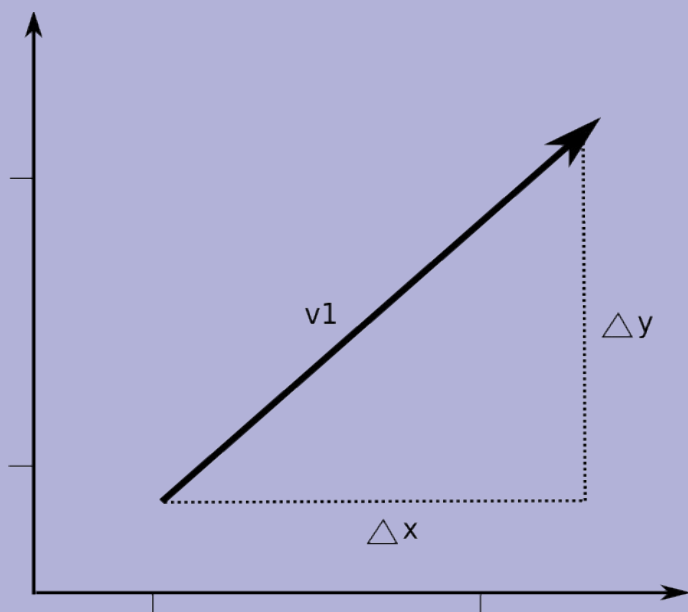


# Vector Analysis in Computational Dialectometry

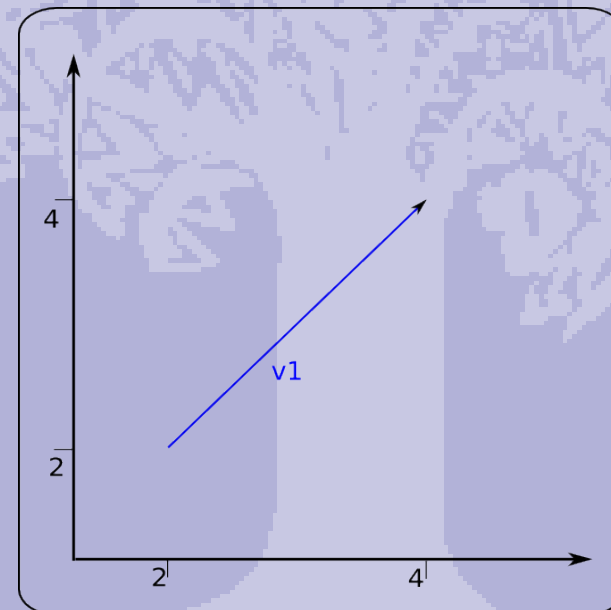


## Outline

- Vector Analysis: basics
- Idea: Vector Analysis in Computational Dialectometry
- Realisation
- Results
- Comparison: VA & Informationtheory (Entropy)
- Future Work in Project Buldialekt

## Basics

- Vector analysis is a subfield of geometry.
- Every array (e.g. vectors) has a starting point and an end point in a two- or more dimensional coordinate system. They determine the *length* and the *direction* of the vector.



