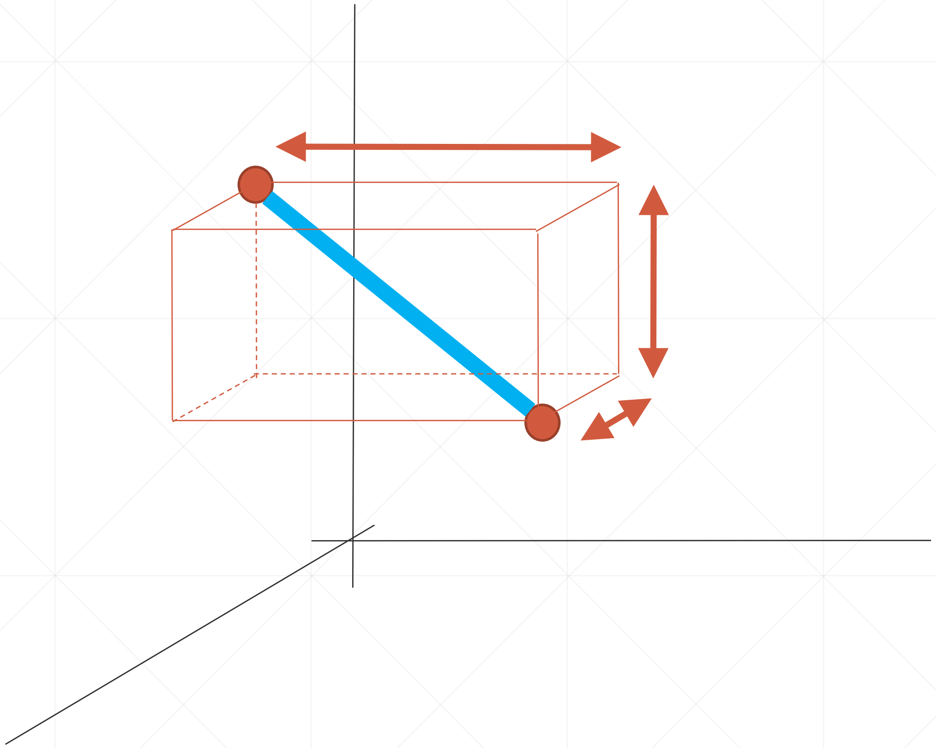
The Euclidean distance

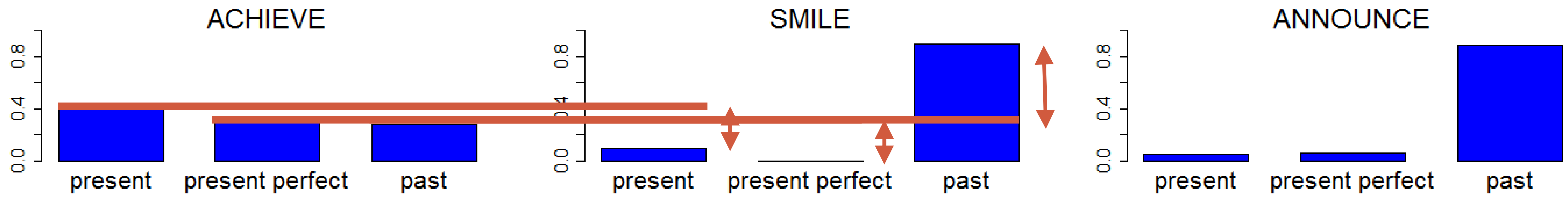
How do we measure the distance between the two dots?

