Dialectology & Dialectometry

Outline

- What is a dialect?
- Dialectology
- Computational Dialectometry
- Project Buldialects: Methods & Results

What is a Dialect?

- "A language is a dialect with an army and navy"

  (Max Weinreich)

- A dialect (from the Greek word **διάλεκτος, dialektos**) is a *variety* of a language characteristic of a particular group of the language's speakers.

  (Wikipedia)

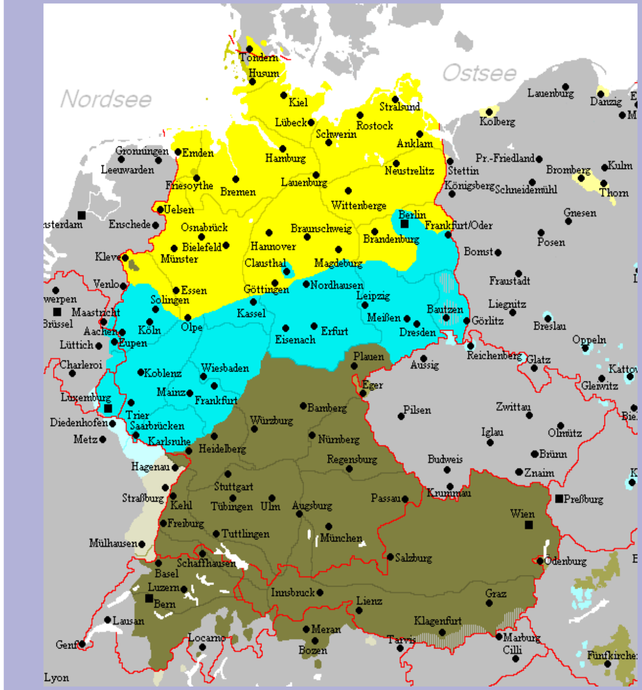
What is a Dialect?

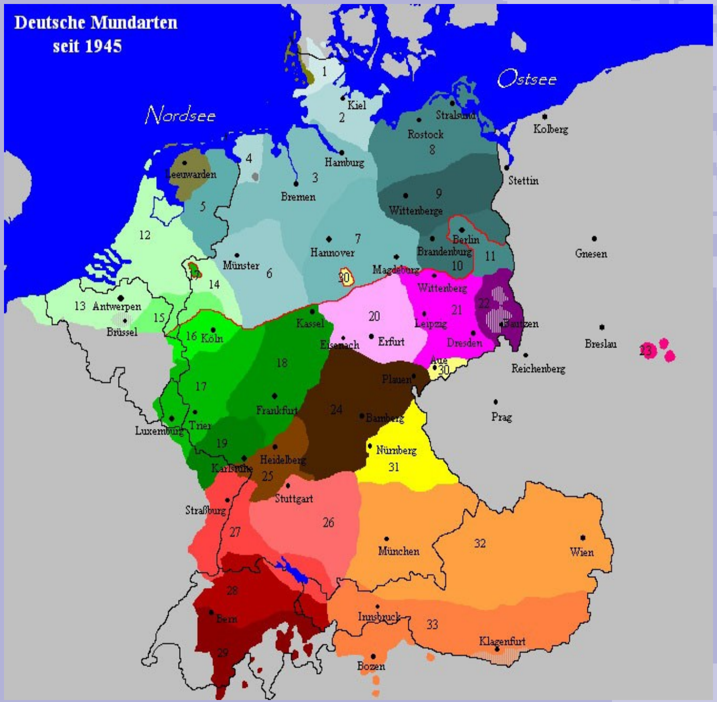
- The term dialects allows definition of individuals dialects in several meanings:
  - Social division of dialects
  - Age specific dialects
  - Cultural, political dialect regions
- But most linguistic research in the field of dialects deals with the *geographical distribution* of dialects

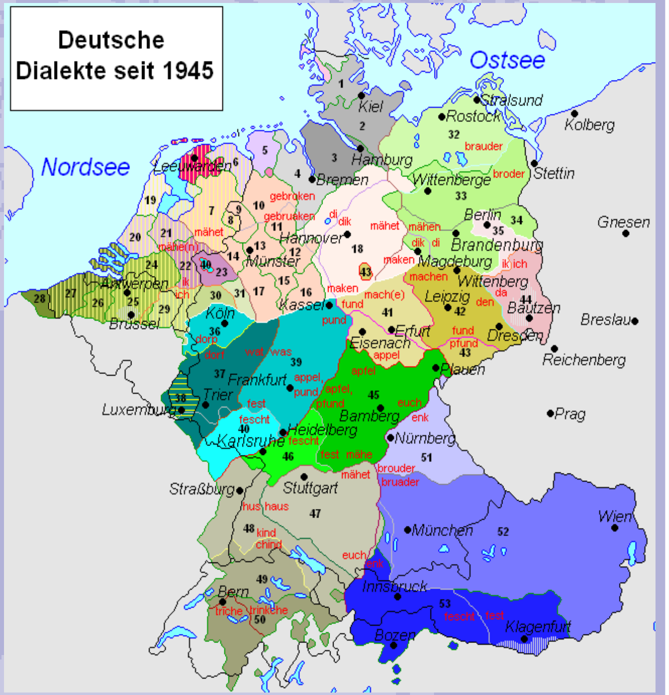
What is a Dialect - remarks

- In contrast to a dialect, the *idiolect* describes the linguistic properties of an individual speaker and not a group of speakers
- Also the *standard language* is a dialect. A standard language is mostly defined by a political or cultural authority

German Dialects







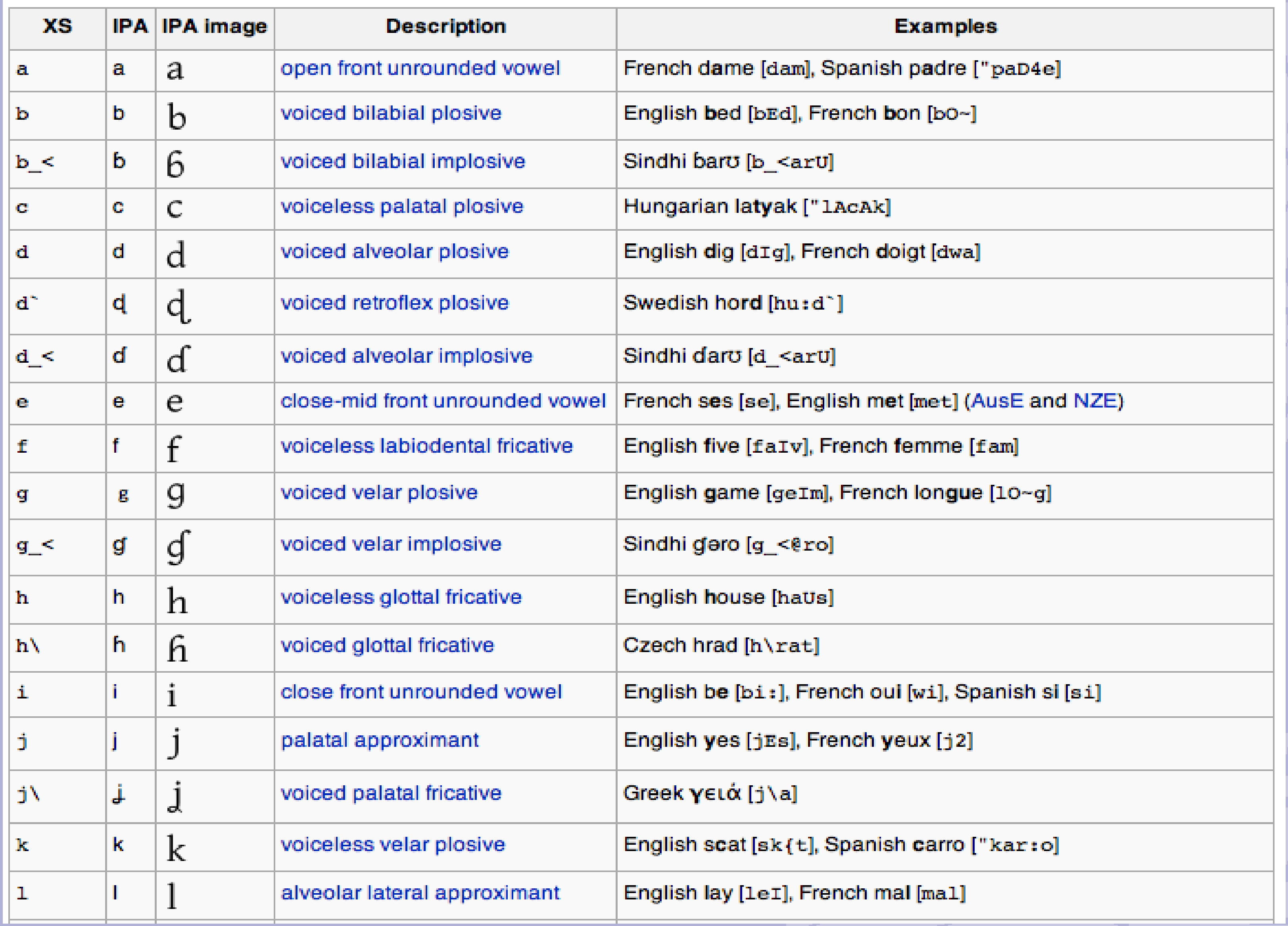
Dialect Data

- Two kinds of dialect data:
- Lexical data:
  - Kartoffel - Grumbeere
- Pronunciation data:
  - Grumbeere - Krombeere

In the rest of the presentation, I'm talking only about pronunciation data!

Annotating Pronunciation Data

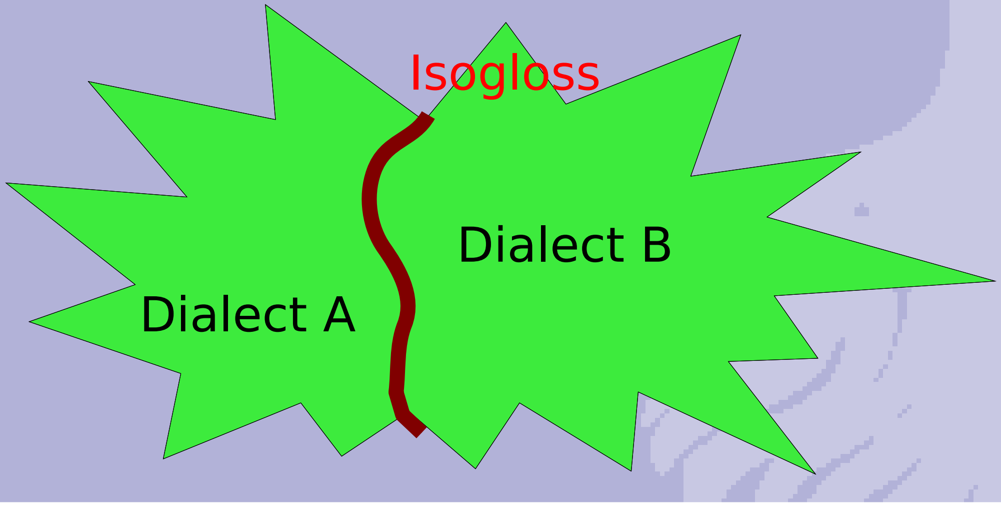
- “The International Phonetic Alphabet” as standard
- The IPA contains 107 vowels & consonants, 31 diacritics and 19 suprasegmentals
- XSampa is an electronic encoding of the IPA



XS	IPA	IPA image	Description	Examples
a	a	a	open front unrounded vowel	French d**a**me [dam], Spanish p**a**dre ["paD4e]
b	b	b	voiced bilabial plosive	English **b**ed [bEd], French **b**on [bO~]
b_<	ɓ	ɓ	voiced bilabial implosive	Sindhi **ɓ**arʊ [b_<arU]
c	c	c	voiceless palatal plosive	Hungarian la**ty**ak ["lAcAk]
d	d	d	voiced alveolar plosive	English **d**ig [dIg], French **d**oigt [dwa]
d`	ɖ	ɖ	voiced retroflex plosive	Swedish ho**rd** [hu:d`]
d_<	ɗ	ɗ	voiced alveolar implosive	Sindhi **ɗ**arʊ [d_<arU]
e	e	e	close-mid front unrounded vowel	French s**e**s [se], English m**e**t [met] (AusE and NZE)
f	f	f	voiceless labiodental fricative	English **f**ive [faIv], French **f**emme [fam]
g	ɡ	ɡ	voiced velar plosive	English **g**ame [geIm], French lon**gu**e [lO~g]
g_<	ʛ	ʛ	voiced velar implosive	Sindhi **ʛ**əro [g_<@ro]
h	h	h	voiceless glottal fricative	English **h**ouse [haUs]
h\	ɦ	ɦ	voiced glottal fricative	Czech hrad [h\rat]
i	i	i	close front unrounded vowel	English b**e** [bi:], French ou**i** [wi], Spanish s**i** [si]
j	j	j	palatal approximant	English **y**es [jEs], French **y**eux [j2]
j\	ʝ	ʝ	voiced palatal fricative	Greek **γ**ειά [j\a]
k	k	k	voiceless velar plosive	English s**c**at [sk{t], Spanish **c**arro ["kar:o]
l	l	l	alveolar lateral approximant	English **l**ay [leI], French ma**l** [mal]

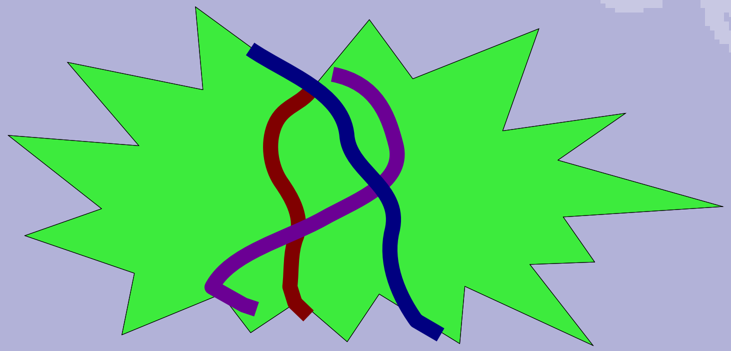
Dialectology

- Scientific dialectology is done since the 17th century
- Dialectology considers extra linguistic knowledge like social sciences, migrations, cultural borders etc.
- Dialect regions are separated by *isglosses* or bunches of isoglosses:



Dialectology - Problems with isoglosses

- In the practice, isoglosses have often a complete different distribution:



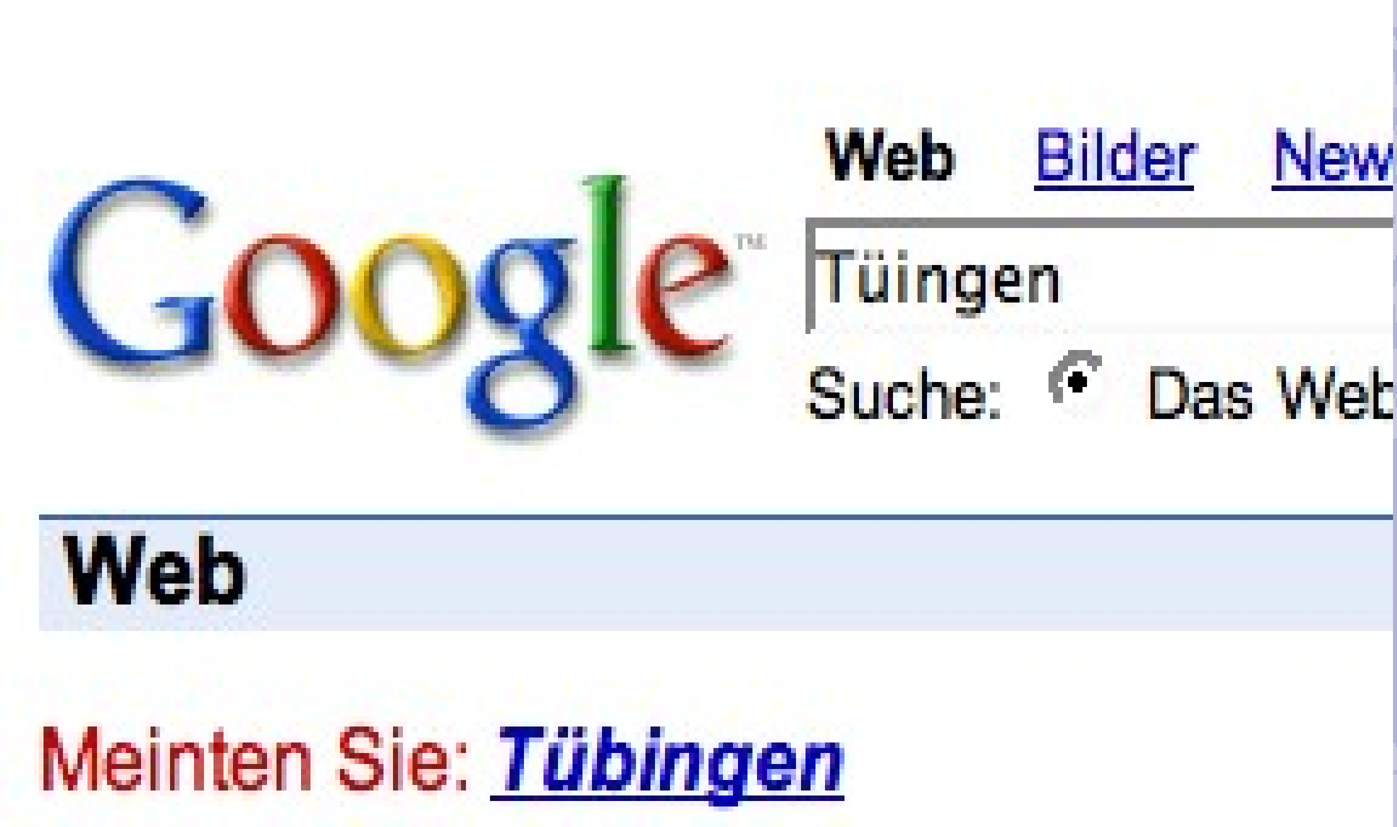
- Some isoglosses are more important than other isoglosses
- Dialect continua can't be represented by isoglosses

Computational Dialectometry

- In contrast to Dialectology, Computational Dialectometry tries to identify dialects by the use of mathematical, statistical and information theoretic approaches
- No extra linguistic information is needed - in a second step, information like distances, population etc. can be considered
- Both dialect continua and isoglosses can be detected

Computational Dialectometry

- Methods of choice in Computational Dialectometry are *Aggregate Methods*
- Aggregate Methods are calculating the differences between two or more words and then aggregating this differences for the whole data record
- Two different kinds of AM: *alignment* and *edit distances*
- Edit distance algorithms like Levenshtein are used in a wide range of fields, for example for correcting typos:



Tüingen

Suche: ● Das Web

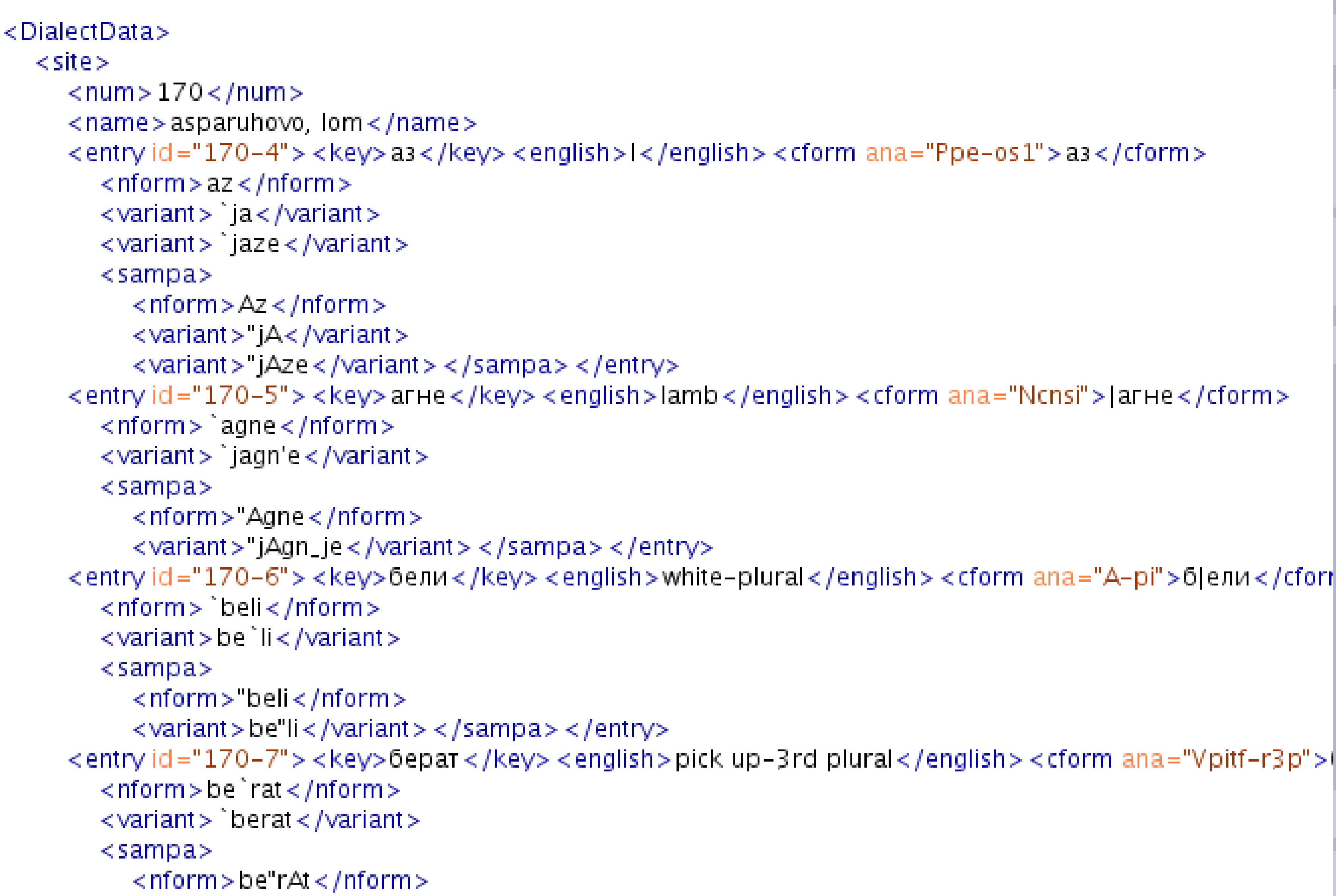
Meinten Sie: ***Tübingen***

Computational Dialectometry - Project Buldialects

- Cooperation between the University Groningen, University Tübingen, Bulgarian Academy of Science and the University Sofia
- Basis: a data set of Bulgarian dialects:
    - ca. 200 sites, spread across the whole region of Bulgaria
    - ca. 150 words per site