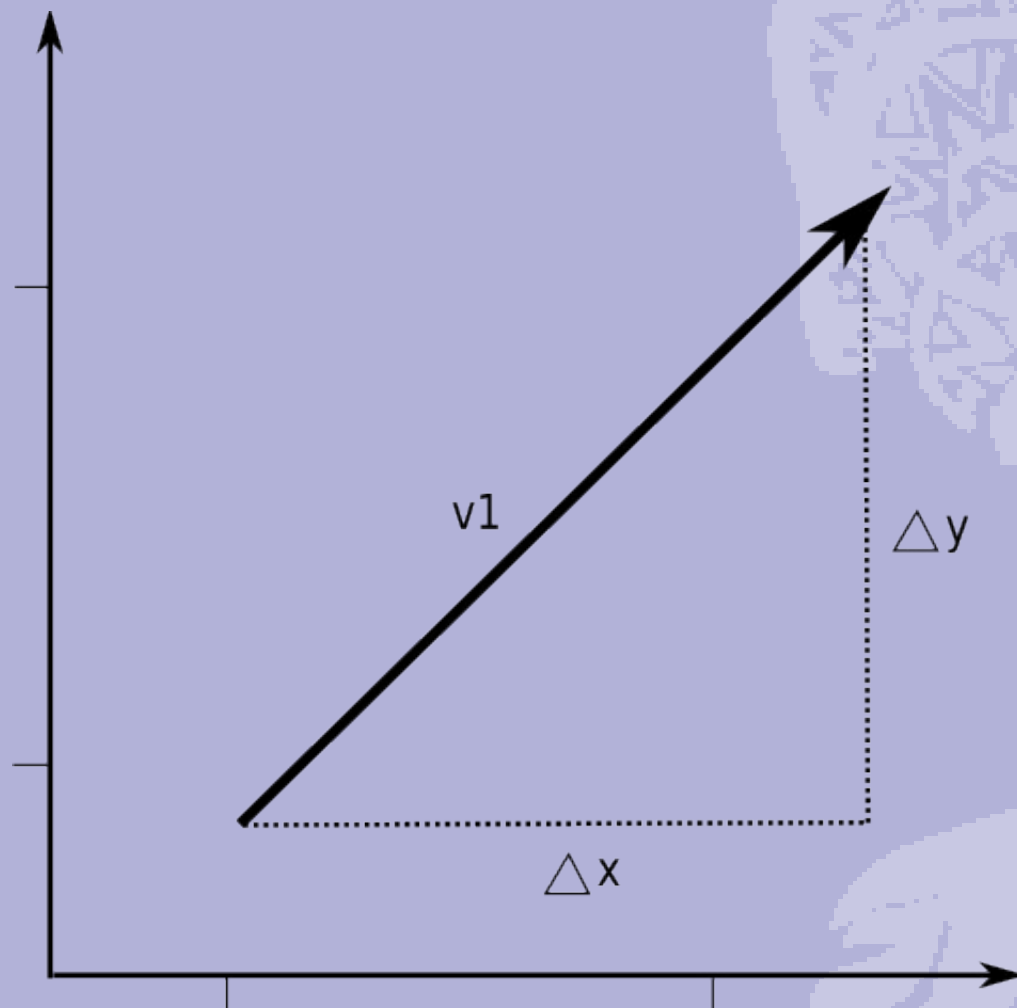
## Basics

- The length of a vector:

$$|\vec{v}| = \sqrt{\Delta x^2 + \Delta y^2}$$

- Angle  $\alpha$  between two vectors:

$$\cos(\alpha) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|}$$

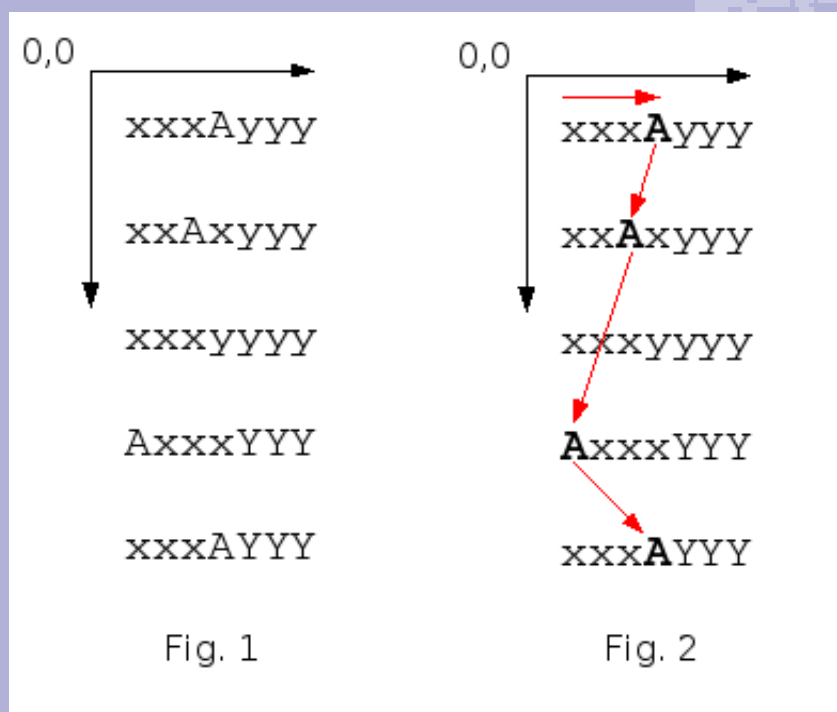


## The Idea

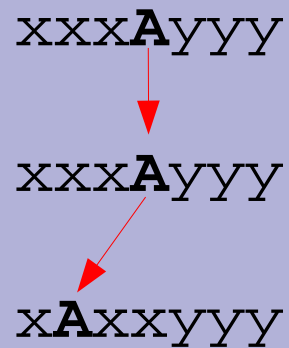
- Keep the focus on one interesting element (uni-, bi-, trigram, ...)
- Follow that element through the whole data-site
- Record the position-changes of that element
- Build a **chain of vectors** through the data
- Compare the position-changes from different data-sites

## In Detail

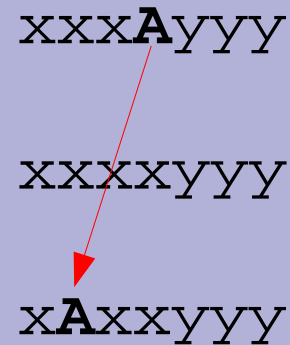
- Following an element (here: A) through the data-site:



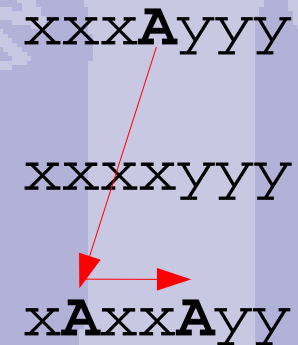
One element per word:



A missing element:



2 elements in one word:





## Example, following the element “e”:

"jA	"jA
"jAgne	"jAgne
be"li	be"li
"berAt	"berAt
"beSe	"beSe
brA"n_je	brA"n_je
"brASno	"brASno
*"br_=Ze	*"br_=Ze
"beme	"beme
veZ"dA	veZ"dA
"vece	"vece
"vet_Ser	"vet_Ser
"vet_SAr	"vet_SAr



## Analysis

**Question:** How to calculate site specific, individual values for every site?

## Answer(s):

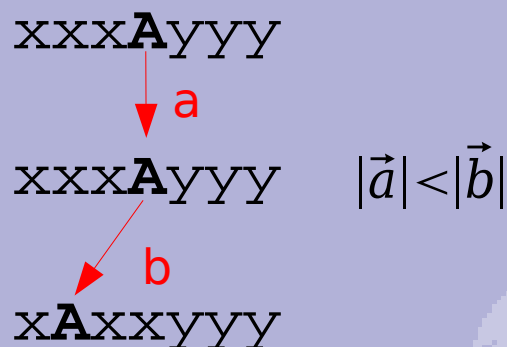
- Using the angle between the single vectors to sum up the movement of the element
- Summing up the length of the single vectors to the length of the whole vector chain

### Question

- Ordering of the word lists?
- At the moment: alphabetic order
- Possible solution: using randomly ordered word lists

## Analysis

- Using the length of a vector chain takes into account the number of the element *and* position changes, while the angle between vectors would just count position changes



## Which element to follow?

- Examinations are possible in two directions:
  - Single-word-all-sites*, a so called SWAS-trace, is an examination of one word in all sites. For example, the word "red" in all sites
  - Single-site-all-words*, a SSAW-trace, examines all different words of a site, for example the complete list of words in site x

SSAW



	Aldomirovci	Asparuhovo	...	Zheravna
агне (lamb)	"jAgne	"Agni	...	"Agni
аз (I)	"jA	"As	...	"As
бели (white-plural)	"beli	"beli	...	"beli
берат (pick up - 3rd plural)	"beru	bi"r7t	...	bi"r7t
...	...	...	...	...
ям (eat, 1st singular)	e"dem	"jAm	...	"jAm

SWAS



## Which element to follow?

- Using the SWAS direction for identifying the elements with the most position changes in the data set:

X-Sampa code	Length of Vector Chain
e	40015.1759910523
stress	35731.207131129
7 (close-mid back, unrounded)	35653.6778159966
A	35432.7572223606
i	34438.756791175
u	34120.3965759371
n	33581.1330654058
s	33038.0473845845
o	32878.0780176776
_j (palatalized)	32317.4612226377