1. Definition:

$$D_E(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_j |x_j - y_j|^2}$$

2. Geometrical interpretation: the straight line





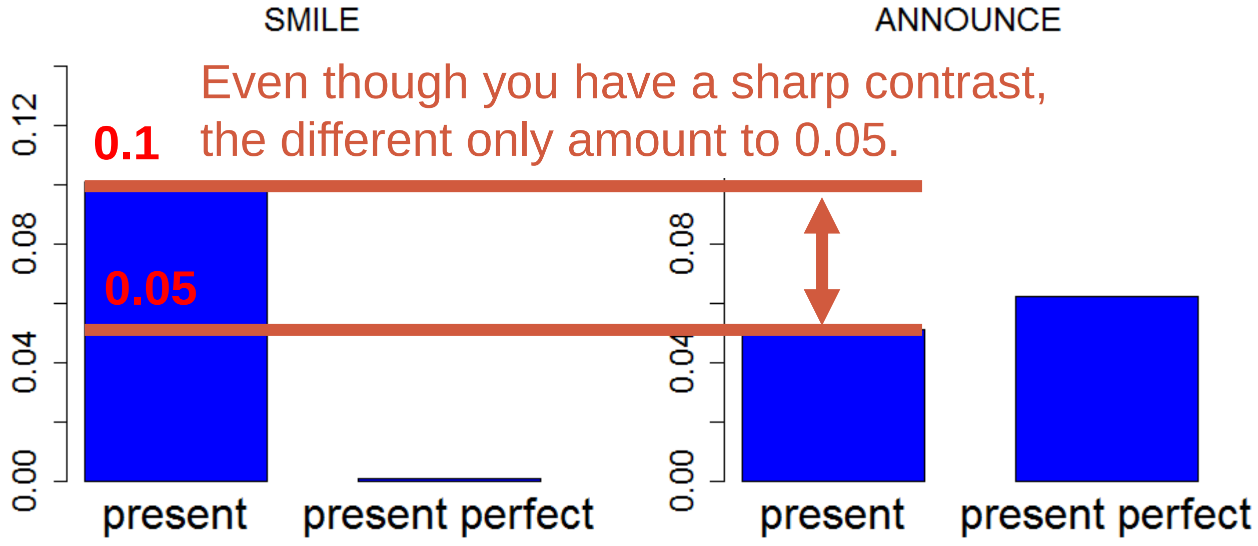
4 Distance between the Euclidean distance and our intuition

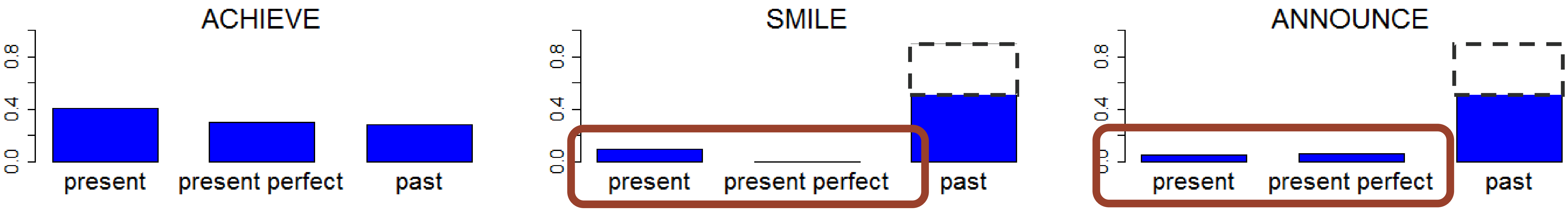
Dependence on the dominant dimension

The Euclidean distance depends too much on the **most dominant dimension**:

Example: the difference between *smile* and *announce*

1. **Preponderance of the past tense** conceals the otherwise detectable contrast.
2. We want to say they are quite **different** in other dimensions.
3. which is totally **ignored** by the Euclidean distance, because of the constraint $\sum p_i = 1$.