Project Buldialects - the Data Set

"jAgne	brA"n_je	u"lAzim
"Agni	"brAni	u"lAz@m
"jAgn_je	brA"n_je	u"lAzim
"jAgne	br@"ne	"flAz@m
"jegn_j@	br@"ne	"fl_jAv@m
"jAgne	be"ren_je	u"lAzim
"jAgne	br@"n_je:	*"fliz@m
"jAgne	brA"n_je	u"lAzim
"jAgne	br@"ne	"flAzem
"Agne	brA"ne	"flA:m



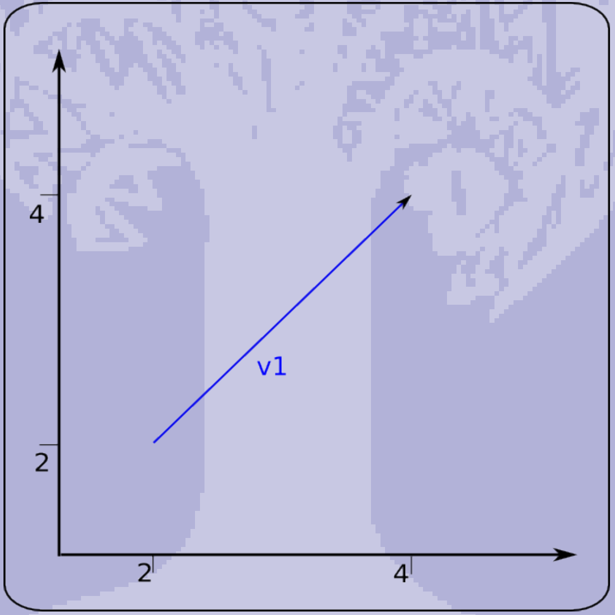
```
<DialectData>
   <site>
      <num>170</num>
      <name>asparuhovo, lom</name>
      <entry id="170-4"><key>аз</key><english>I</english><cform ana="Ppe-os1">аз</cform>
         <nform>az</nform>
         <variant>`ja</variant>
         <variant>`jaze</variant>
         <sampa>
            <nform>Az</nform>
            <variant>"jA</variant>
            <variant>"jAze</variant></sampa></entry>
      <entry id="170-5"><key>агне</key><english>lamb</english><cform ana="Ncnsi">|агне</cform>
         <nform>`agne</nform>
         <variant>`jagn'e</variant>
         <sampa>
            <nform>"Agne</nform>
            <variant>"jAgn_je</variant></sampa></entry>
      <entry id="170-6"><key>бели</key><english>white-plural</english><cform ana="A-pi">б|ели</cfor
         <nform>`beli</nform>
         <variant>be`li</variant>
         <sampa>
            <nform>"beli</nform>
            <variant>be"li</variant></sampa></entry>
      <entry id="170-7"><key>берат</key><english>pick up-3rd plural</english><cform ana="Vpitf-r3p">
         <nform>be`rat</nform>
         <variant>`berat</variant>
         <sampa>
            <nform>be"rAt</nform>
```

Computational Dialectometry - Project Buldialects

- In Tübingen: two new approaches to Computational Dialectometry were developed in the last two years:
    - Information theory related: Entropy
    - Vector Analysis

Vector Analysis

- 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.

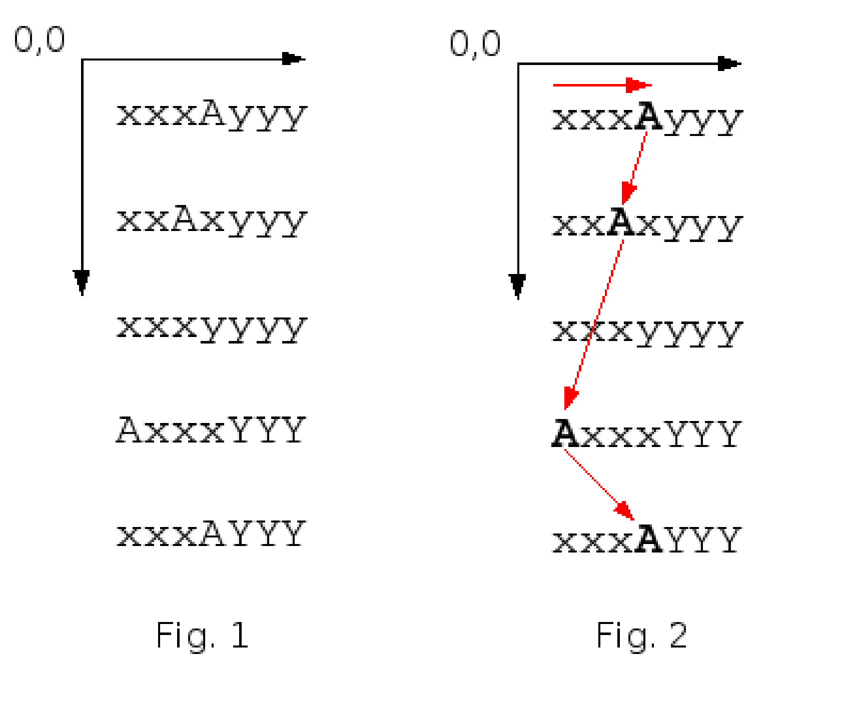


The Idea: how to use vector analysis in CD

- Keep the focus on one interesting element
- Follow that element through the whole data record
- Record the position changes of that element
- Build a ***chain of vectors*** through the data
- Compare the vector chains from different data records

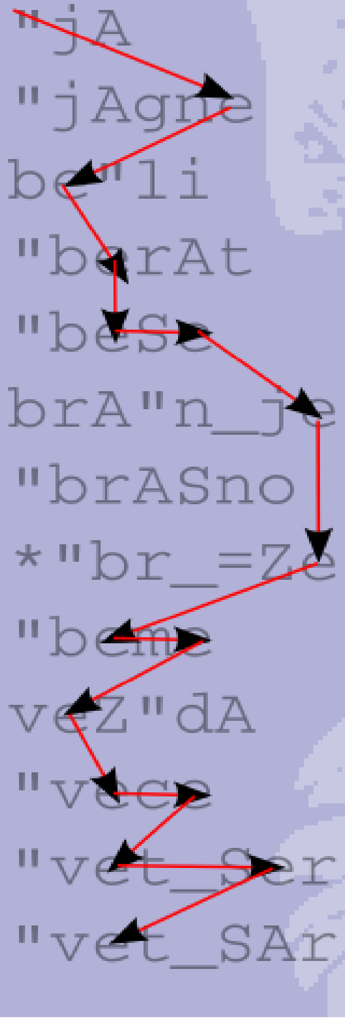
Vector Analysis

- Artificial example of a vector chain:

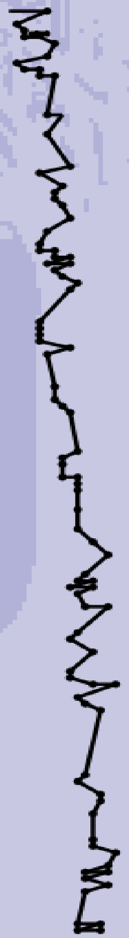


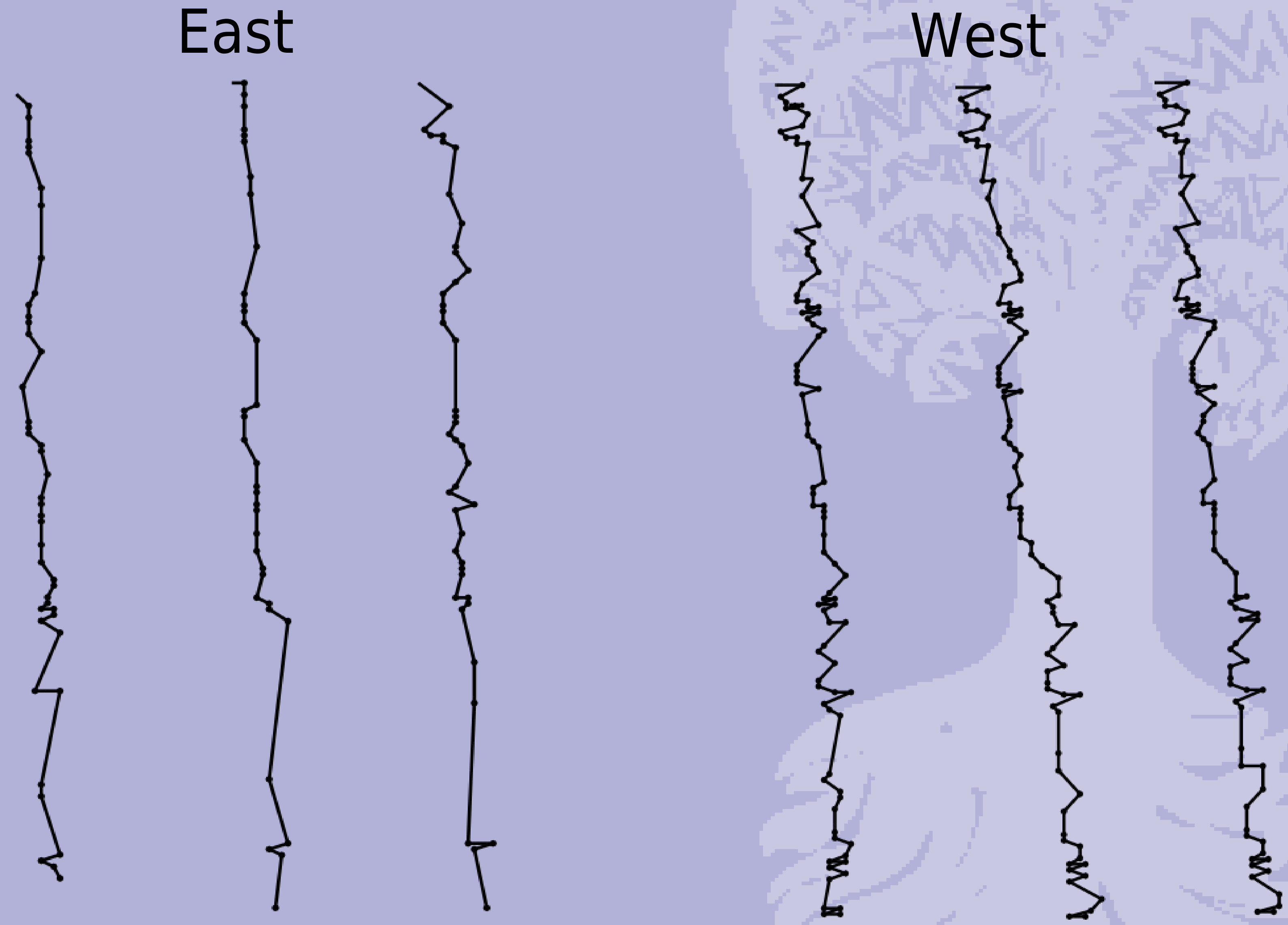
Vector Analysis

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



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



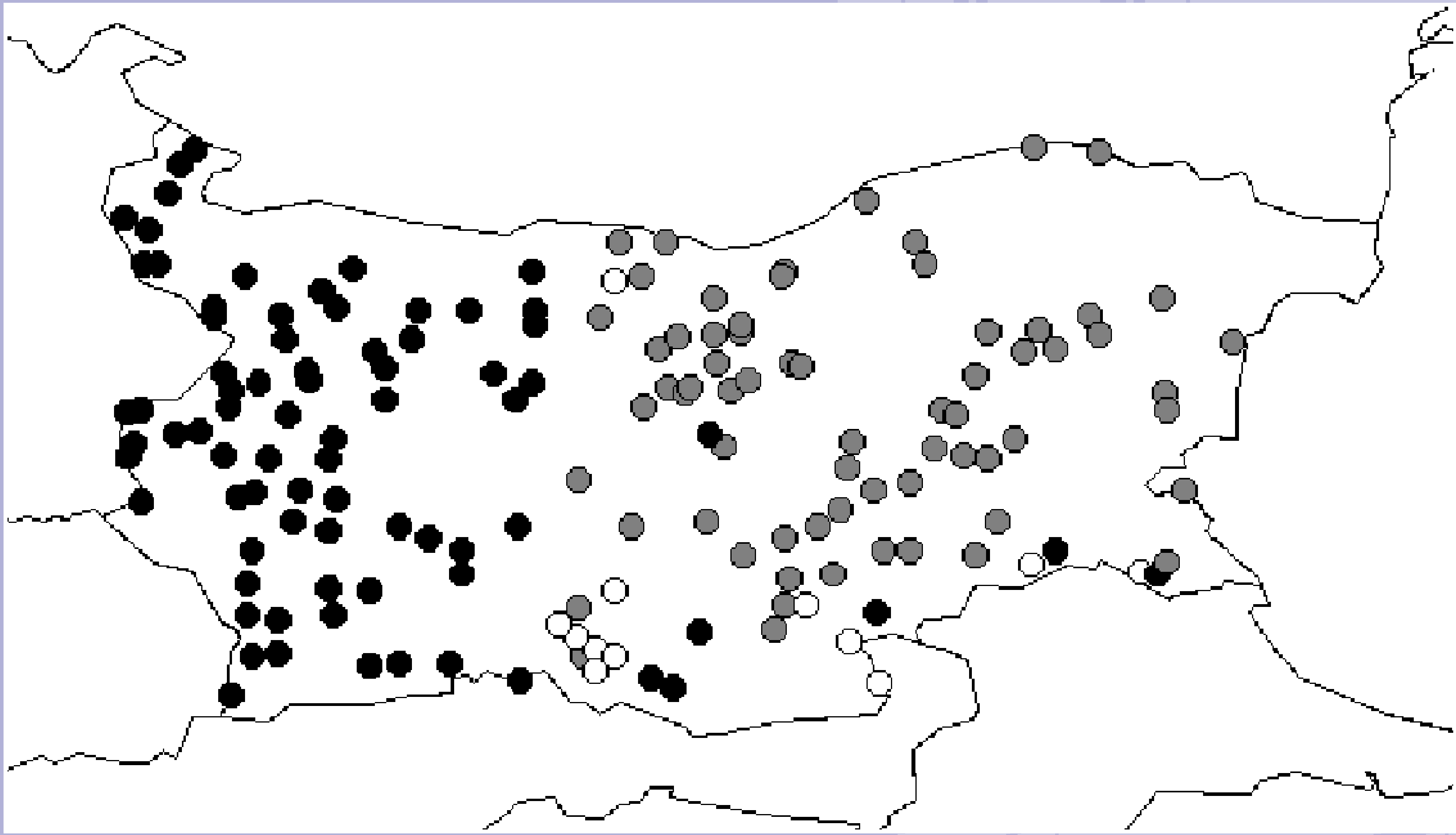

East
West

Vector 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



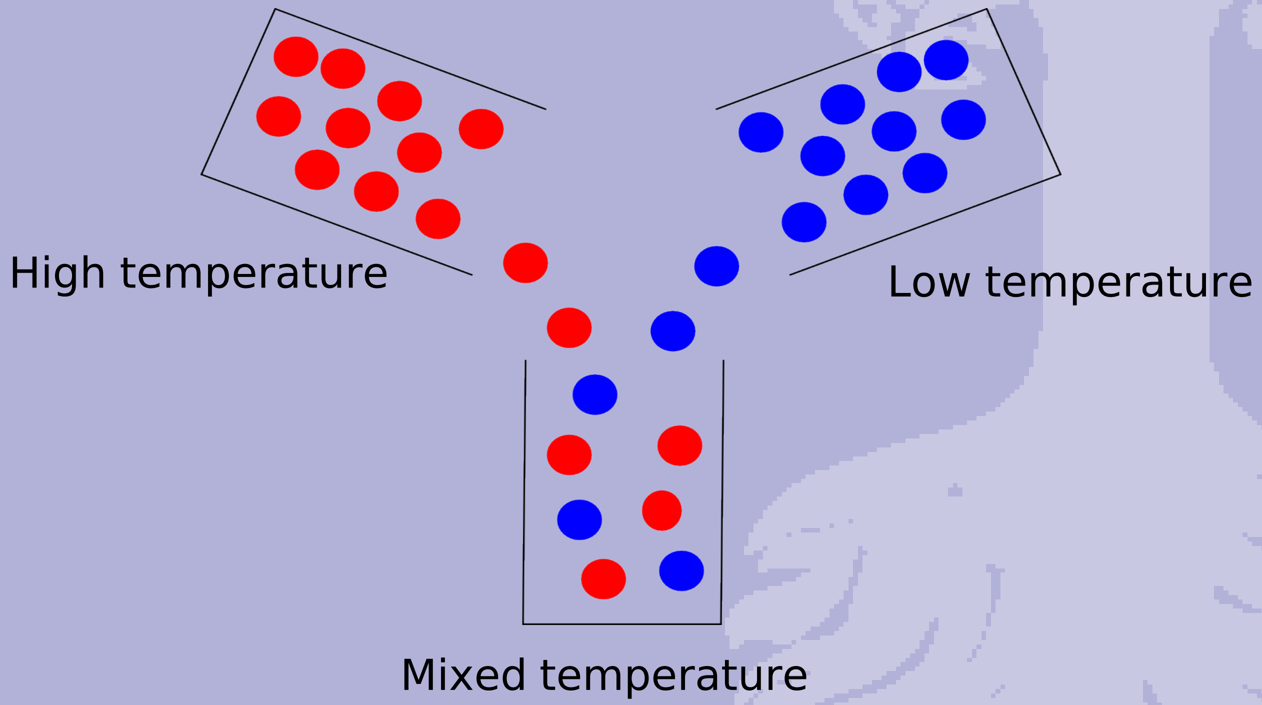
Second Approach: Entropy

The entropy is a concept which is used in a wide range of scientific fields:

- Classical physics
- Biology
- **Informationtheory**
- Entropy and ... *daemons*
- Cosmology

The Entropy in Physics

- In Thermodynamics: mixing up liquids or gas with different temperatures



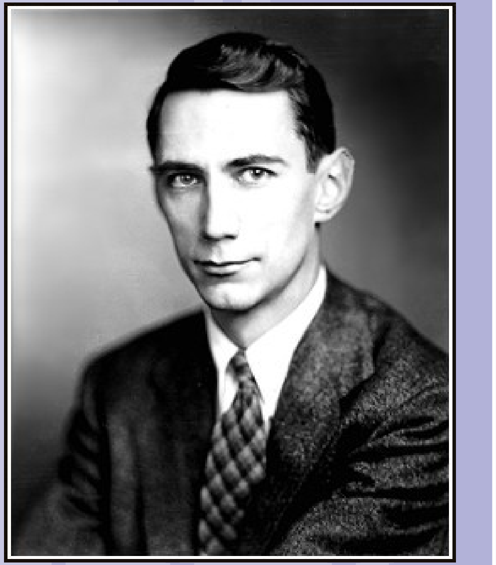
From Physics to Informationtheory

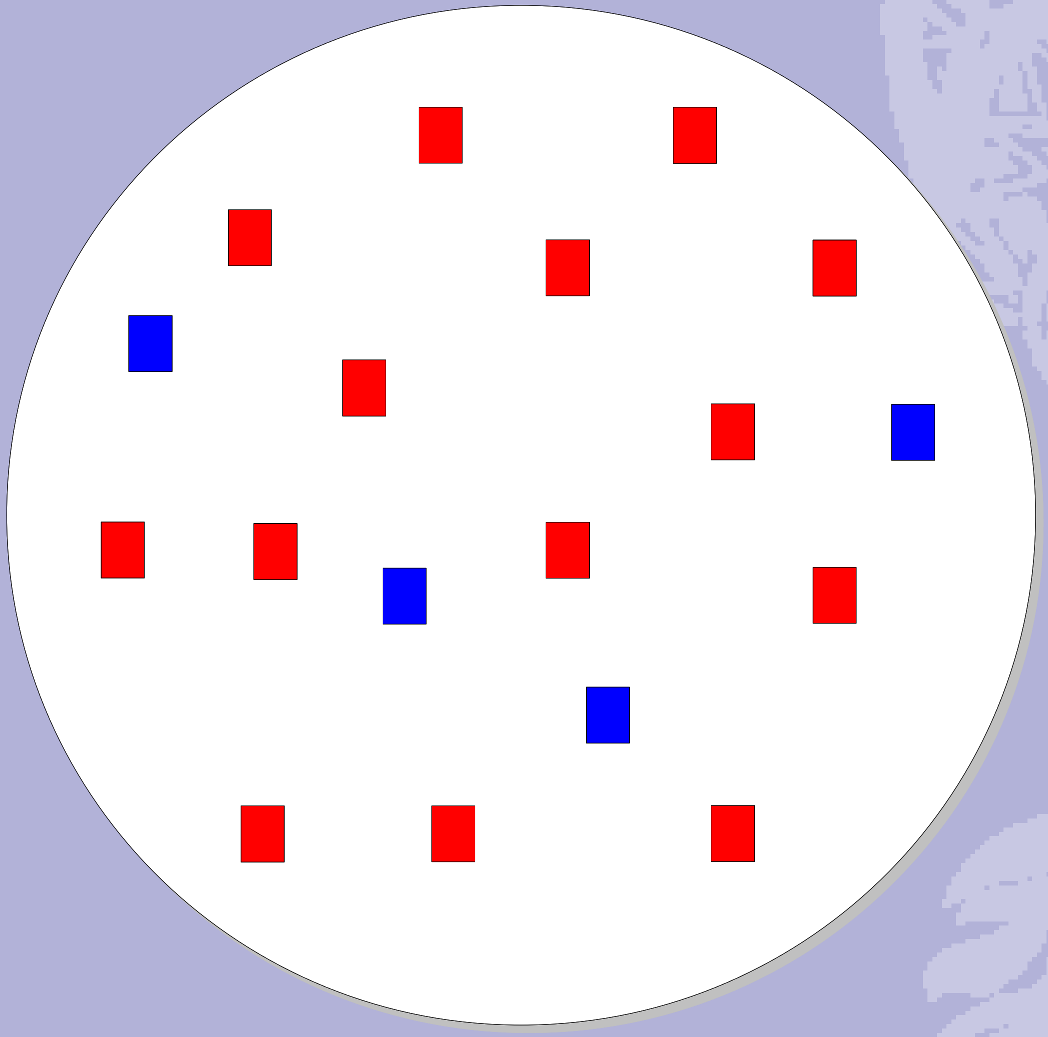
- In the 1940s, the quality of telephone systems was low
- Claude Elwood Shannon, 1948:

A Mathematical Theory of Communication

- What is the relation between

Information and ***Noise*** within data?