## First results

- Uni-grams
- Analysis of vowels is more informative than consonants
- Clear distinction between the east and the west of Bulgaria

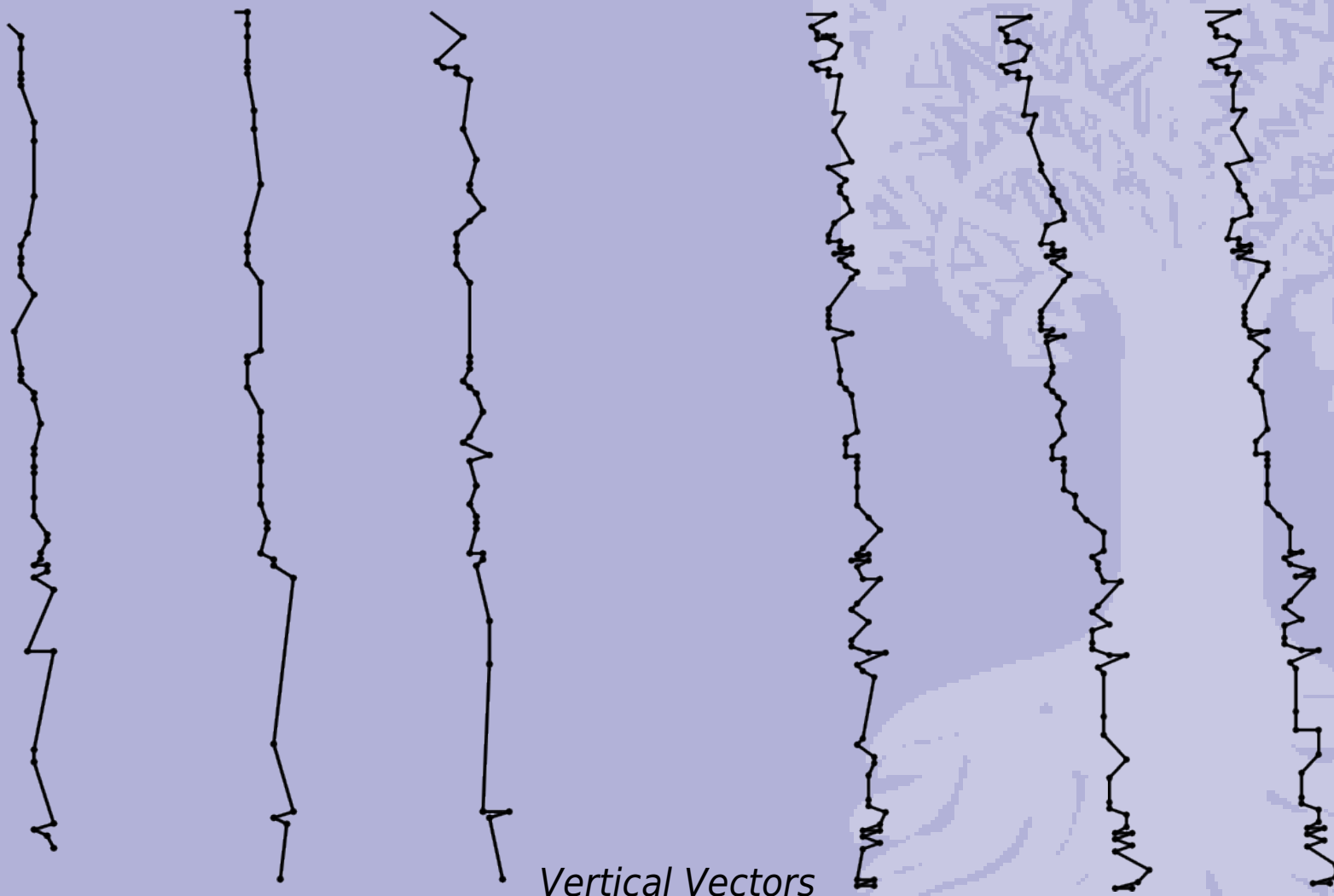
# First results: East- west

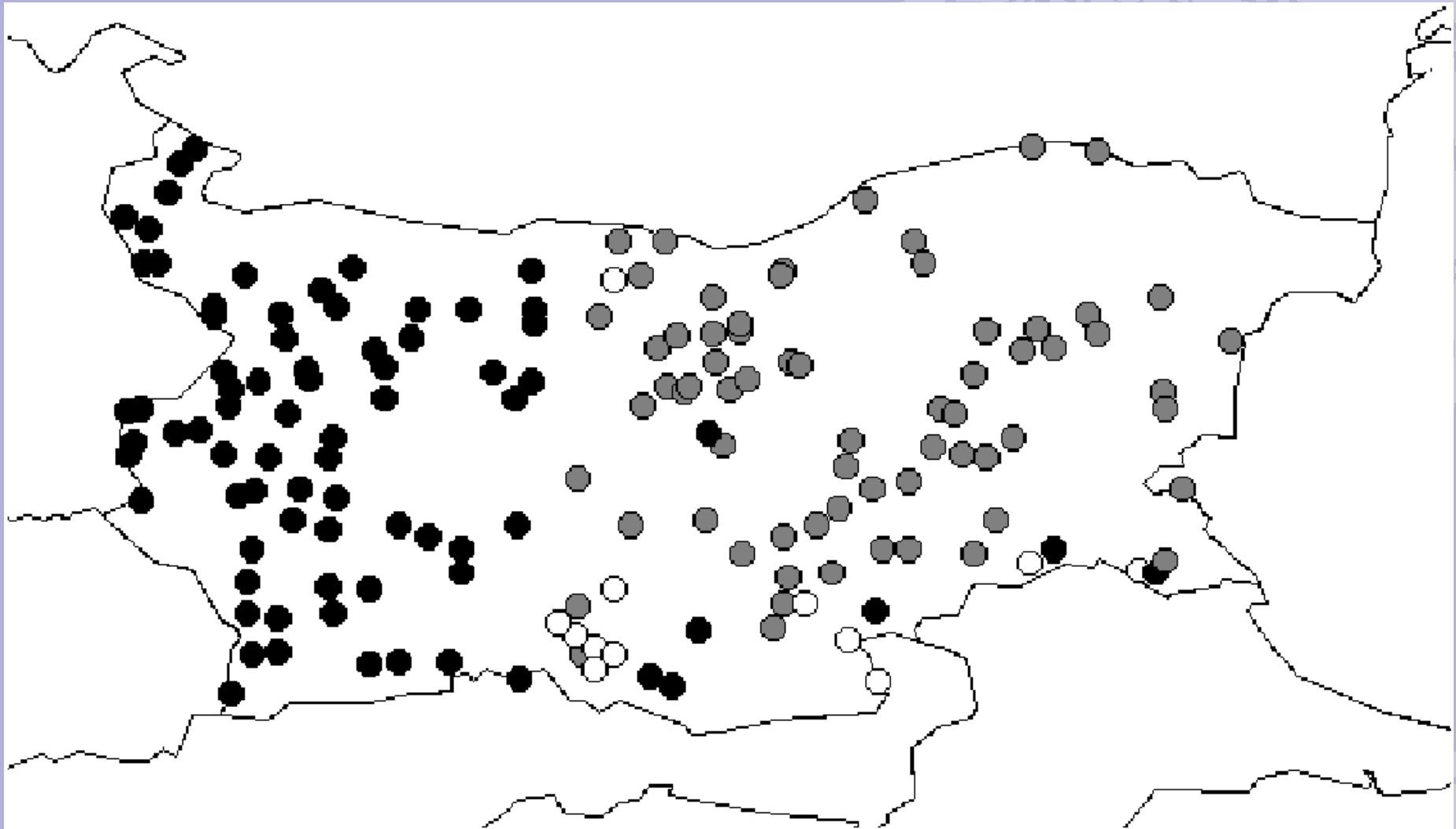




East

West





## Comparison: Information Theory and Vector Analysis

### Information Theory

Bigrams  
Corpus based  
Takes all elements into account  
Ignores position of elements  
Number of elements are measured

### Vector Analysis

Focused Element  
Word based  
Uses just single elements  
Tracks position of elements  
Number of elements influences the VC

## Future work in *Vector Analysis*

- Using other structures than Uni-grams
- Use of randomly ordered word lists
- Combining the vector-based-approach with other approaches

## Future work in *general*

- Different clustering methods
- Classifiers instead of clustering
- More complex analysis in GIS