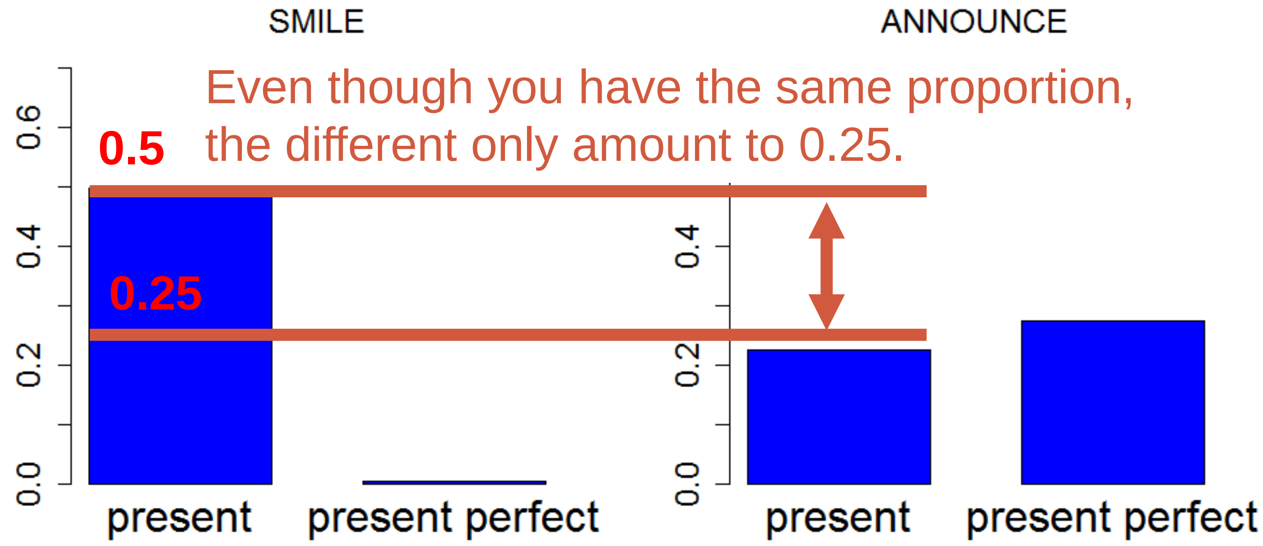


4 Distance between the Euclidean distance and our intuition

Dependence on the dominant dimension

The Euclidean distance depends too much on the **most dominant dimension**:

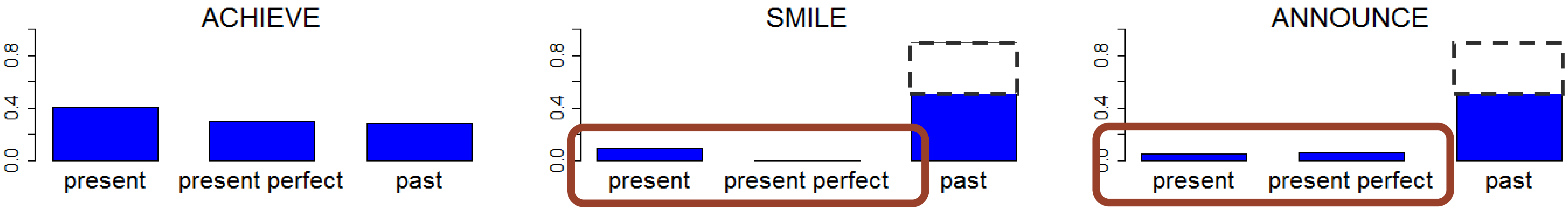
Example: the difference between *smile* and *announce*



1. **Preponderance of the past tense** conceals the otherwise detectable contrast.
2. We want to say they are quite **different** in other dimensions.

The meaning of 0.05 distance is different!

3. which is totally **ignored** by the Euclidean distance, because of the constraint $\sum p_i = 1$.

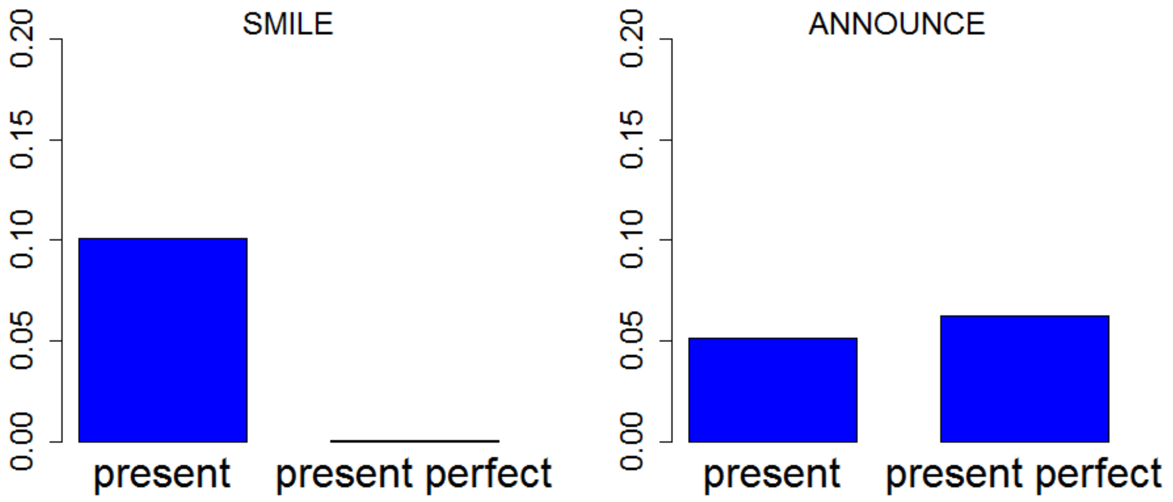


4 Distance between the Euclidean distance and our intuition

The Hellinger distance (philosophy)

The Euclidean distance depends too much on the **most dominant dimension**:

➡ Let's listen to the voice of **minorities**!!