🟦 Information

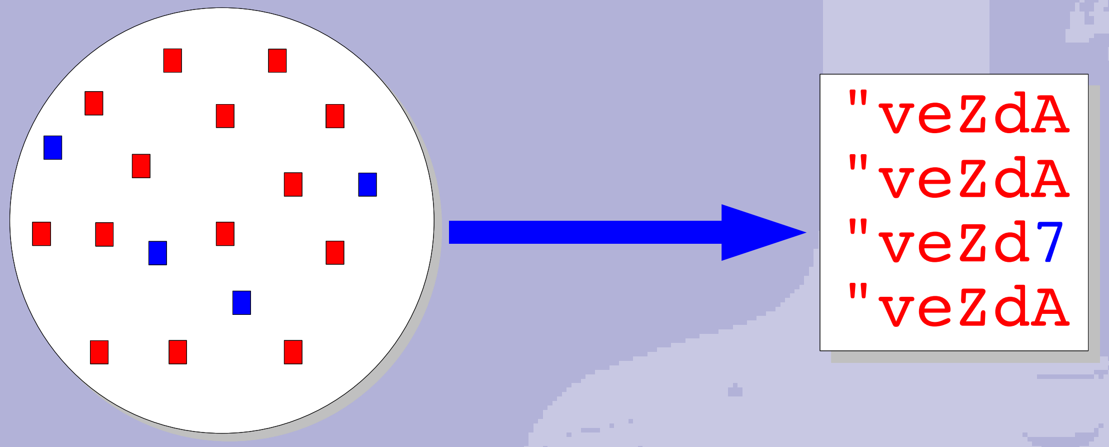
🟥 Noise

Entropy in Computational Dialectometry

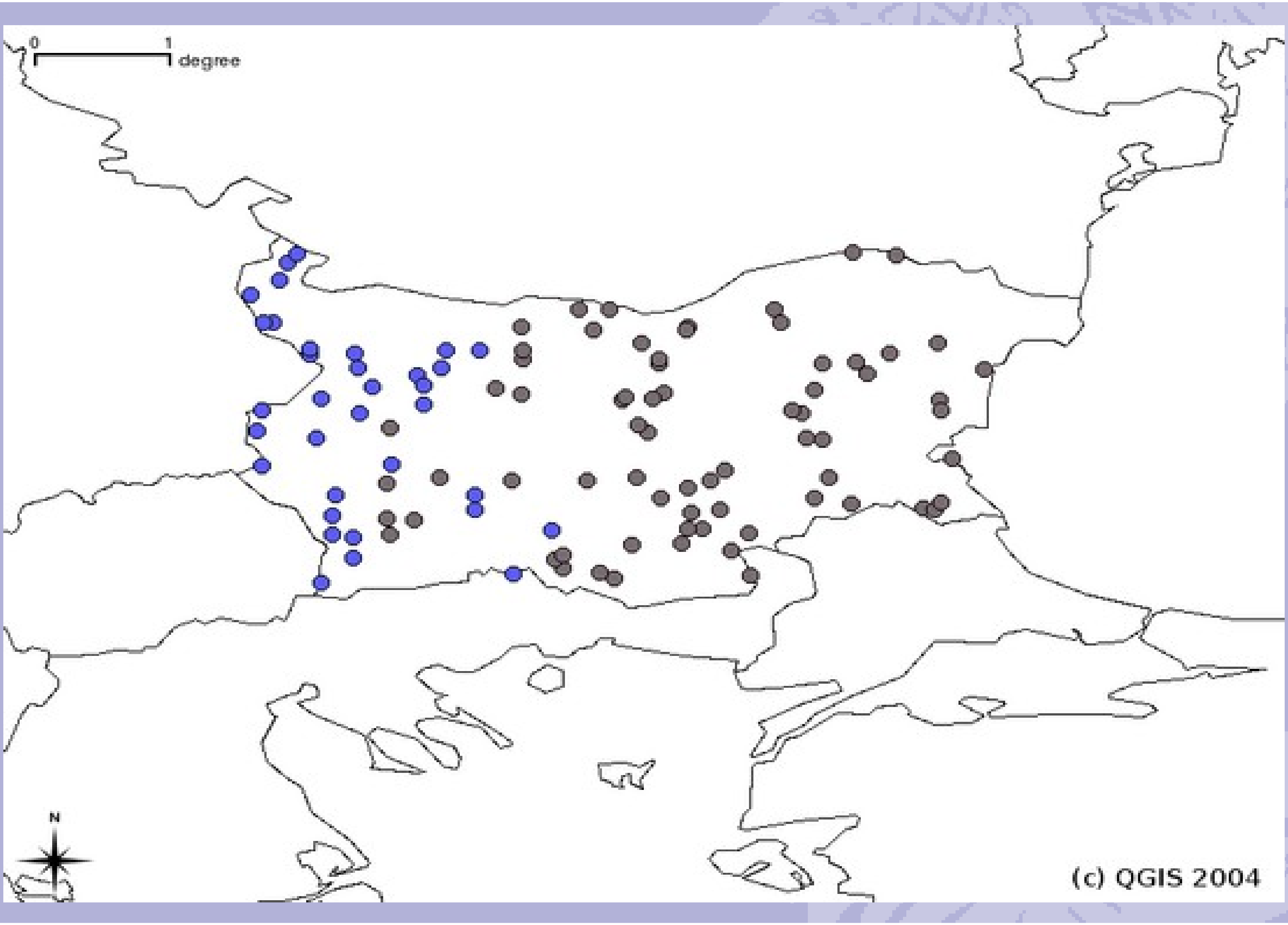
- The phonetic data of all sites are building the *data set*

🟦 Information

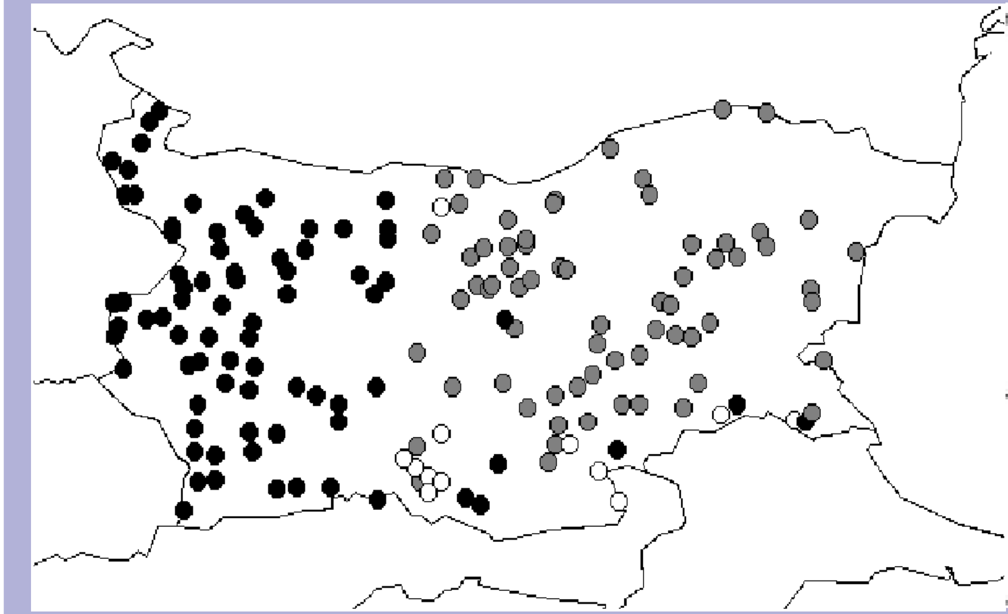
🟥 Noise



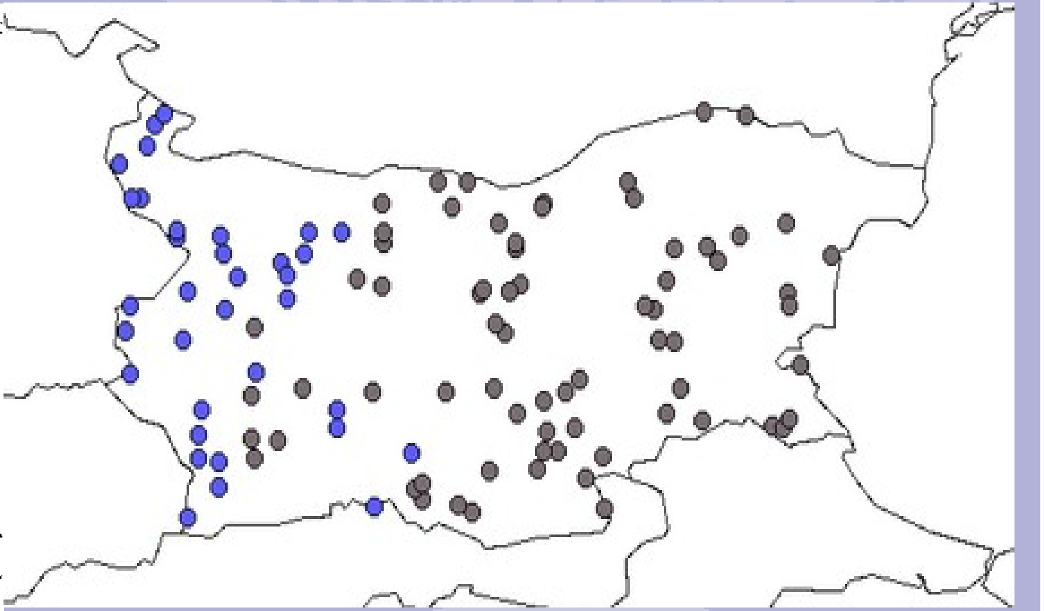
"veZdA
"veZdA
"veZd7
"veZdA


0
1
degree
N
(c) QGIS 2004

Vector Analysis



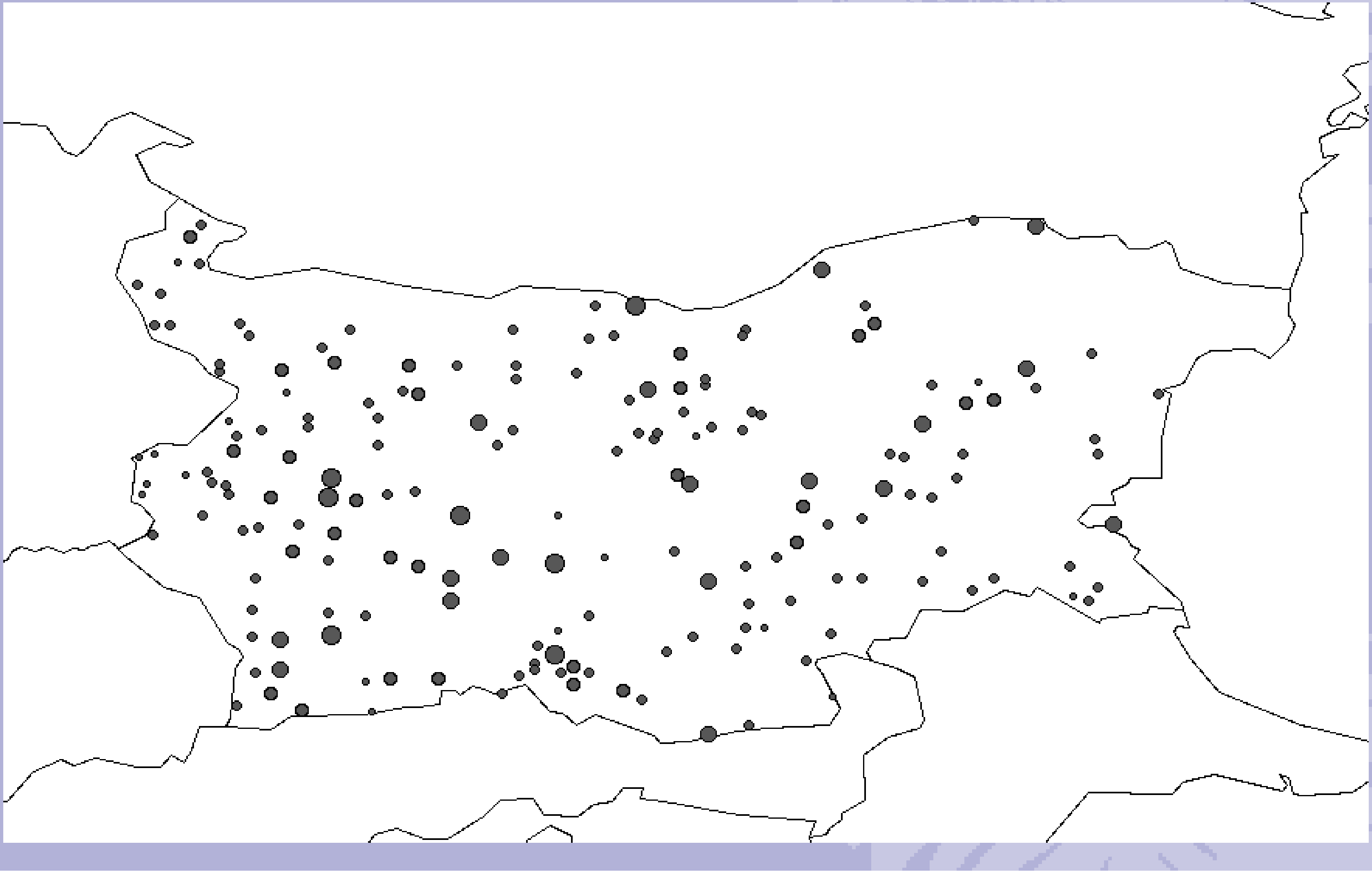
Entropy



Conclusions

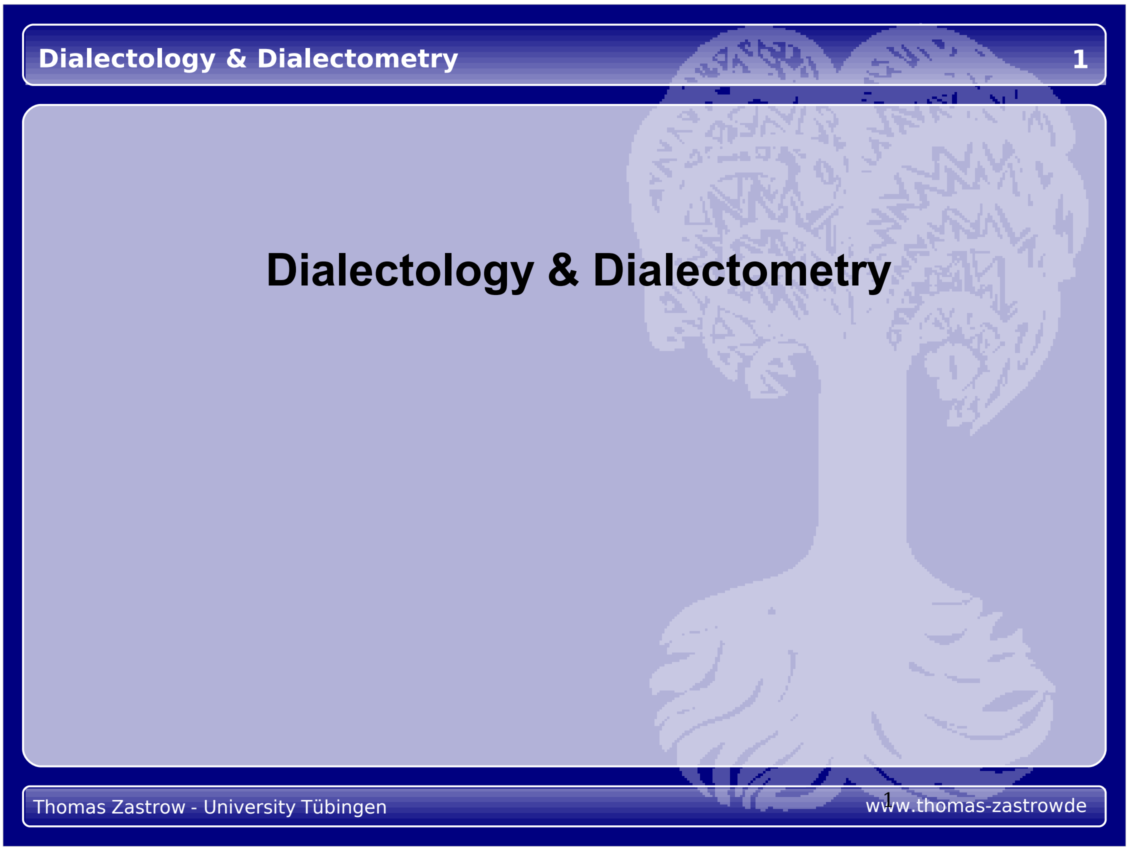
- Both methods are showing similiar, but not identical results
- An isogloss divides Bulgaria in a west and an east part
- These results are conform with the traditional dialectology

Future Work: Distances & Population



Future Work

- Extending the pronunciation data set with lexical data
- More complex clustering algorithms
- Integration of geographer's methods and data (streets, rivers, etc.)



Dialectology & Dialectometry 1

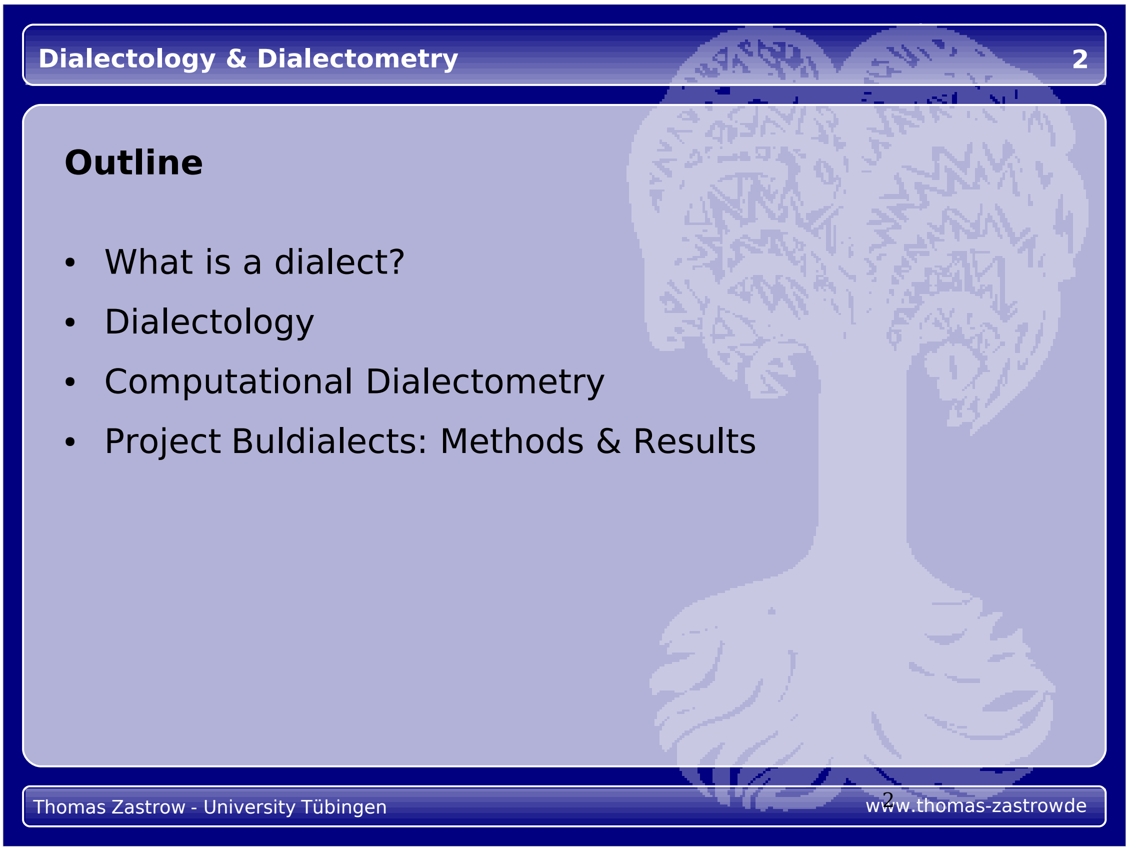
Dialectology & Dialectometry

Thomas Zastrow - University Tübingen

www.thomas-zastrowde

1

- Today, I wanna tell you something about dialects. ...