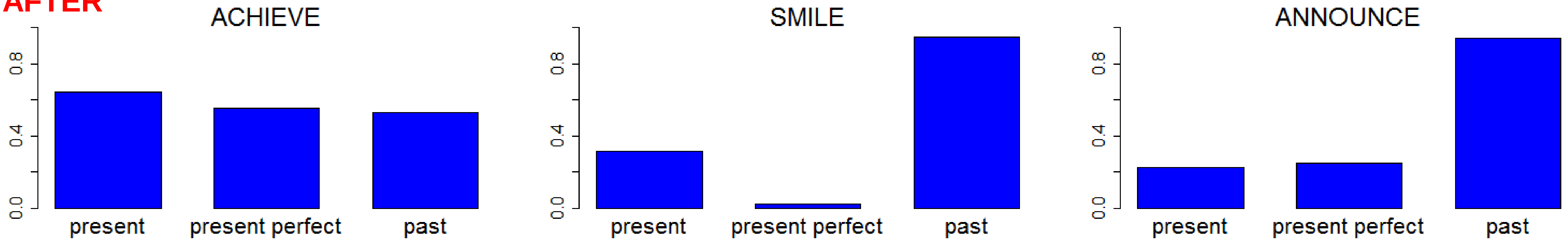
1. What we want: putting more emphasis on the **minorities**
2. **Transform** each bar *s.t.*, the lower bar gets relatively bigger:
3. One of such convex technique is to take the sqrt of each height.

$\sqrt{0.9} = 0.95$ $\quad 0.05\ up!$

$\sqrt{0.1} = 0.32$ $\quad \mathbf{0.22}\ up!$

BEFORE

AFTER



5 Example 1: English Tense and Aspect system

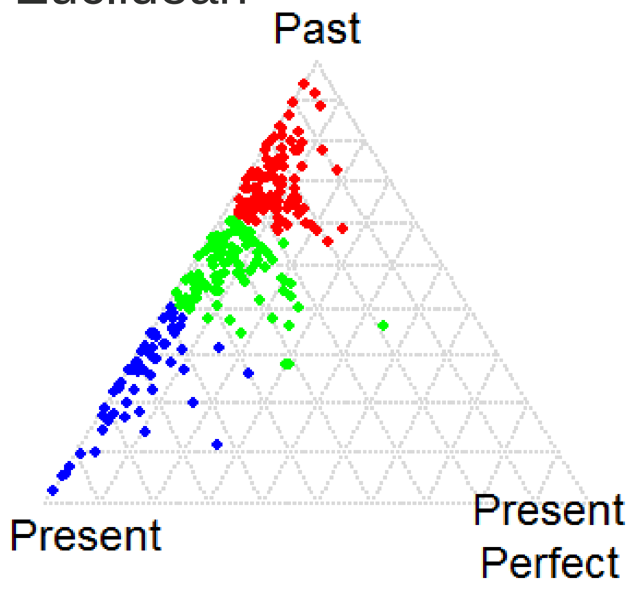
5 Example 1: Tense and Aspect in English

Let us see how the Hellinger distance disagrees with the Euclidean distance.

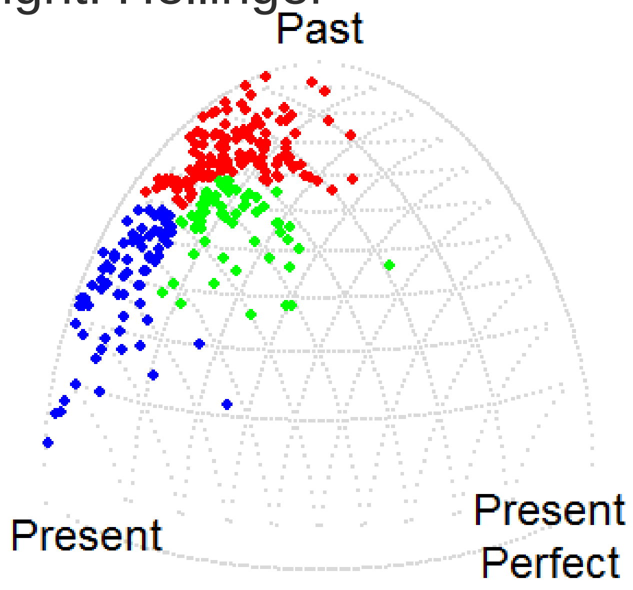
1. **Dendrogram** does not help us a lot.