Dialectology & Dialectometry 2

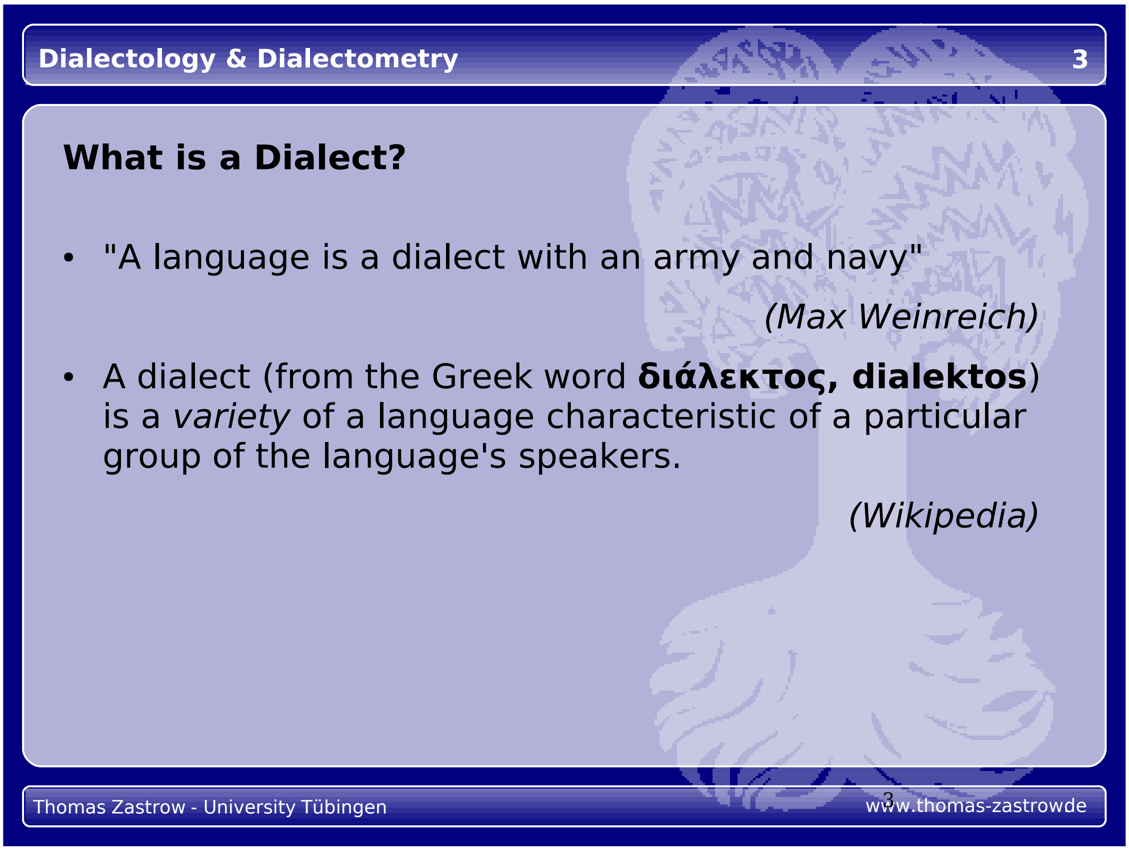
Outline

- What is a dialect?
- Dialectology
- Computational Dialectometry
- Project Buldialects: Methods & Results

Thomas Zastrow - University Tübingen 2 www.thomas-zastrowde

At first, here is the outline.
I want to explain what is a dialect? How are the borders between dialects defined and so on ...
Next topic is “traditional” dialectology
Then, Computational dialectometry which is the field where I'm working in. What is it, and what is the difference to dialectology and so on.
Then, and this is the main part of my talk, I wanna present you the project where I'm working for and - especially - the methods we have developed here in Tübingen, some results and so on.

Lets start.



Dialectology & Dialectometry 3

What is a Dialect?

- "A language is a dialect with an army and navy"

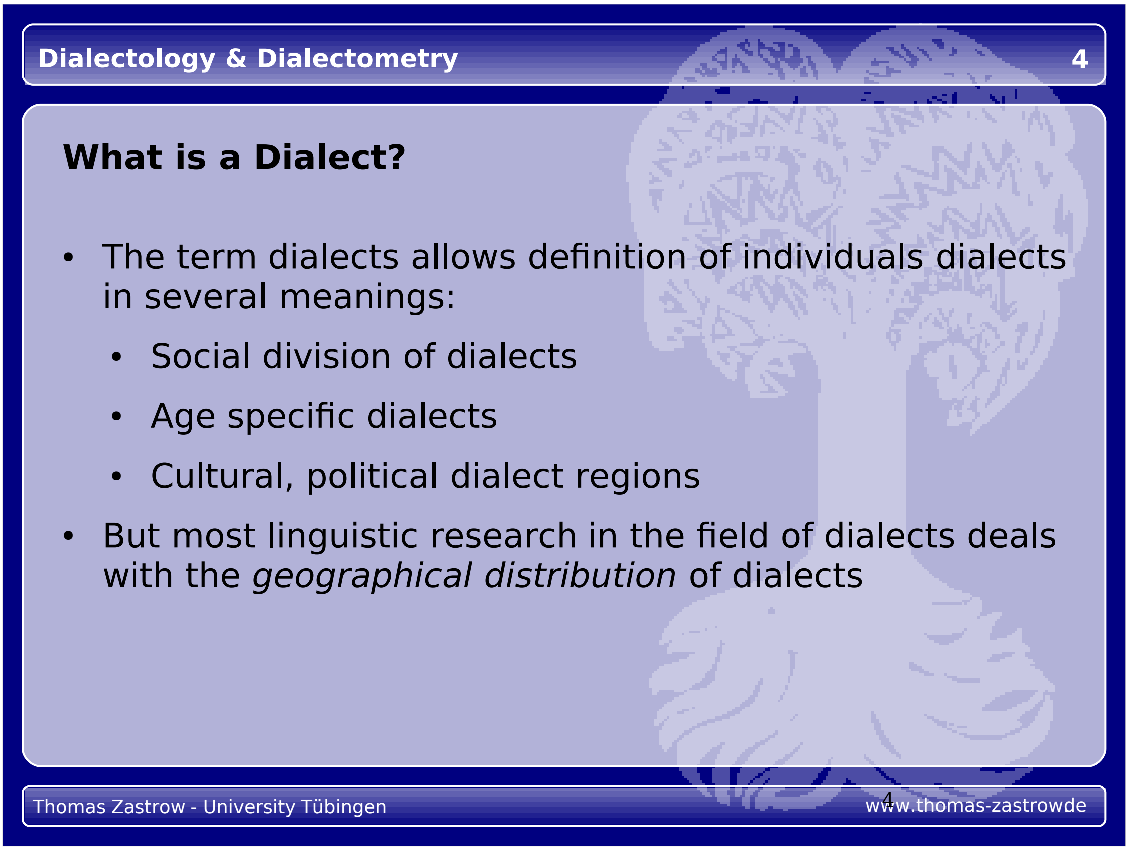
  (Max Weinreich)

- A dialect (from the Greek word **διάλεκτος, dialektos**) is a *variety* of a language characteristic of a particular group of the language's speakers.

  (Wikipedia)

Thomas Zastrow - University Tübingen 3 www.thomas-zastrowde

First and most important question: What is a dialect?
If you look at the literature, you will find explanations like “A language ...”
But is this realy correct? Think of Luxemburg: They have an army I'm not sure about the navy, but is “Letzemburgerisch” an own language ..? In my oppinnion, it is not: As a german, I understand them perfectly, they use the same vocabulary, the same grammar, the only difference is in the pronunciation.
So, wen need another definition of what a dialect is.
If you look at the wikipedia, that a dialect, from the greek word dialektos, is a variety of a language ... speakers.
And this is also the difinition which I want to use here: It is a very wide definition which allows a lot of interpretations.



Dialectology & Dialectometry 4

What is a Dialect?

- The term dialects allows definition of individual dialects in several meanings:
  - Social division of dialects
  - Age specific dialects
  - Cultural, political dialect regions
- But most linguistic research in the field of dialects deals with the *geographical distribution* of dialects

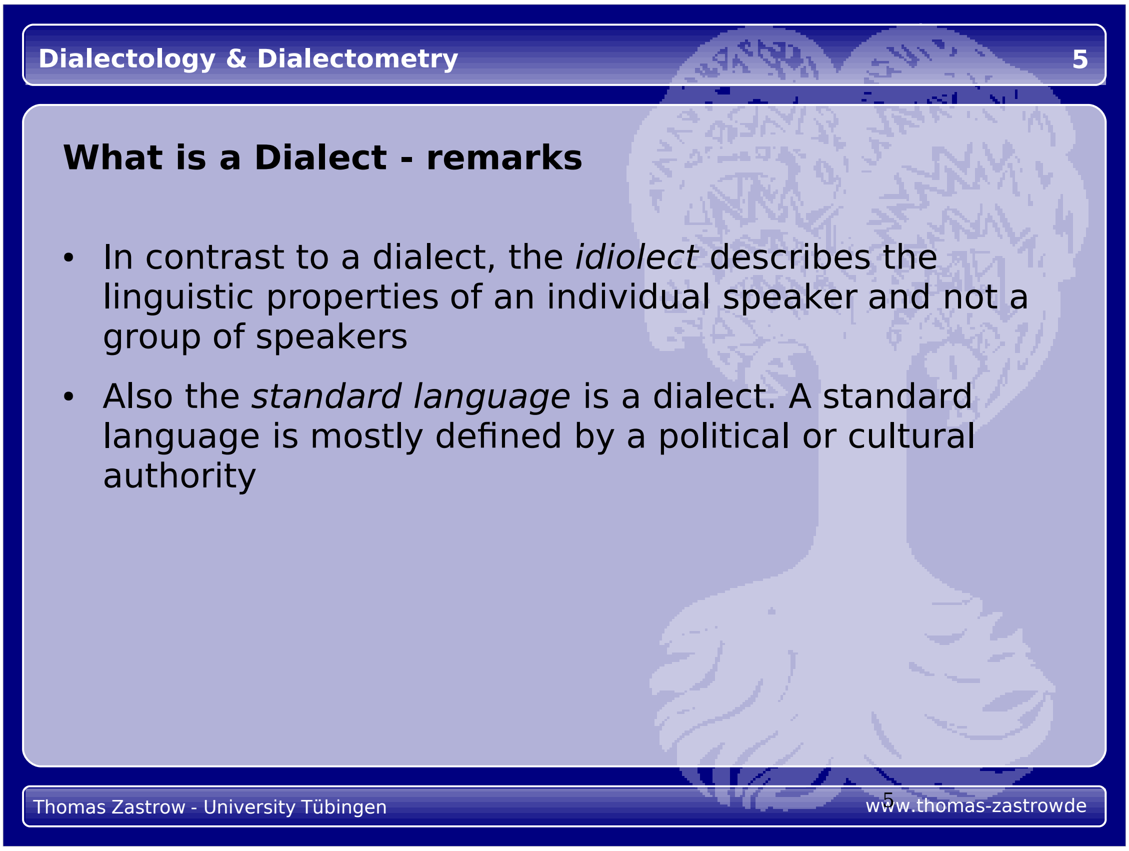
Thomas Zastrow - University Tübingen 4 www.thomas-zastrowde

Dialects can be defined in several ways: As a social division, age specific and so on and so on.

But when we speak of “dialects” in linguistic research, we normally refere to the geographical division of dialects.

That means, how can we divide a given region, for example a country like Germany with it's language German, into several dialect regions?

But please keep in mind that is also the field of socio linguistics or other specialists who are dealing with other kinds of dialects or let me say, divisions of languages.



Dialectology & Dialectometry 5

What is a Dialect - remarks

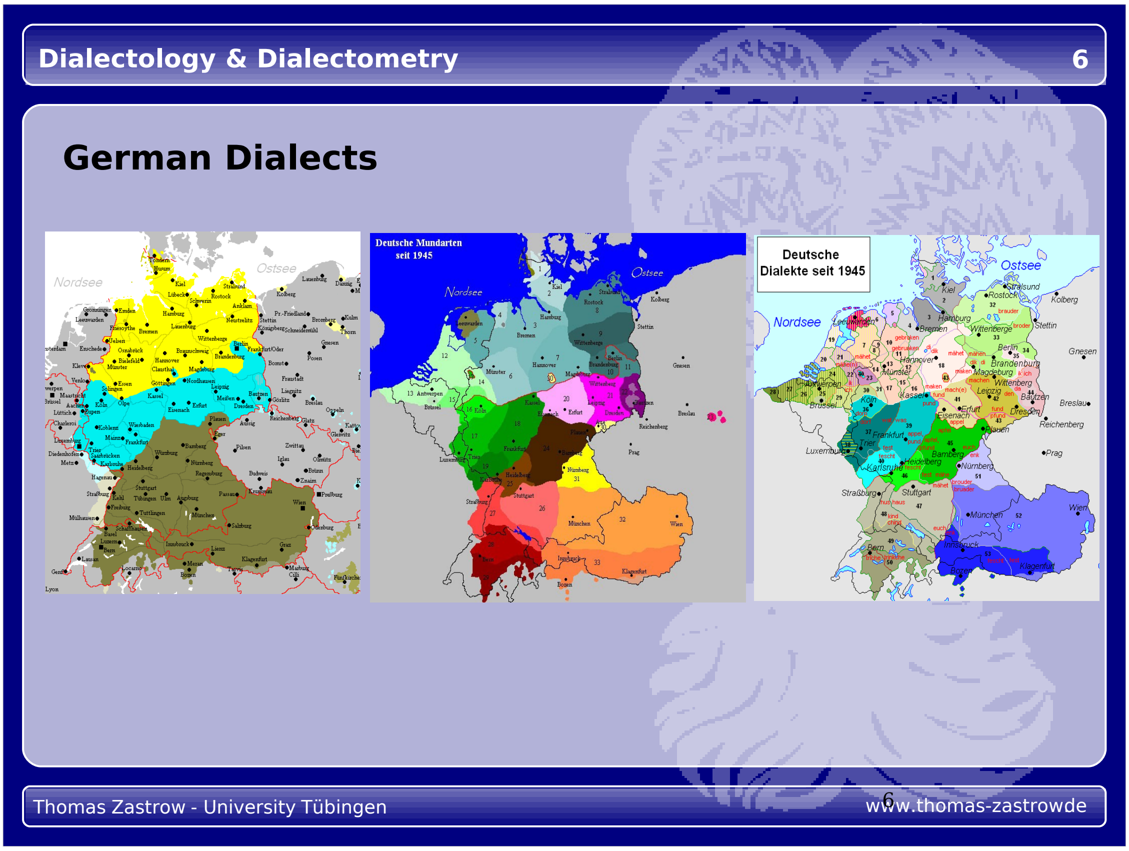
- In contrast to a dialect, the *idiolect* describes the linguistic properties of an individual speaker and not a group of speakers
- Also the *standard language* is a dialect. A standard language is mostly defined by a political or cultural authority

Thomas Zastrow - University Tübingen 5 www.thomas-zastrowde

Here are just two short remarks on our definition of a dialect: at first, the idiolect describes the speaker's individual, linguistic properties and not, as the dialect does, the properties of a whole group.

And second, which is very important, the so called standard language, for example in german the Hochdeutsch, is also nothing else than a dialect. For political, cultural or whatever reasons, this dialect has a special status within the sum of the languages dialects: it is used in books, in the press, in official statements and is also normaly taught at school.

Do you know from where the Hochdeutsch comes? And why?



Dialectology & Dialectometry 6

German Dialects

Thomas Zastrow - University Tübingen 6 www.thomas-zastrowde

Another important thing can be seen here on this slide.

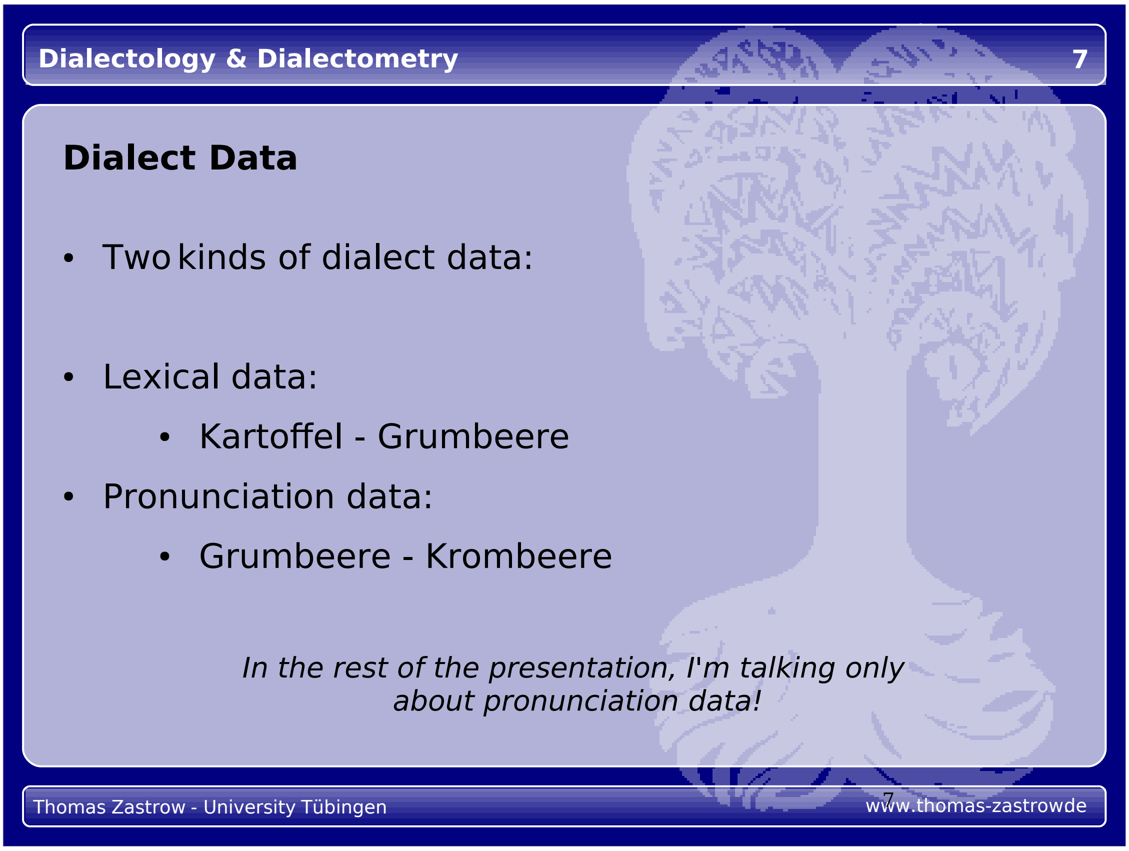
Here are three maps, all of them are representing the distribution of german dialects after WW2

The first one: Only 3 big dialects.

The scond and the third one are similiar, but not - ver important - identically.

-> See Austria!

None of this three maps are wrong. So, dividing a language into dialect regions is often an ambigous thing, depending on what criteria were used for dividing dialects.



Dialectology & Dialectometry 7

Dialect Data

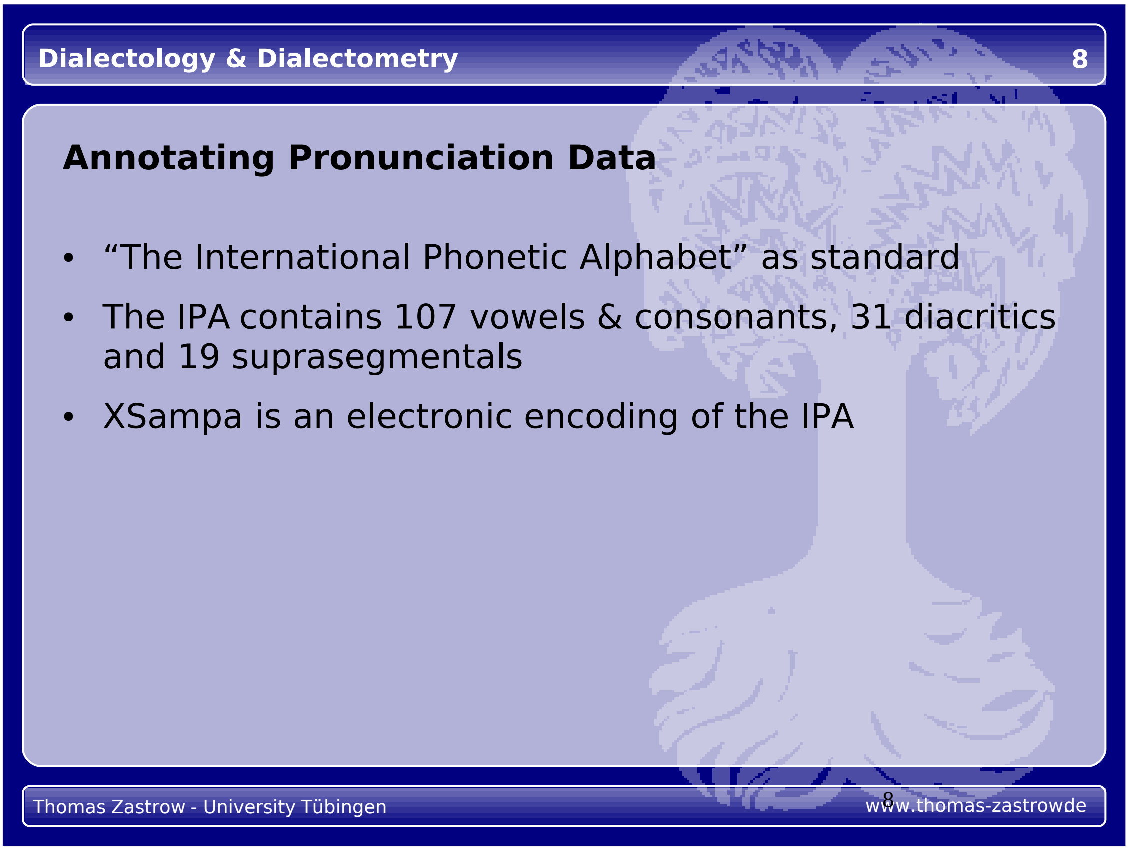
- Two kinds of dialect data:
- Lexical data:
    - Kartoffel - Grumbeere
- Pronunciation data:
    - Grumbeere - Krombeere

In the rest of the presentation, I'm talking only about pronunciation data!

Thomas Zastrow - University Tübingen 7 www.thomas-zastrowde

Lets talk about the basis of all these dialect examinations. In eneral, we know two different kinds of dialect data: Lexical data and pronunciation data. Here are examples for both of them

Kartoffel is the german word for potatoe.



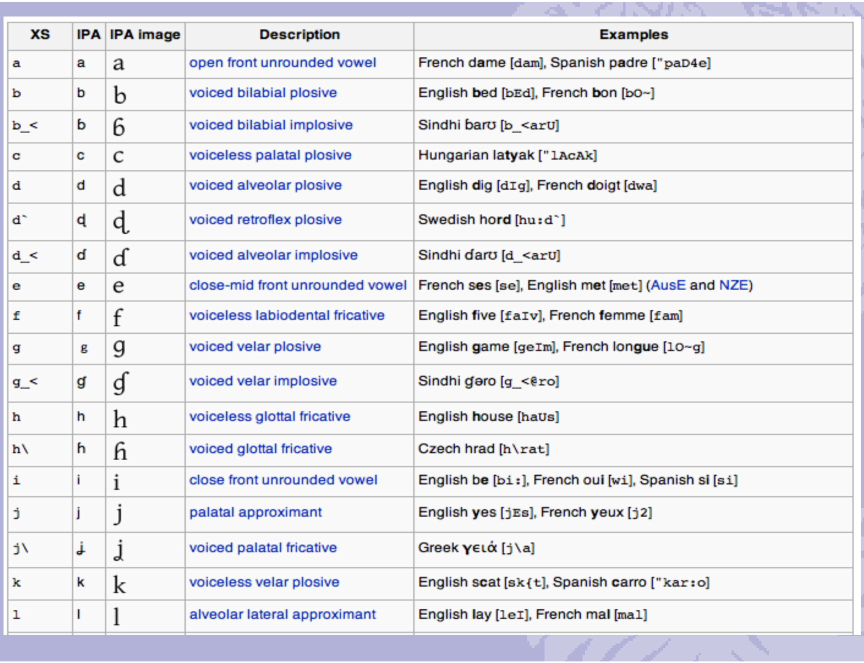
Annotating Pronunciation Data

- “The International Phonetic Alphabet” as standard
- The IPA contains 107 vowels & consonants, 31 diacritics and 19 suprasegmentals
- XSampa is an electronic encoding of the IPA

When we collect pronunciation data, we use the IPA for writing down the data. The IPA is an international standard which contains ..... in fact, with the IPA it should be possible to write down every phone which can be produced by a human beeing.

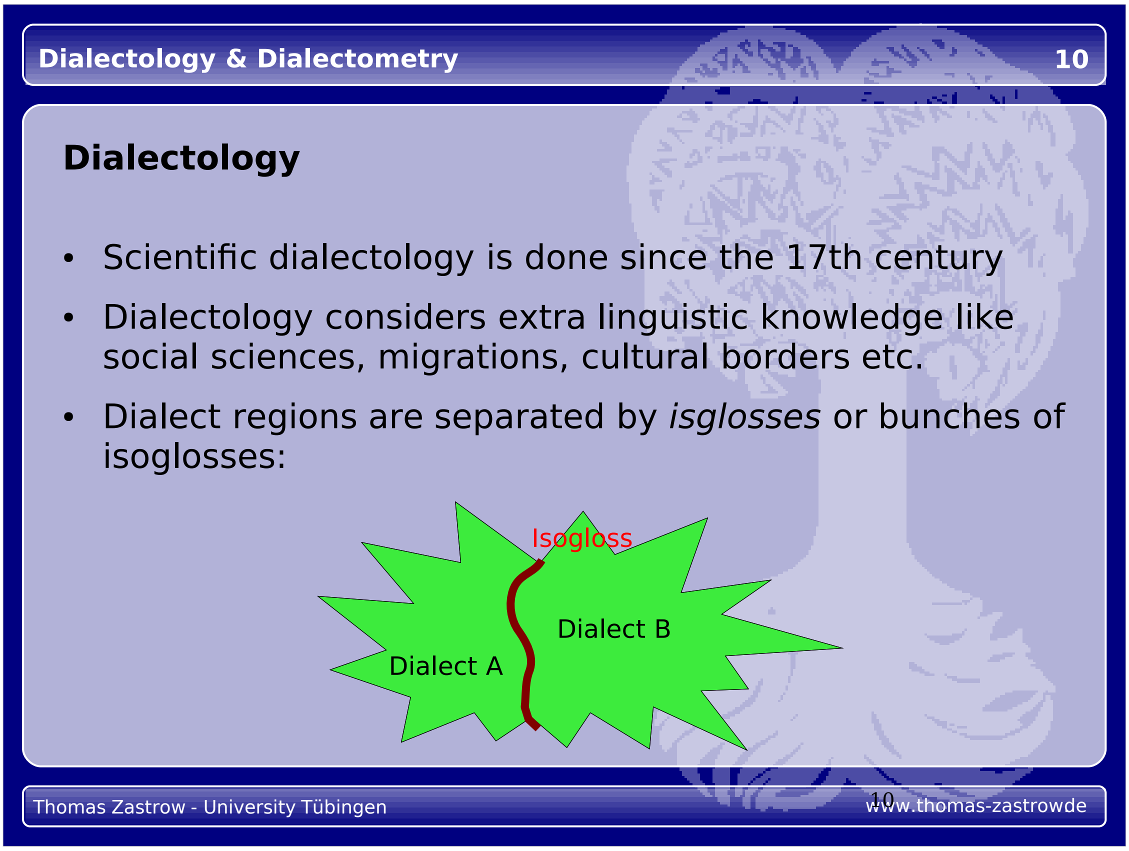
The IPA contains a lot of special symbols which can't displayed on a standard computer. So, the X-Sampa standard is a one-to-one encoding of the IPA into an electronic readable form.

Dialectology & Dialectometry



XS	IPA	IPA image	Description	Examples
a	a	a	open front unrounded vowel	French dame [dam], Spanish padre ["paD4e]
b	b	b	voiced bilabial plosive	English bed [bEd], French bon [bO~]
b_<	ɓ	ɓ	voiced bilabial implosive	Sindhi ɓarʊ [b_<arU]
c	c	c	voiceless palatal plosive	Hungarian latyak ["lAcAk]
d	d	d	voiced alveolar plosive	English dig [dIg], French doigt [dwa]
d`	ɖ	ɖ	voiced retroflex plosive	Swedish hord [hu:d`]
d_<	ɗ	ɗ	voiced alveolar implosive	Sindhi ɗarʊ [d_<arU]
e	e	e	close-mid front unrounded vowel	French ses [se], English met [met] (AusE and NZE)
f	f	f	voiceless labiodental fricative	English five [faIv], French femme [fam]
g	ɡ	ɡ	voiced velar plosive	English game [geIm], French longue [lO~g]
g_<	ɠ	ɠ	voiced velar implosive	Sindhi ɠəro [g_<@ro]
h	h	h	voiceless glottal fricative	English house [haUs]
h\	ɦ	ɦ	voiced glottal fricative	Czech hrad [h\rat]
i	i	i	close front unrounded vowel	English be [bi:], French oui [wi], Spanish si [si]
j	j	j	palatal approximant	English yes [jEs], French yeux [j2]
j\	ʝ	ʝ	voiced palatal fricative	Greek γειά [j\a]
k	k	k	voiceless velar plosive	English scat [sk{t], Spanish carro ["kar:o]
l	l	l	alveolar lateral approximant	English lay [leI], French mal [mal]

Here we have some symbols and their explanation of the IPA. The XSampa code is here in the first ....



Dialectology

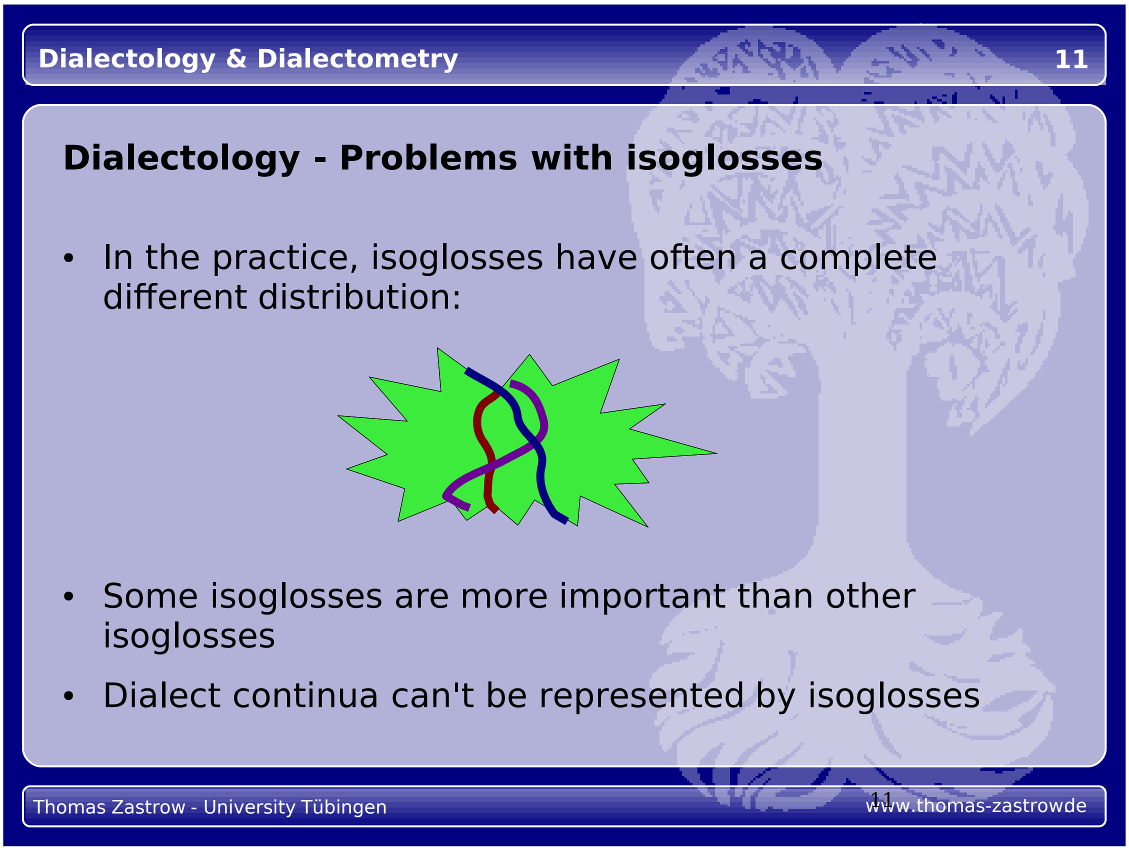
- Scientific dialectology is done since the 17th century
- Dialectology considers extra linguistic knowledge like social sciences, migrations, cultural borders etc.
- Dialect regions are separated by *isglosses* or bunches of isoglosses:

Now, lets come to Dialectology. D is an old science, the first dialectological investigations were done in the 17th century. In the middle of the 19th century, the first, realy scientific researchs were done.

In D, extralinguistic knowledge was always involved to identify dialect regions. This includes knowledge about migration movements, cultural and social borders and so on.

These knowledge, together with the dialect data, is used to divide one dialect region from the other. And the border between those regions is called an isogloss, similiar to the terms which are used in geograpic to identify different levels of a landscape.

So, for example, one very important isogloss here in the south of germany is the “le” isogloss. The use of “le” in the end of a word divides the swabian dialect from the surrounding ones.

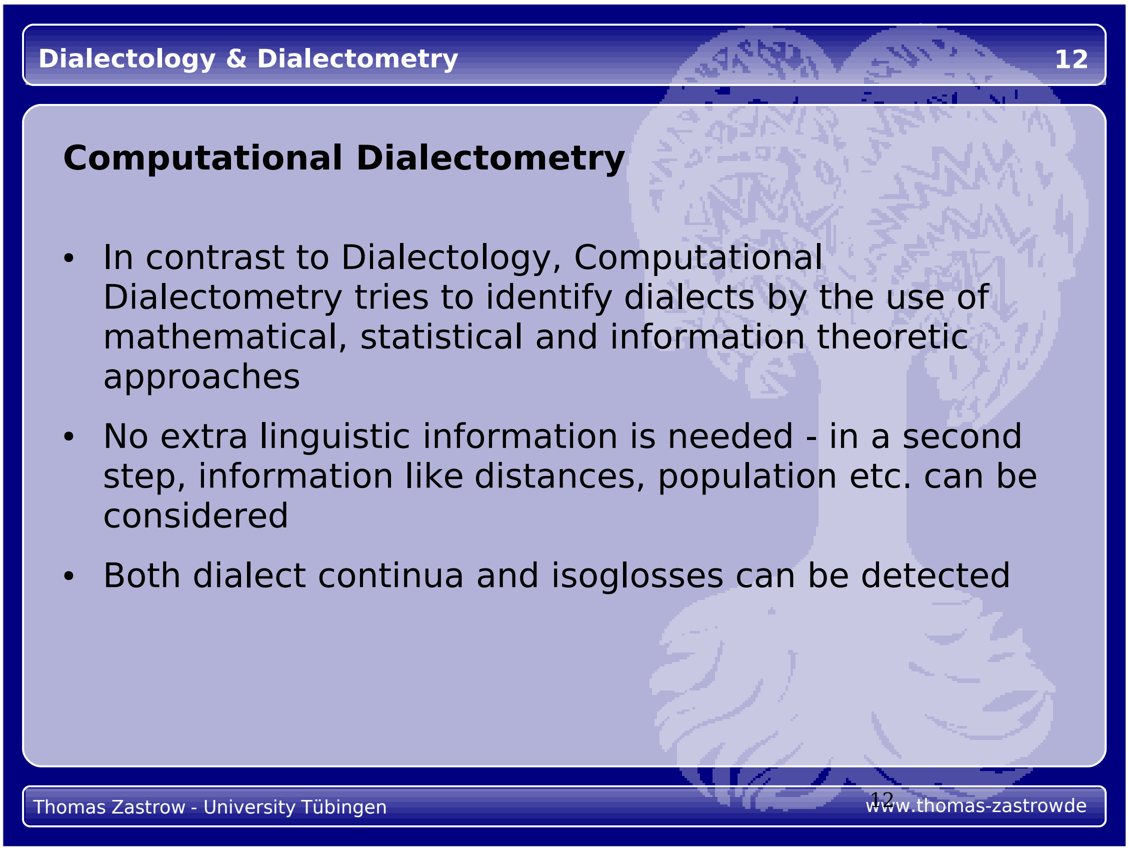


Dialectology - Problems with isoglosses

- In the practice, isoglosses have often a complete different distribution:
- Some isoglosses are more important than other isoglosses
- Dialect continua can't be represented by isoglosses

When using isoglosses, some problems appears. At first, different isoglosses are often showing a complete different division of the language's area.
Second, it is a fact that some isoglosses are more important than other ones, for example the “le” isogloss is a very important one here in the south.

And the most important problem when using isoglosses is the fact, that they are not able to display dialect continua. In a dialect continua, the transition from one dialect to the other is not a hard break, instead of that it is soft, floating one and you can't say exactly where a dialect ends and where the next one begins.
So, we need some other methods for identifying dialects.



Dialectology & Dialectometry 12

Computational Dialectometry

- In contrast to Dialectology, Computational Dialectometry tries to identify dialects by the use of mathematical, statistical and information theoretic approaches
- No extra linguistic information is needed - in a second step, information like distances, population etc. can be considered
- Both dialect continua and isoglosses can be detected

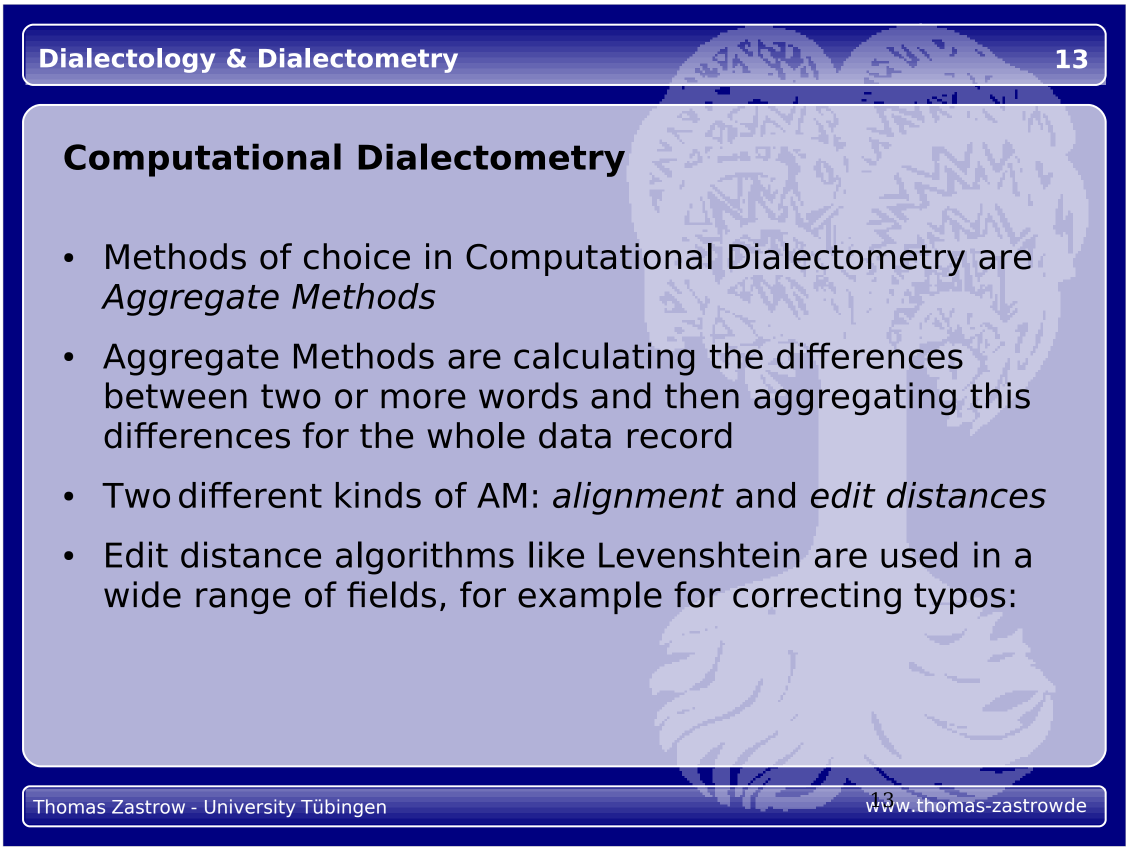
Thomas Zastrow - University Tübingen 12 www.thomas-zastrowde

Computational Dialectometry is a slightly other approach to dialects than dialectology. The name “CD” contains the string “metry” which means that we try to measure the division of dialects.

We try to do this with mathematical, statistical and information theoretic apporaches. So, now it starts to get a little bit more mathematical, but don't worry, I have no formular in this presentation.

In CD, we use only linguistic data and no extra linguistic data. This can be done in a second step.

And - very important - with these mathematical methods we are able to detect both, isoglosses and dialect continua.



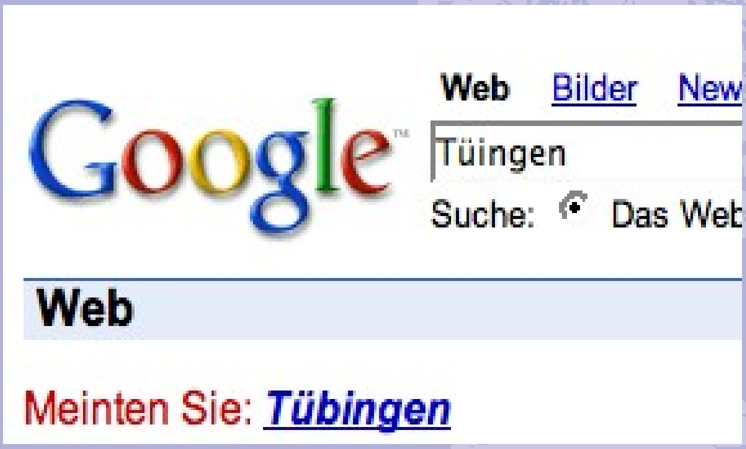
Computational Dialectometry

- Methods of choice in Computational Dialectometry are *Aggregate Methods*
- Aggregate Methods are calculating the differences between two or more words and then aggregating this differences for the whole data record
- Two different kinds of AM: *alignment* and *edit distances*
- Edit distance algorithms like Levenshtein are used in a wide range of fields, for example for correcting typos:

At the moment, aggregate methods like alignment and edit distances are the methods of choice. Especially the edit distances were adopted to CD by John Nerbonne and his team at the university Groningen, we are working together with them.

This methods are called aggregate methods because they are calculating the differences between two or more words and then these differences are summed up for the whole data set.