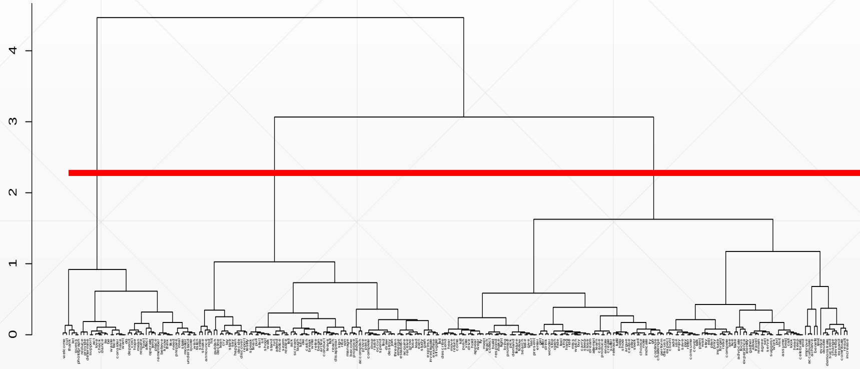
2. **Scatterplot** does.

Left: Euclidean

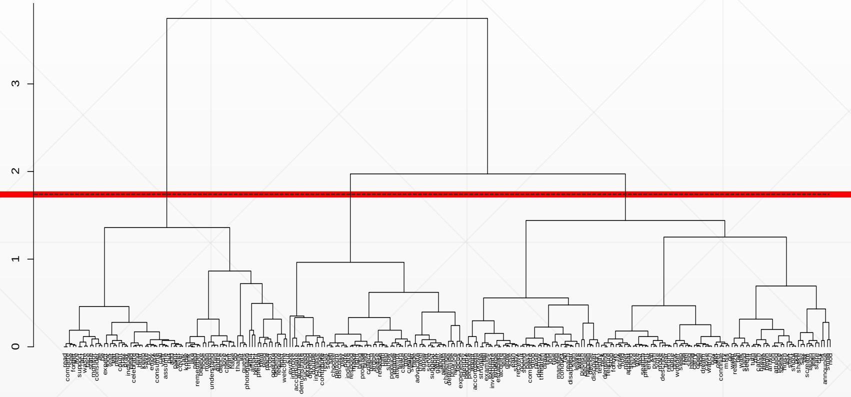


Right: Hellinger





Ward's method/Euclidean distance



Ward's method/Hellinger distance

5 Example 1: Tense and Aspect in English

Let us see how the Hellinger distance disagrees with the Euclidean distance.

1. **Dendrogram** does not help us a lot.

2. **Scatterplot** does.

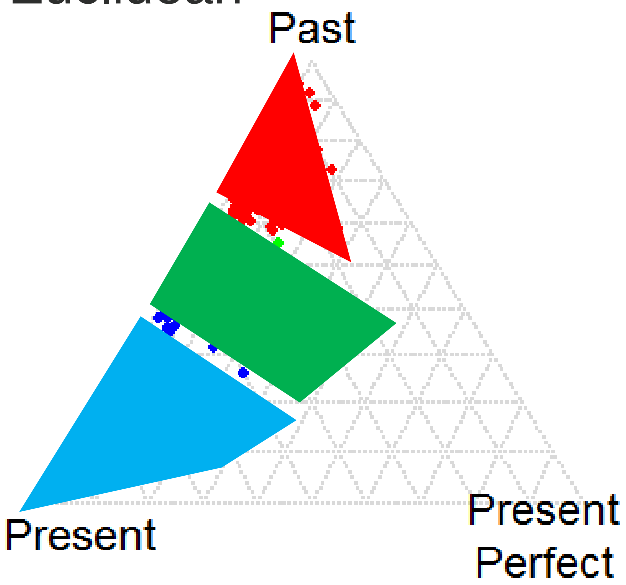
3. This is why the Euclidean distance is not appealing in corpus linguistics.