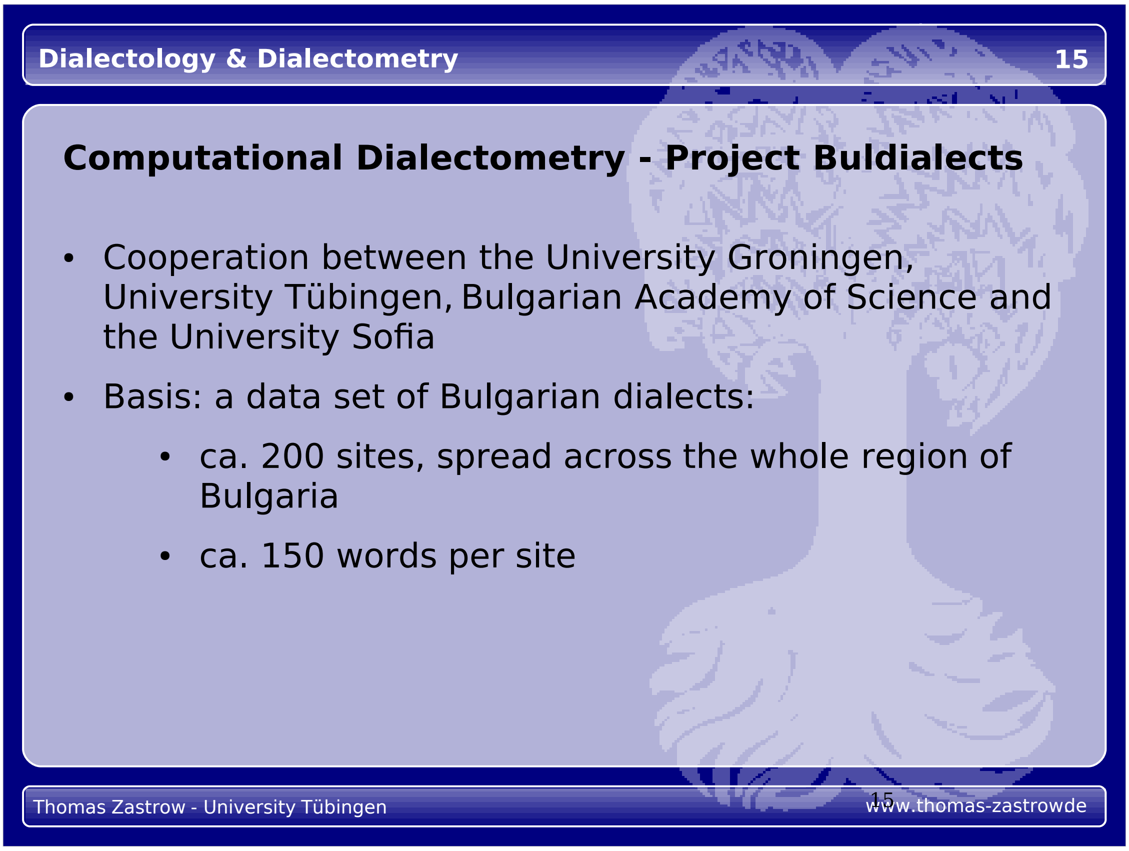
I think that everybody of you had already contact with edit distance algorithms: Often they are used in word processors, search engines etc. for detecting typos.



Tüingen

Meinten Sie: *Tübingen*

So, here is an example how a big search engines identifies typos: TÜINGEN TÜBINGEN, the algorithm which is used here to detect similiar words to what was typed in the text field is nothing else than the Levenshtein Algorithm.



Dialectology & Dialectometry 15

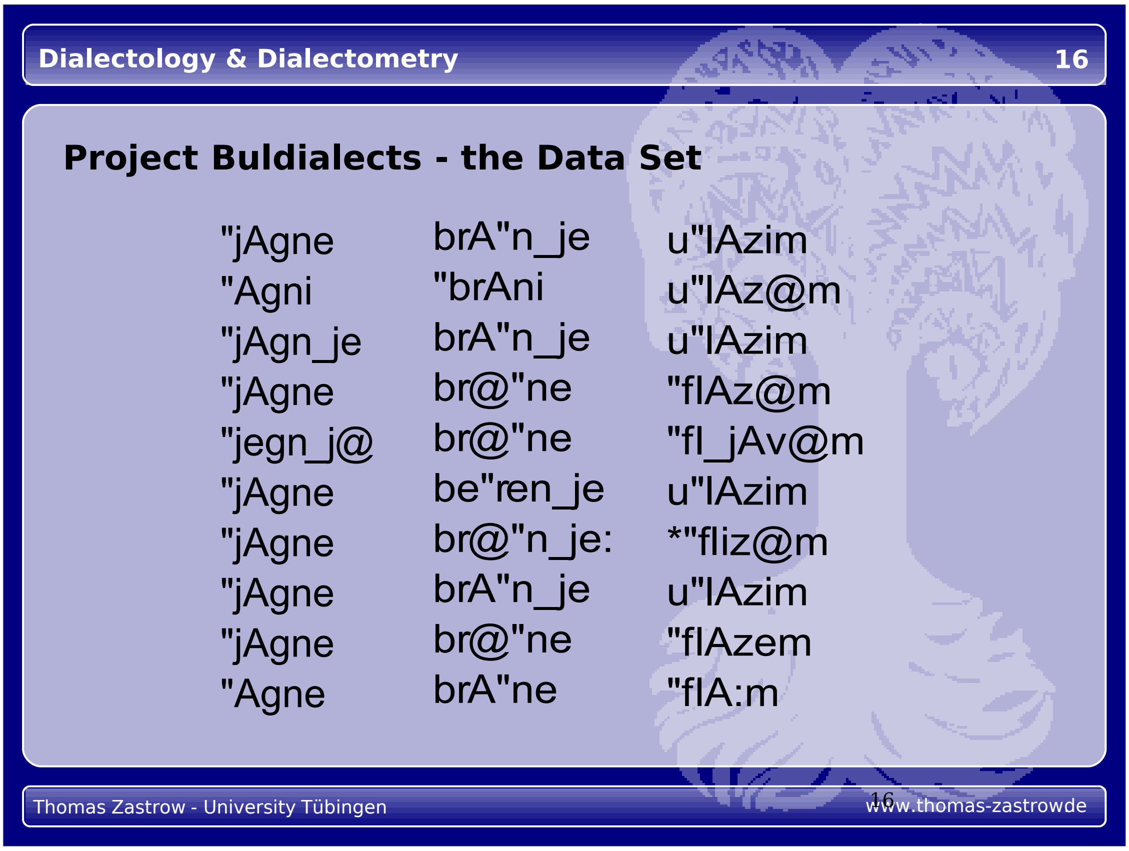
Computational Dialectometry - Project Buldialects

- Cooperation between the University Groningen, University Tübingen, Bulgarian Academy of Science and the University Sofia
- Basis: a data set of Bulgarian dialects:
  - ca. 200 sites, spread across the whole region of Bulgaria
  - ca. 150 words per site

Thomas Zastrow - University Tübingen www.thomas-zastrowde

Now, I'm comming to the project Buldialects, where I'm working in here at the university tübingen.
The Budlialects project is a cooperation between ....

Our Bulgarian colleguas providing us with a big data set of Bulgarian dialects: We have 197 sites, which can be small villages or even big cities in Bulgaria and 151 words for every site, annotated as pronunciation data in XSampa.



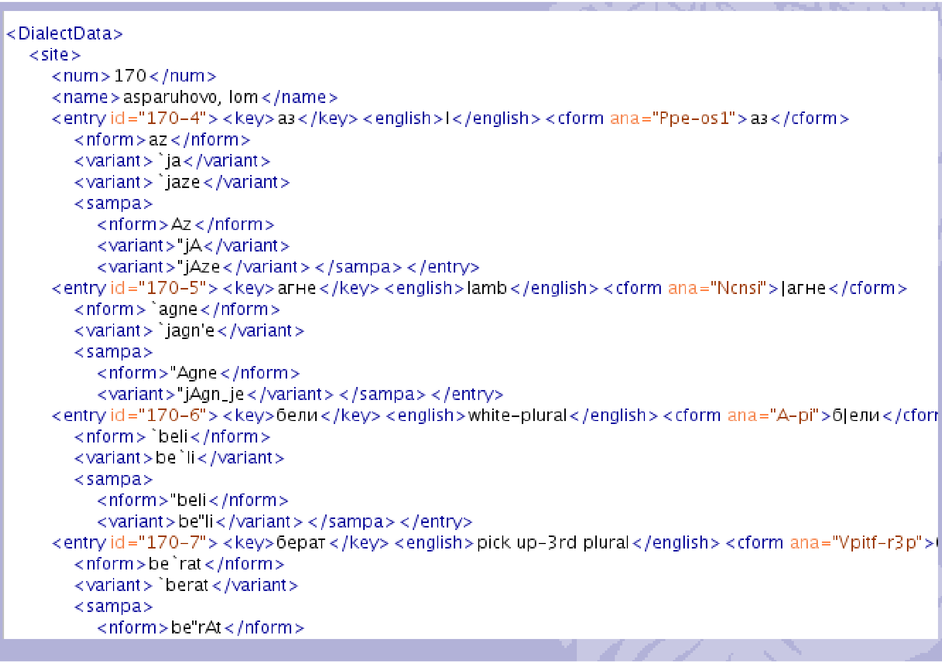
Dialectology & Dialectometry 16

Project Buldialects - the Data Set

"jAgne	brA"n_je	u"lAzim
"Agni	"brAni	u"lAz@m
"jAgn_je	brA"n_je	u"lAzim
"jAgne	br@"ne	"flAz@m
"jegn_j@	br@"ne	"fl_jAv@m
"jAgne	be"ren_je	u"lAzim
"jAgne	br@"n_je:	*"fliz@m
"jAgne	brA"n_je	u"lAzim
"jAgne	br@"ne	"flAzem
"Agne	brA"ne	"flA:m

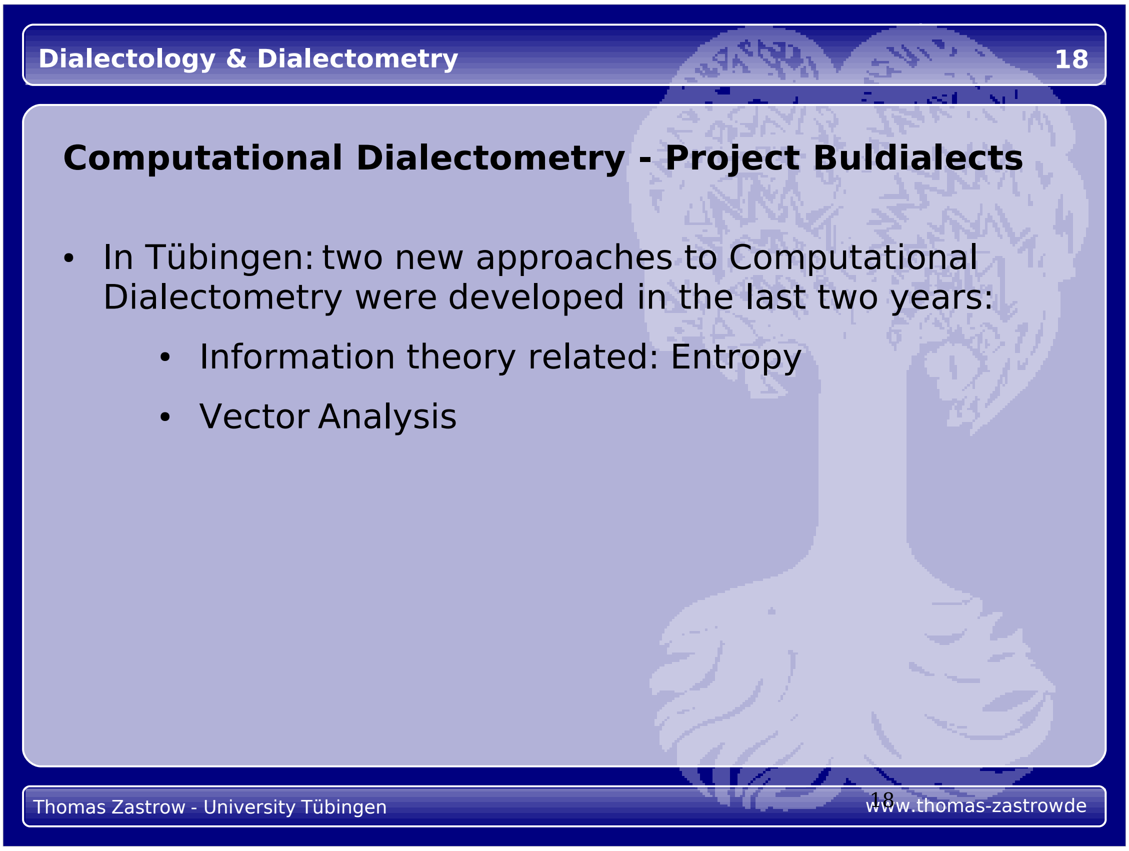
Thomas Zastrow - University Tübingen 16 www.thomas-zastrowde

So here is an example of the Bulgarian data set, I'm not able to read any word in Bulgarian, so I'm just the programmer, but perhaps one of you can tell us something about the words which are used here.



```
<DialectData>
  <site>
    <num>170</num>
    <name>asparuhovo, lom</name>
    <entry id="170-4"><key>аз</key><english>I</english><cform ana="Ppe-os1">аз</cform>
      <nform>az</nform>
      <variant>`ja</variant>
      <variant>`jaze</variant>
      <sampa>
        <nform>Az</nform>
        <variant>"jA</variant>
        <variant>"jAze</variant></sampa></entry>
    <entry id="170-5"><key>агне</key><english>lamb</english><cform ana="Ncnsi">|агне</cform>
      <nform>`agne</nform>
      <variant>`jagn'e</variant>
      <sampa>
        <nform>"Agne</nform>
        <variant>"jAgn_je</variant></sampa></entry>
    <entry id="170-6"><key>бели</key><english>white-plural</english><cform ana="A-pi">б|ели</cfor
      <nform>`beli</nform>
      <variant>be`li</variant>
      <sampa>
        <nform>"beli</nform>
        <variant>be"li</variant></sampa></entry>
    <entry id="170-7"><key>берат</key><english>pick up-3rd plural</english><cform ana="Vpitf-r3p">
      <nform>be`rat</nform>
      <variant>`berat</variant>
      <sampa>
        <nform>be"rAt</nform>
```

Just a short excourse into the technics: The data is encoded in XML which is a very wide used data format in Computer Linguistics. I think, during the next years you will often have to deal with XML data ;-)



Dialectology & Dialectometry 18

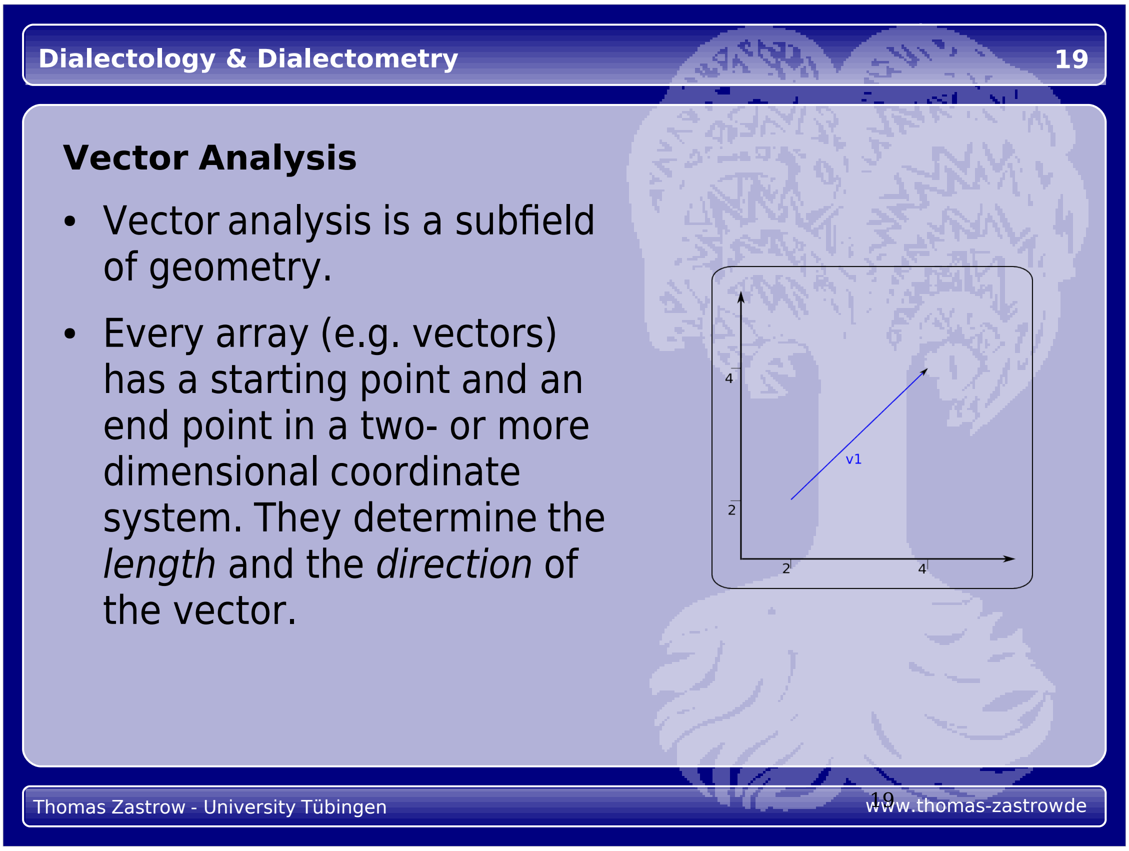
Computational Dialectometry - Project Buldialects

- In Tübingen: two new approaches to Computational Dialectometry were developed in the last two years:
    - Information theory related: Entropy
    - Vector Analysis

Thomas Zastrow - University Tübingen 18 www.thomas-zastrowde

Now I'm starting with the main part of my preentation.

Here in Tübingen, we developed two new approaches to CD in the last two years. One of them is the Entropy which is an information theory related approach and the other one is a geometrical approach. Vector analysis.



Dialectology & Dialectometry 19

Vector Analysis

- 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.

Thomas Zastrow - University Tübingen

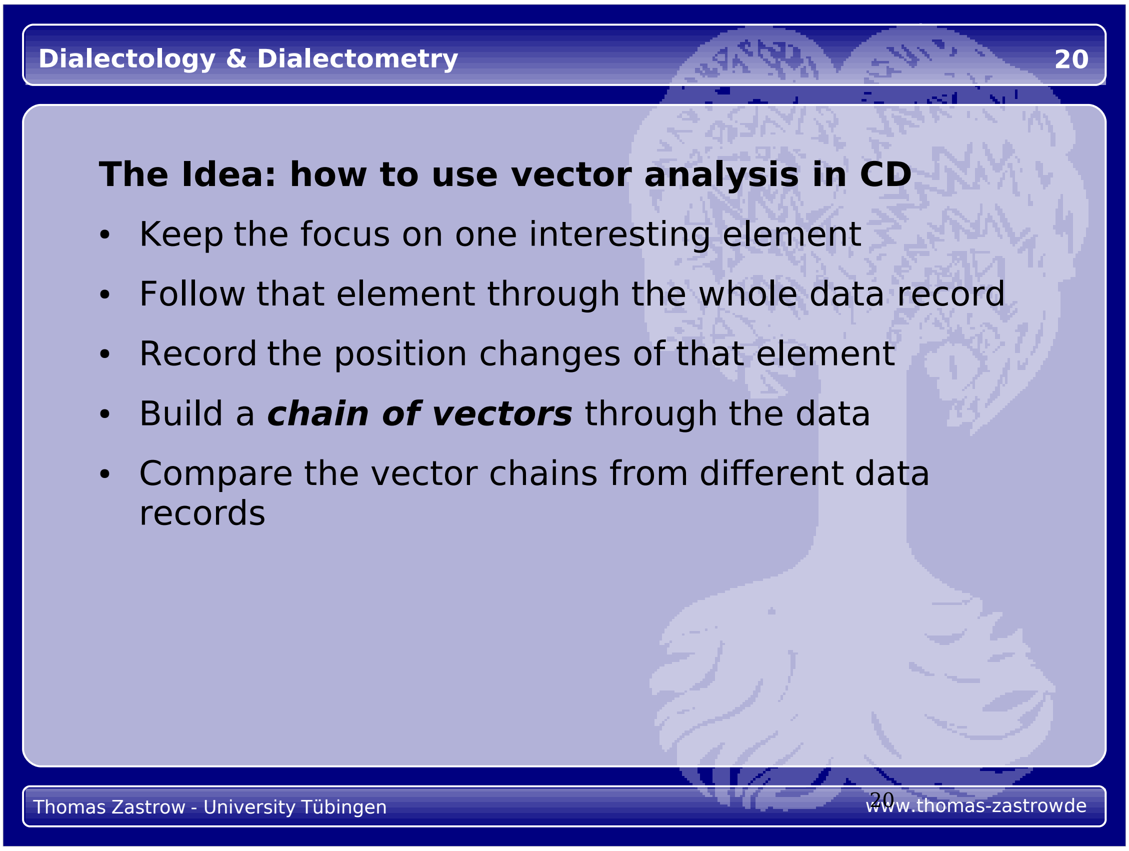
19

www.thomas-zastrowde

At first, the vector analysis.

VA is a subfield of geometry. And a vector is nothing else than an array: Each array has a starting- and an end point in a two- or more dimensional space. With these points and the direction of the vector, the vector is completely defined.

Here in CD we only use 2 dimensional vectors, but vectors in higher dimensions are also possible.



Dialectology & Dialectometry 20

The Idea: how to use vector analysis in CD

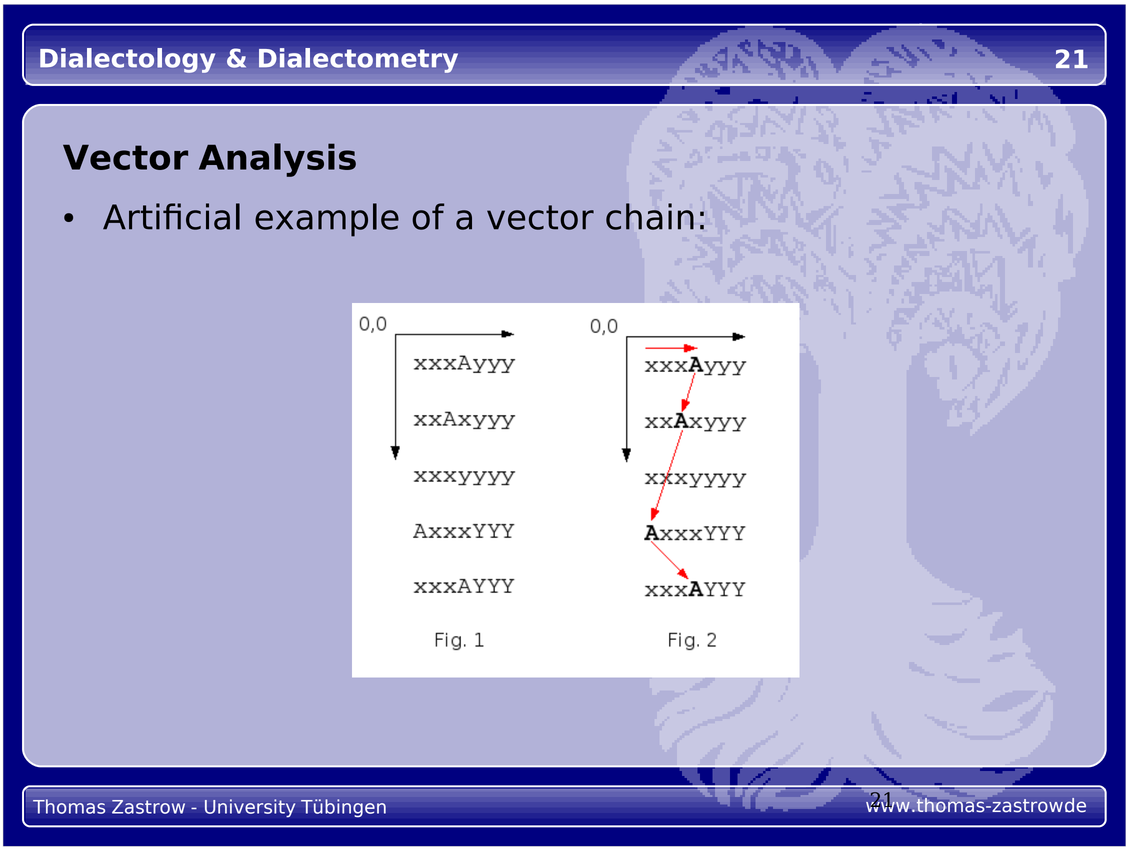
- Keep the focus on one interesting element
- Follow that element through the whole data record
- Record the position changes of that element
- Build a ***chain of vectors*** through the data
- Compare the vector chains from different data records

Thomas Zastrow - University Tübingen 20 www.thomas-zastrowde

The idea I had now, is to keep the focus on one interesting element, for example a vowel. And then we draw vectors from the first to the second appearance of that element and so on and so on. In this way, we record the position changes of that element through the data set.

If we are finished, we can build a chain of vectors by chaining all these single vectors together to one big one.

And last not least, if we did that for every site or wordlist, we can compare the vector chains with mathematical methods.



Dialectology & Dialectometry 21

Vector Analysis

- Artificial example of a vector chain:

0,0

xxxAyyy

xxAxyyy

xxxyyyy

AxxxYYY

xxxAYYY

Fig. 1

0,0

xxx**A**yyy

xx**A**xyyy

xxxyyyy

AxxxYYY

xxx**A**YYY

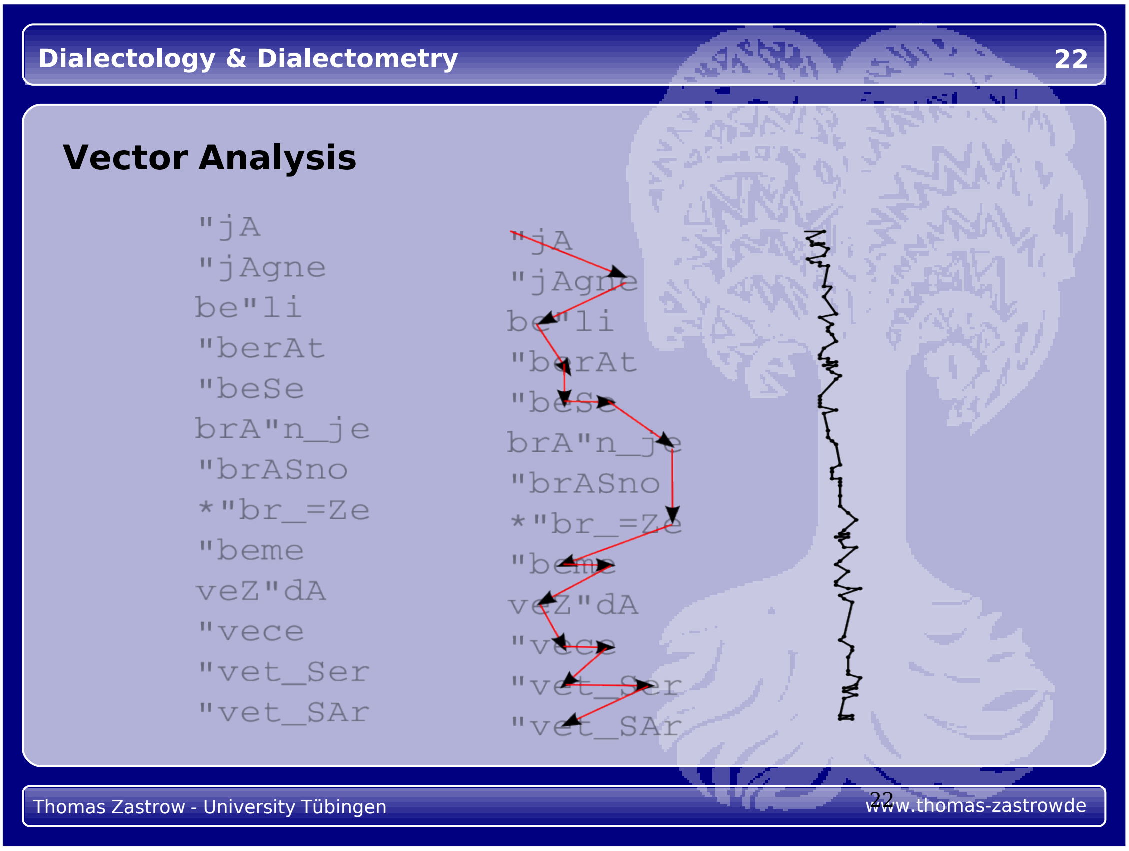
Fig. 2

Thomas Zastrow - University Tübingen

www.thomas-zastrowde

So here is an artifical example. Think of this here beeing a site in Bulgaria and these lines here are representing the words we have for this site.

If we now want to follow the element A through the data set, we start in the upper left corner of the data set, drawing the first vector to the first appearance of A and so on. In the end we will have a vector cahin.



Dialectology & Dialectometry 22

Vector Analysis

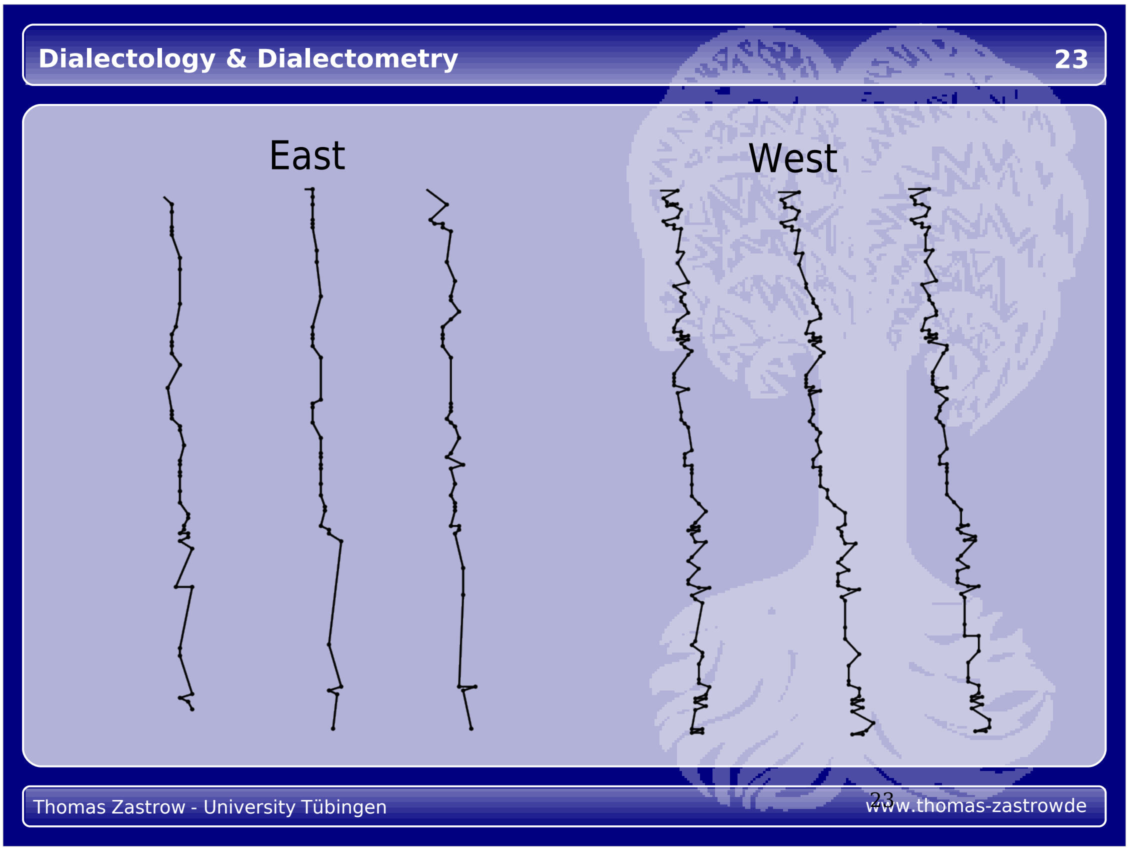
"jA
"jAgne
be"li
"berAt
"beSe
brA"n_je
"brASno
*"br_=Ze
"beme
veZ"dA
"vece
"vet_Ser
"vet_SAr

Thomas Zastrow - University Tübingen

www.thomas-zastrowde

This is now a real example, showing the first words of the site “aldomirivcic” which is in the west of Bukgaria. And you see here, that we are following the vowel “e” here with the vectors through the data set.

On the right side, you see the complete chain of vowel e for the site aldomirovci.

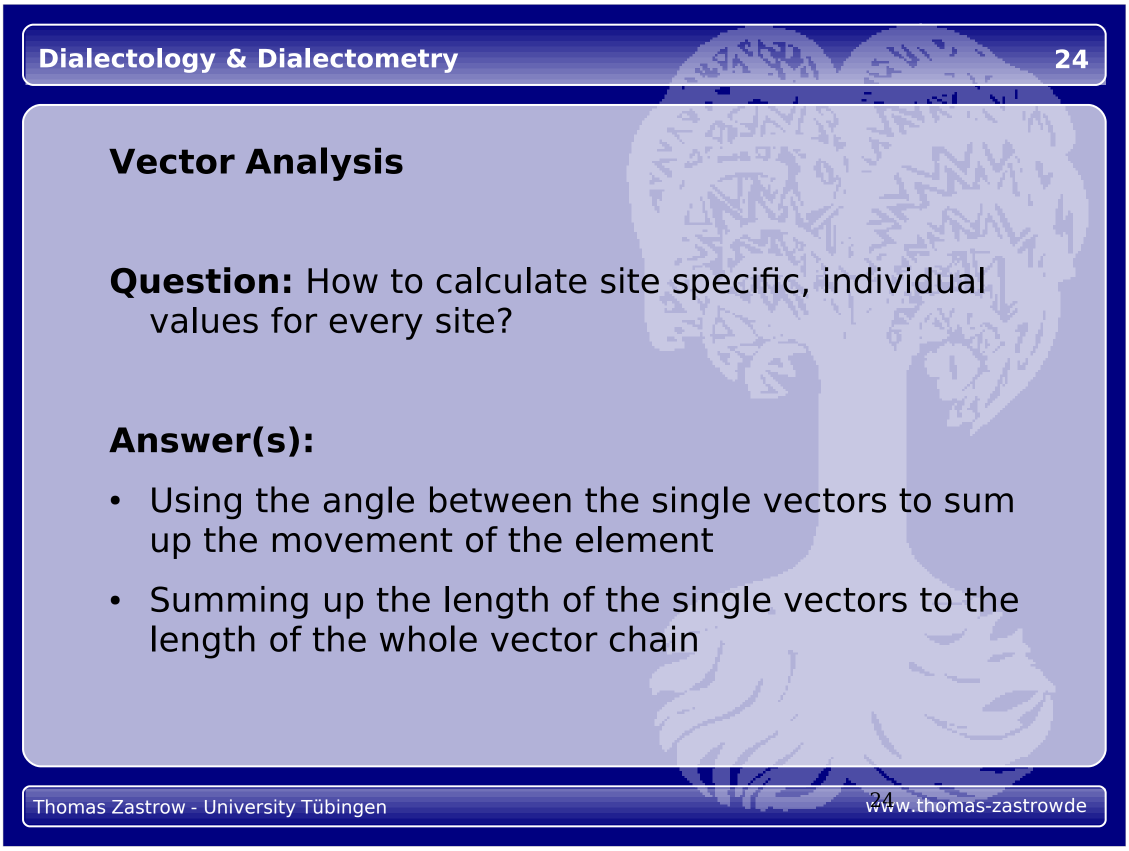


Dialectology & Dialectometry 23

Thomas Zastrow - University Tübingen 23 www.thomas-zastrowde

Here you can see now VCs for six sites - every VC is representing the element "e" though a site.
The three sites on the left here are in the east and the three on the right side are in the west of Bulgaria.

And just from a visual view, you can see that the VCs from the East and these from the West are looking similiar.



Dialectology & Dialectometry 24

Vector 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

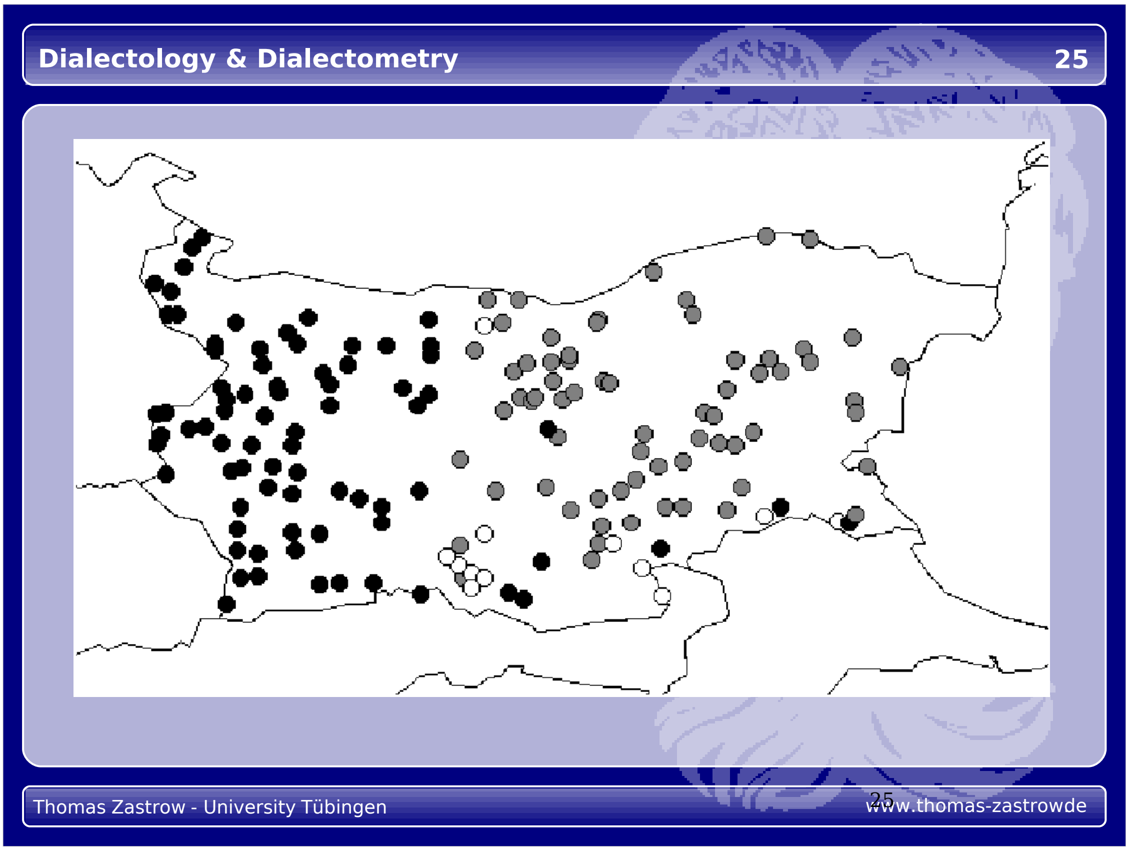
Thomas Zastrow - University Tübingen

www.thomas-zastrowde

But this visual view is certainly niot enough for us. We must ask the question, how we can calculate dialect values in a mathematical way from these VCs.

The asnwer is that we have two possibilities: at first, we can use the angle between the vectors and second one is to sum up the length of the VCs.

What I done so far is to sum up the lengths of VCs.



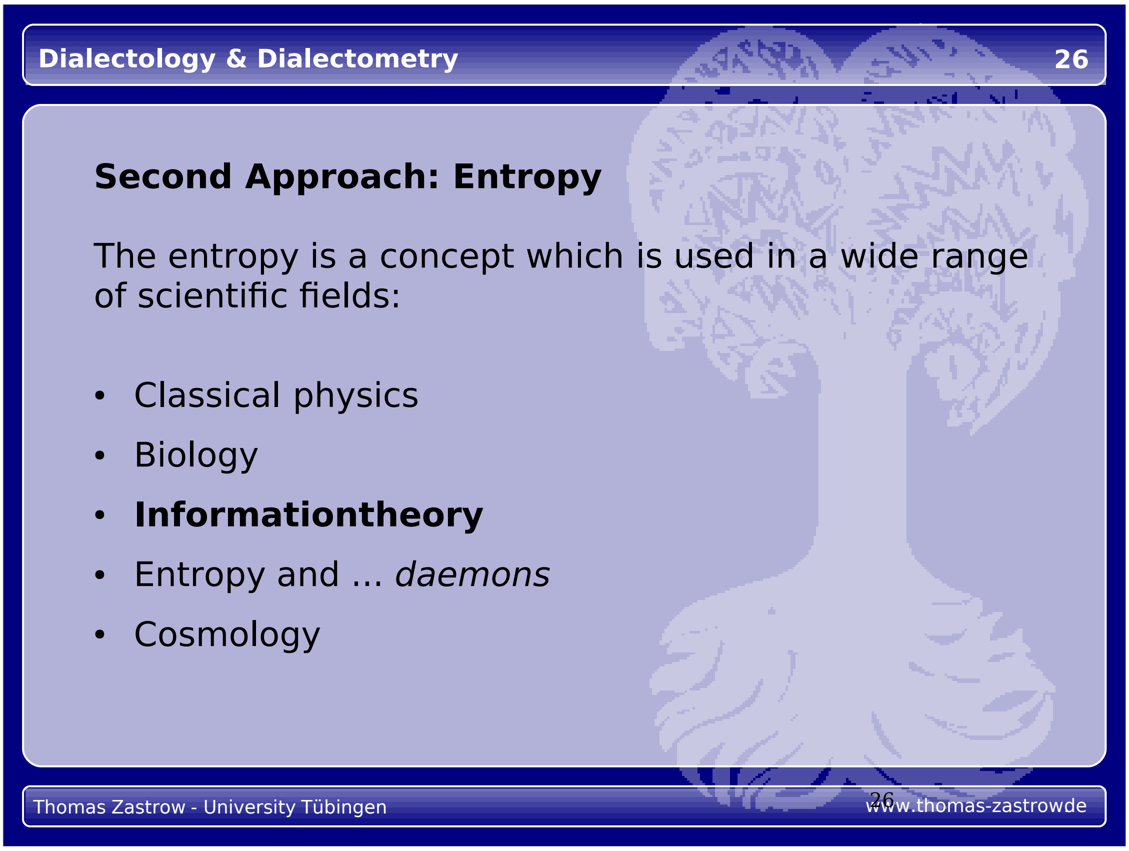
Dialectology & Dialectometry 25

Thomas Zastrow - University Tübingen

25

www.thomas-zastrowde

And this is the result: here you can see in principal three different dialect regions: A strong border between the east and the west and than there is another, small regions here in the south of Bulgaria.



Dialectology & Dialectometry 26

Second Approach: Entropy

The entropy is a concept which is used in a wide range of scientific fields:

- Classical physics
- Biology
- **Informationtheory**
- Entropy and … *daemons*
- Cosmology

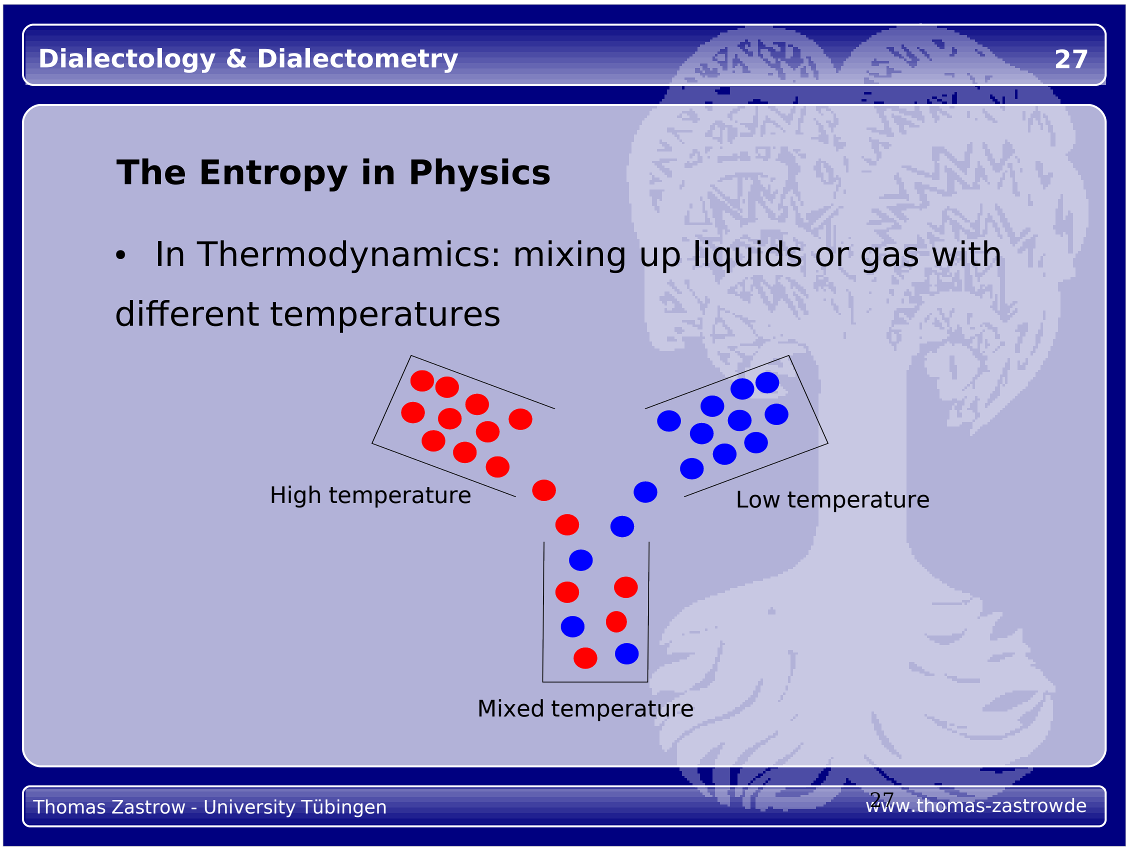
Thomas Zastrow - University Tübingen

www.thomas-zastrowde

Lets talk about our second approach, the Entropy.