4. Important caveat:
   The Euclidean distance **does** give us a perspective.

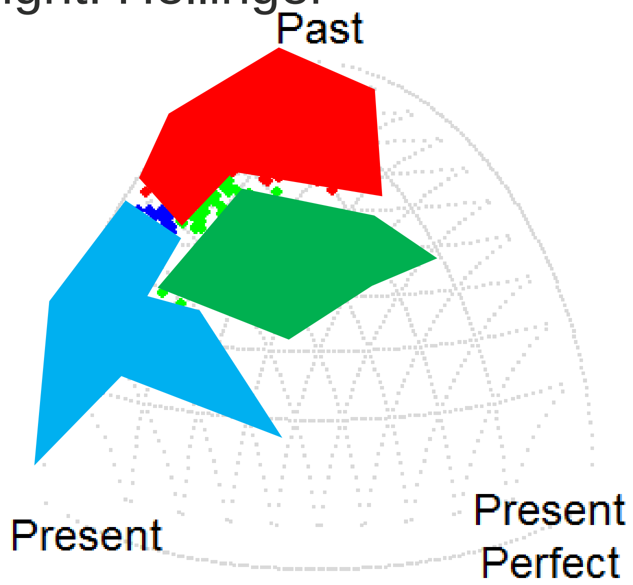
5. Our choice reflects our **subjective** attitude/perspective toward the data.

6. It is good to **compare** results!

Left: Euclidean



Right: Hellinger



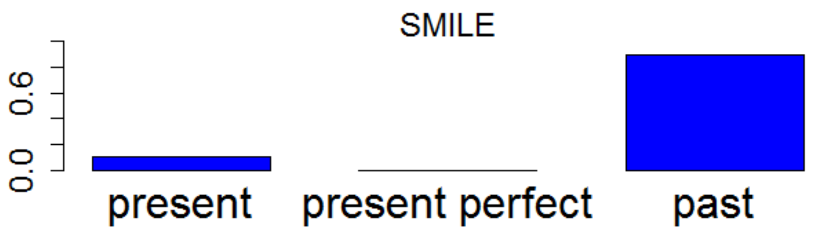
	Euclidean	Hellinger
Commonality	As for the extreme cases, they have similar opinions.	
Emphasis	Dominant dimension	Dominant & minor dimension
Classification	(i) Present (ii) Past (iii) Neither	(i) Present (ii) Past (iii) Present Perfect
Interpretability	Not easy	Quite intuitive

5 Example 1: Tense and Aspect in English

Multifaceted thinking

1. **Robustness**:
   Classification that both approaches agree on.
   Prototypes that **hate** PP.

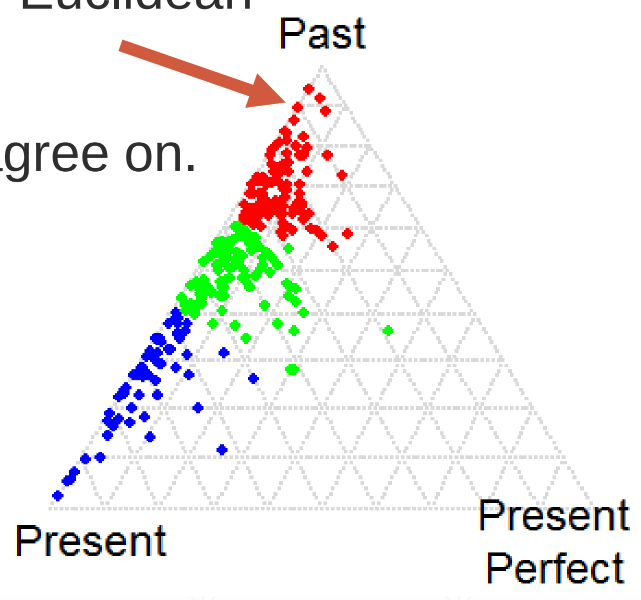
Example:



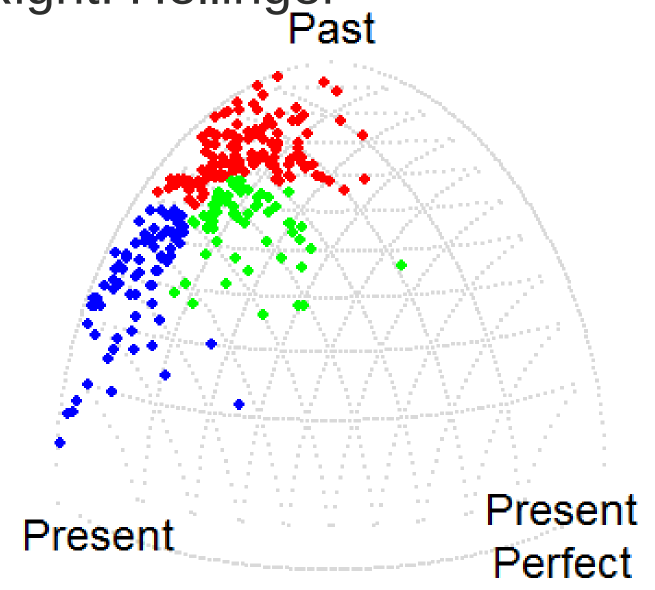
(1) a. when good things happen, we are certain fortune ***has smiled*** on us.

b. Though his expression is serious now, the crinkles at the corners of his eyes make me think he ***has smiled*** a lot. He looks kind.

Left: Euclidean



Right: Hellinger



	Euclidean	Hellinger
Commonality	As for the extreme cases, they have similar opinions.	
Emphasis	Dominant dimension	Dominant & minor dimension
Classification	(i) Present (ii) Past (iii) Neither	(i) Present (ii) Past (iii) Present Perfect
Interpretability	Not easy	Quite intuitive

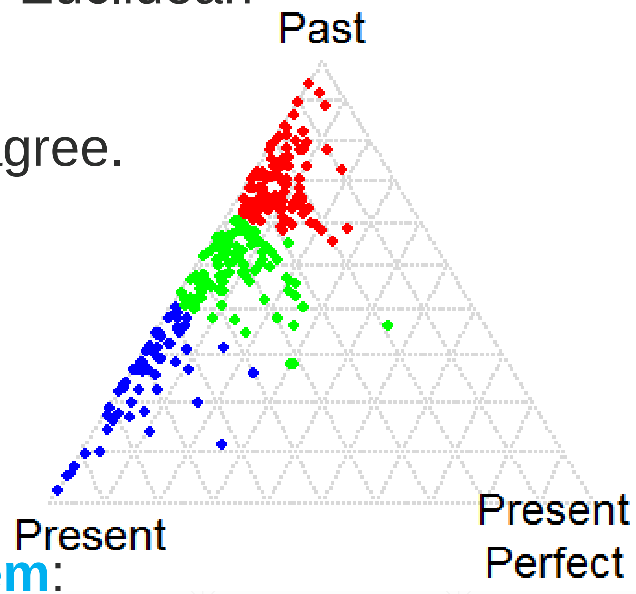
5 Example 1: Tense and Aspect in English

Multifaceted thinking

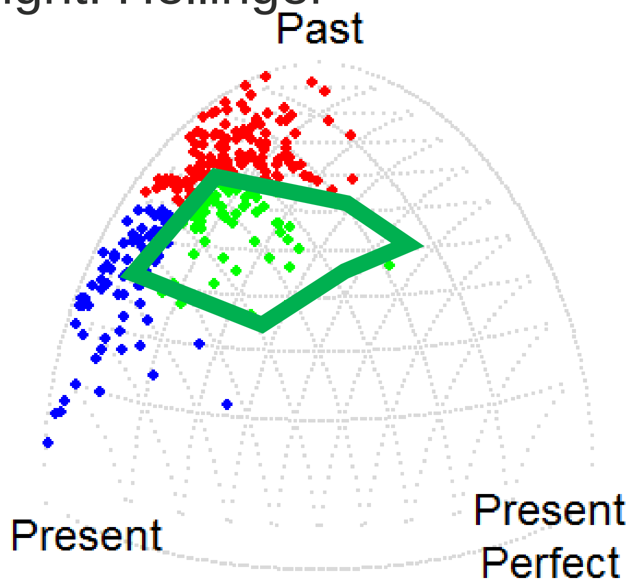
1. **Robustness**:
   Classification that both approaches agree.
   Prototypes that **hate** PP.

announce	lay	scream
cry	lean	shake
hit	nod	smile
laugh	say	stare

Left: Euclidean



Right: Hellinger



2. **Classification w.r.t. three T/A system**:
   Prototypes that **love** PP.

accumulate	demonstrate	expand
achieve	develop	improve
change	double	increase
contribute	evolve	result
		succeed

	Euclidean	Hellinger
Commonality	As for the extreme cases, they have similar opinions.	
Emphasis	Dominant dimension	Dominant & minor dimension
Classification	(i) Present (ii) Past (iii) Neither	(i) Present (ii) Past (iii) Present Perfect
Interpretability	Not easy	Quite intuitive

5 Example 1: Tense and Aspect in English

Multifaceted thinking

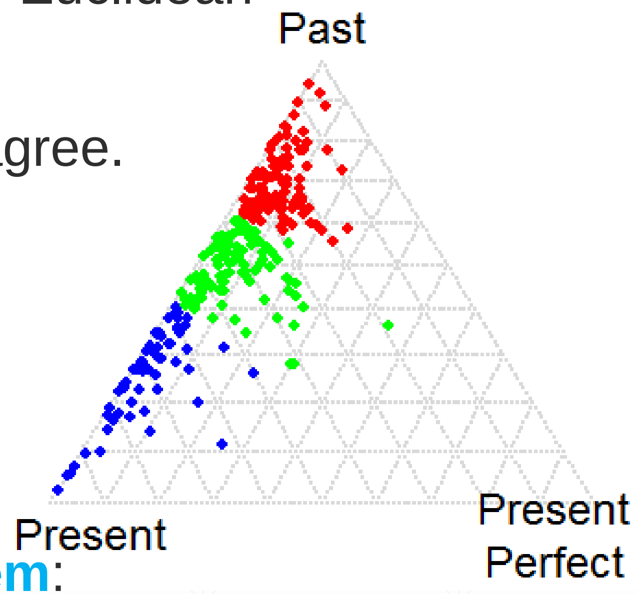
1. **Robustness**:
   Classification that both approaches agree.
   Prototypes that **hate** PP.

announce	lay	scream
cry	lean	shake
hit	nod	smile
laugh	say	stare

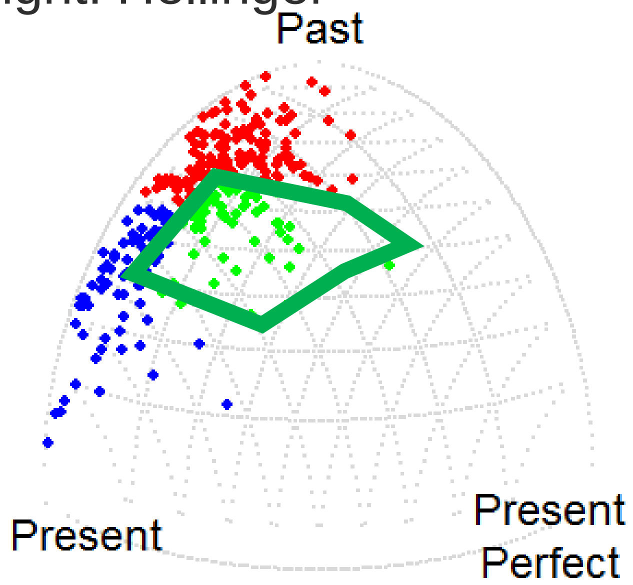
2. **Classification w.r.t. three T/A system**:
   Prototypes that **love** PP.

accumulate	demonstrate	expand
achieve	develop	improve
change	double	increase
contribute	evolve	result
		succeed

Left: Euclidean



Right: Hellinger



Previous theories	(Portner 2011)
(A) Indefinite past theories	
(B) Perfect state theories	
(C) Extended now theories	

Example 2

So far, we have seen an example in which we only have **three dimensions** (= past, present and pp).

→ What about the data with **higher dimensions**?

Example 2 is a case-study in which we have 112 dimensions.

Take-home lessons

1) Good visualization helps us understand the distribution.

2) If compared with the Hellinger distance,
the Euclidean distance gives us a result in which **the highest dimension** is appreciated too much.

3) **Comparison** between the two metrics gives us a better understanding of the data.

Questions are welcome! But let me first conclude this talk …

Conclusion

In this presentation:
I have demonstrated
(a) how we **compare the results** from different metrics
and
(b) how we should **connect** the results **with** the findings in the theoretical linguistics.

In so doing, …
Main claims: (i) our familiar **Euclidean distance** is not the only choice; and, in most cases, not the best choice.
(ii) The **Hellinger distance** is an underdiscussed but promising alternative.
(iii) The information lost in clustering can be recovered by a good visualization.

→ Good comparison of the matrices/visualization

→ Better understanding of the data!

Thank you very much for listening!!