The entropy is an information theiretical approach which is used in a wide range of scientific researchs. For example, it is used in Biology as well as in Cosmology and - in a very specific matter - in conjunctin with daemons.

But most important for us is here the information theoretic approach.

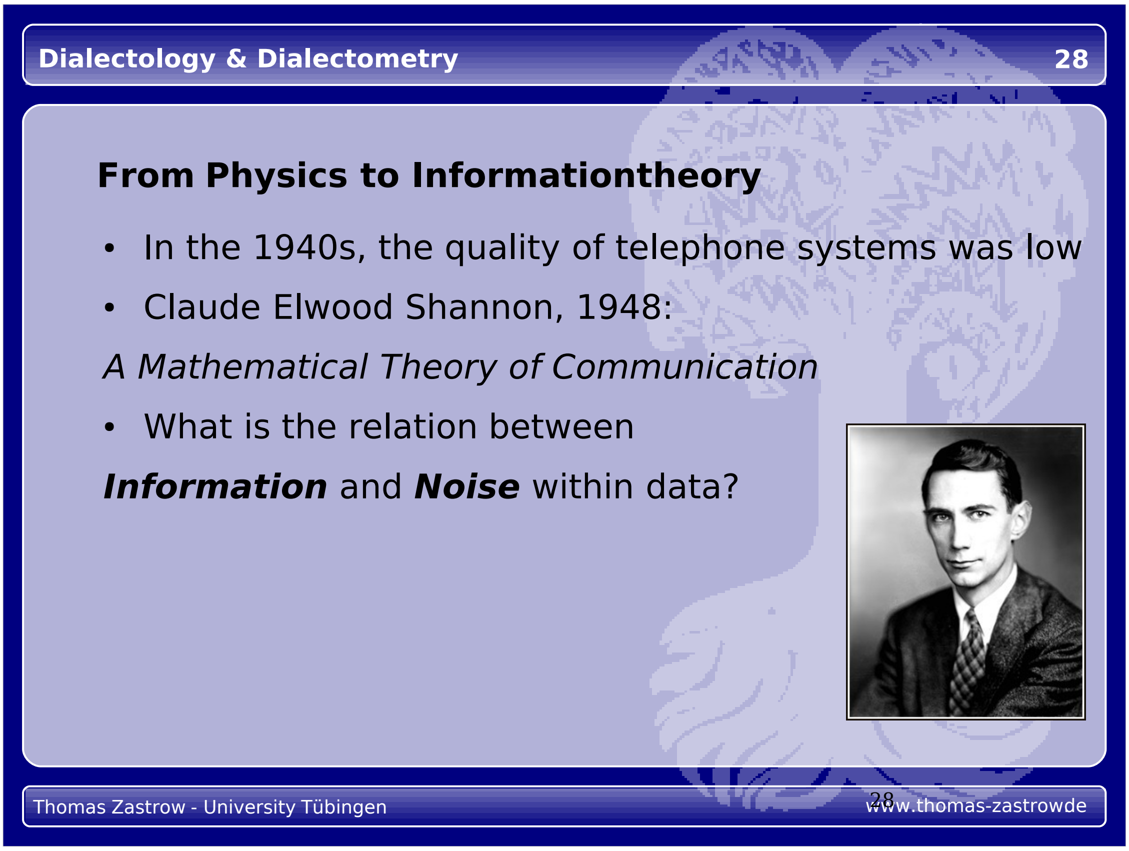


The Entropy in Physics

- In Thermodynamics: mixing up liquids or gas with different temperatures

Just a very short excourse: The entropy has it's origin in Thermodynamics. Here it is used to calculate what happens when you mix up a cold and a warm liquid or gas.

Keep in mind: In CD, we are examining how different dialects are mixed up :-)



From Physics to Informationtheory

- In the 1940s, the quality of telephone systems was low
- Claude Elwood Shannon, 1948:

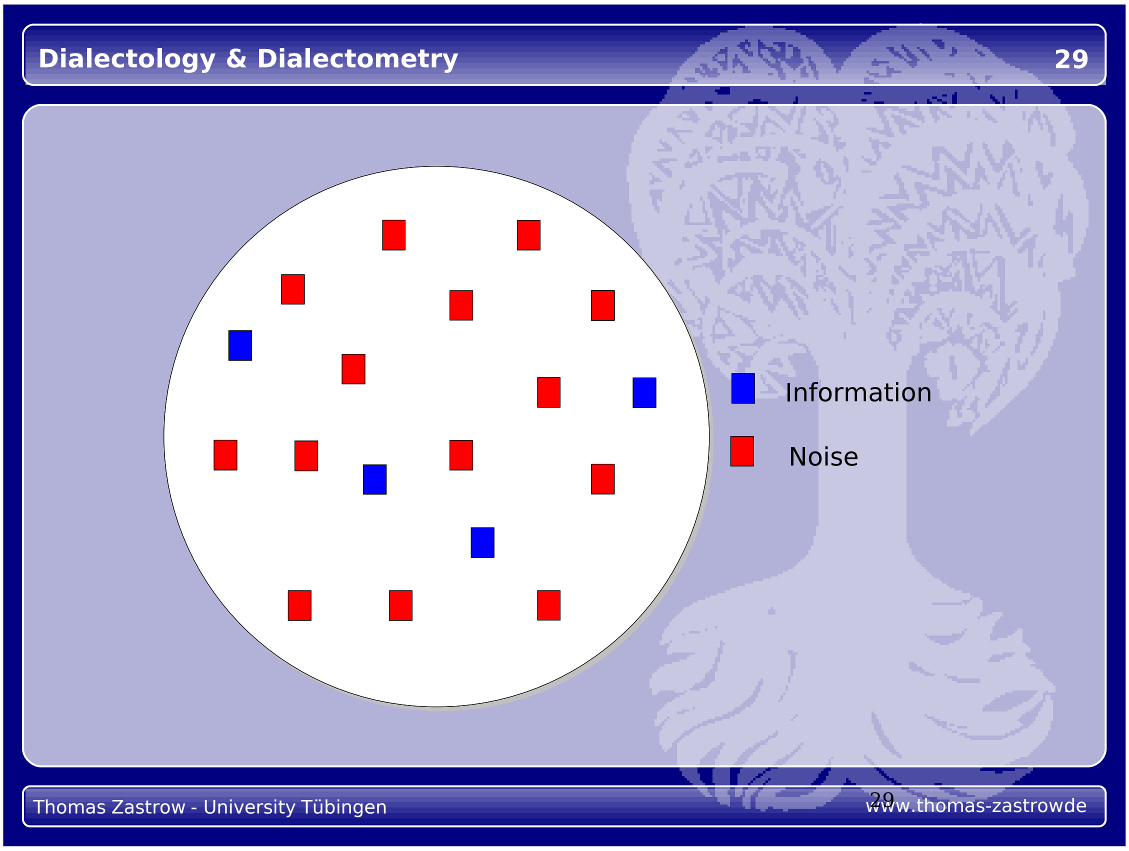
A Mathematical Theory of Communication

- What is the relation between

Information and ***Noise*** within data?

In the 1940s, the american telephone company AT&T asked a researcher whose name was CES, to develope a mahematical method to calculate the quality of telephone lines.

And he came up with the adoption of the phisical principal of entropy to information: What is the difference between Information and noise in a given data set?



Dialectology & Dialectometry

29

🟦 Information

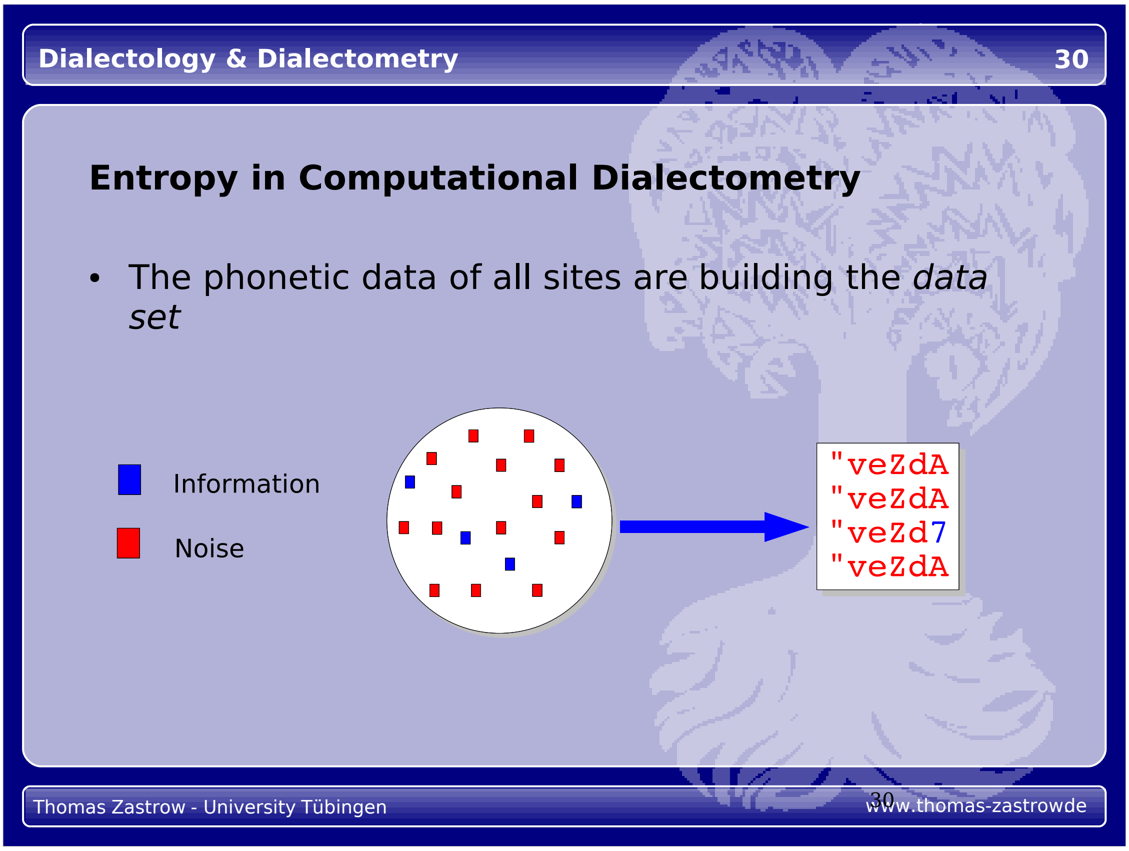
🟥 Noise

Thomas Zastrow - University Tübingen

www.thomas-zastrowde

In a telephone chat, the telephone line communicates the spoken words - information - and noise. The higher the noise, the lower the quality of the telephone line.

In CD, we apply now this function of mixing up noise and information to dialects: The information is the difference between the dialects, that is what we try to measure, and the noise is the unchanged data.



Entropy in Computational Dialectometry

- The phonetic data of all sites are building the *data set*

🟦 Information

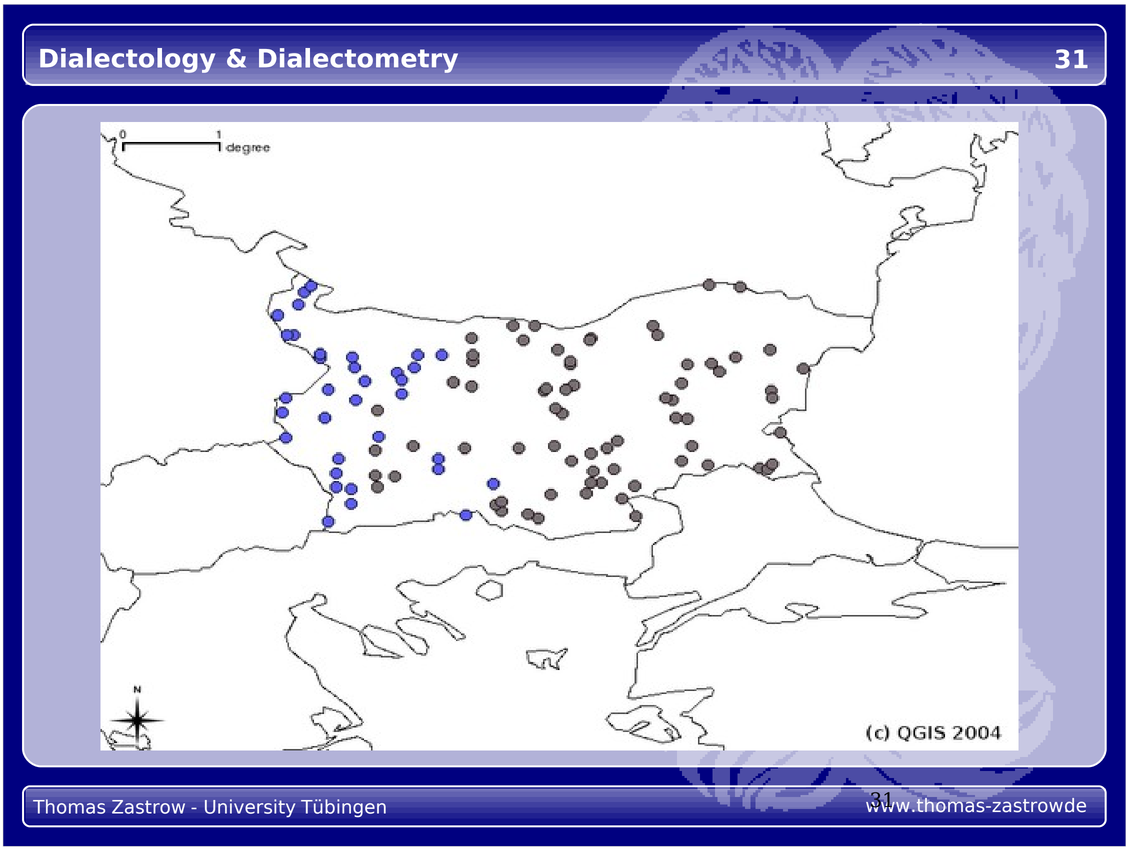
🟥 Noise

"veZdA
"veZdA
"veZd7
"veZdA

So, here we have another example of that. On the right, you see a very short section of one of our wordlists.

You see that all the elements here are the same - instead of one, the XSampa code “7” here. So, from an information theoretical view, the 7 is the only information here and the rest is noise.

That means, that if the difference in information between two sites is high, also the dialect difference is high.



Dialectology & Dialectometry 31

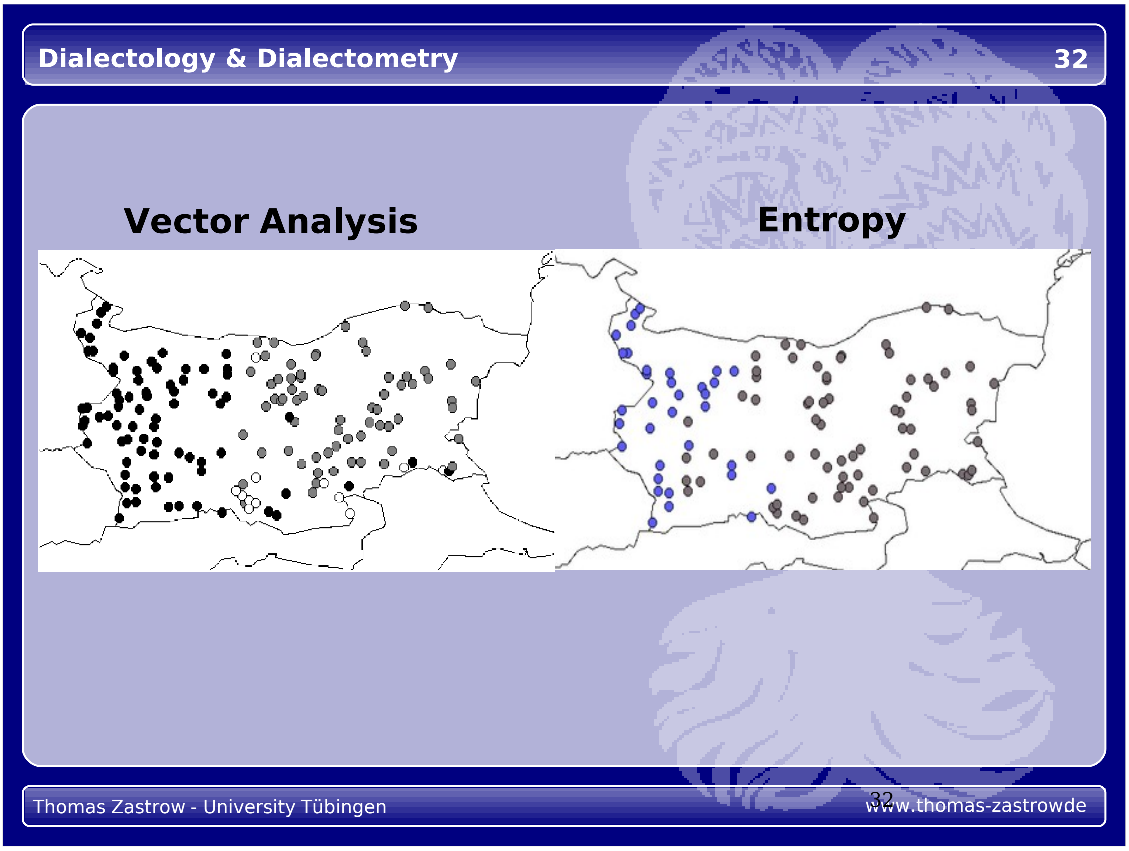
Thomas Zastrow - University Tübingen

www.thomas-zastrowde

Here you can see now a map of the information distribution of our sites.

In fact, we are using not only “The Entropy” as a dialect measure here, but the entropy itself has a lot of variations. What you see here is the so called “Partial Information” of the sites.

Where PI is an entropy related concept.



Dialectology & Dialectometry 32

Vector Analysis

Entropy

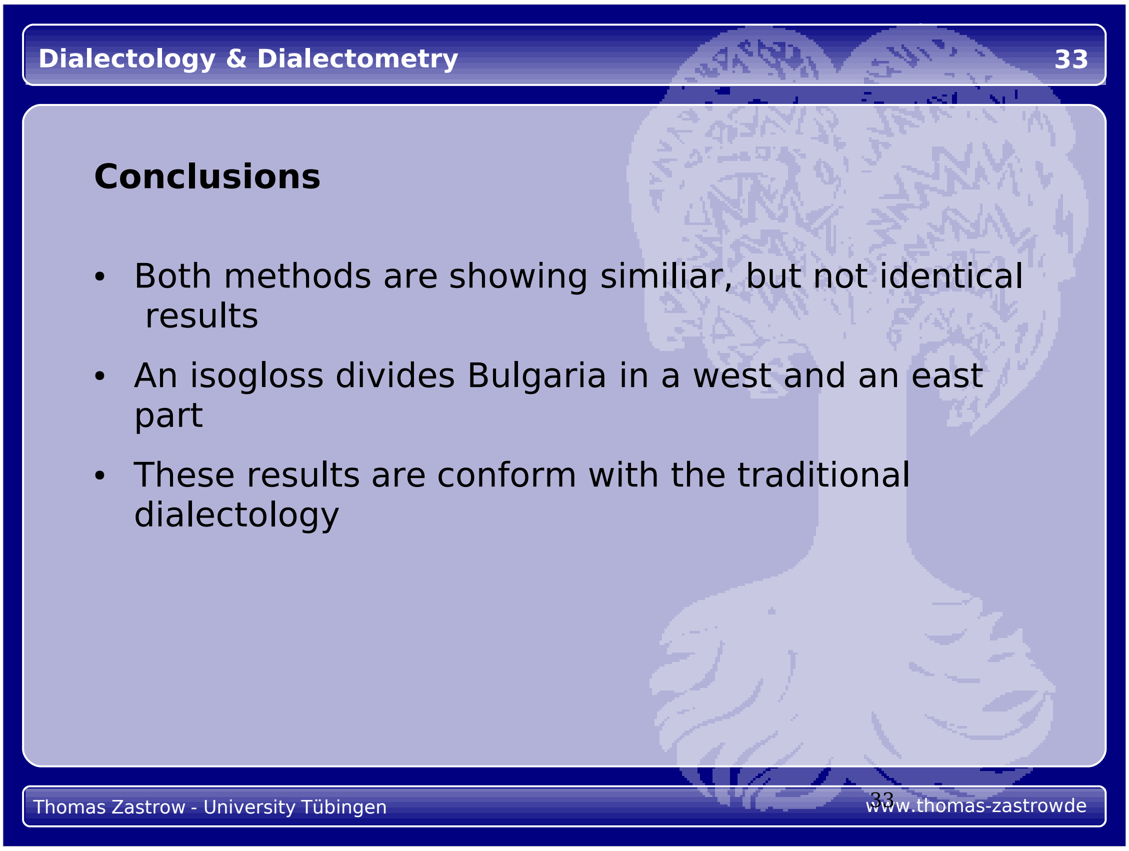
Thomas Zastrow - University Tübingen

32

www.thomas-zastrowde

And here we have the two maps together. On the left, the result from the VA and on the right the results from the Entropy examination.

You can see, both maps are showing similiar results.



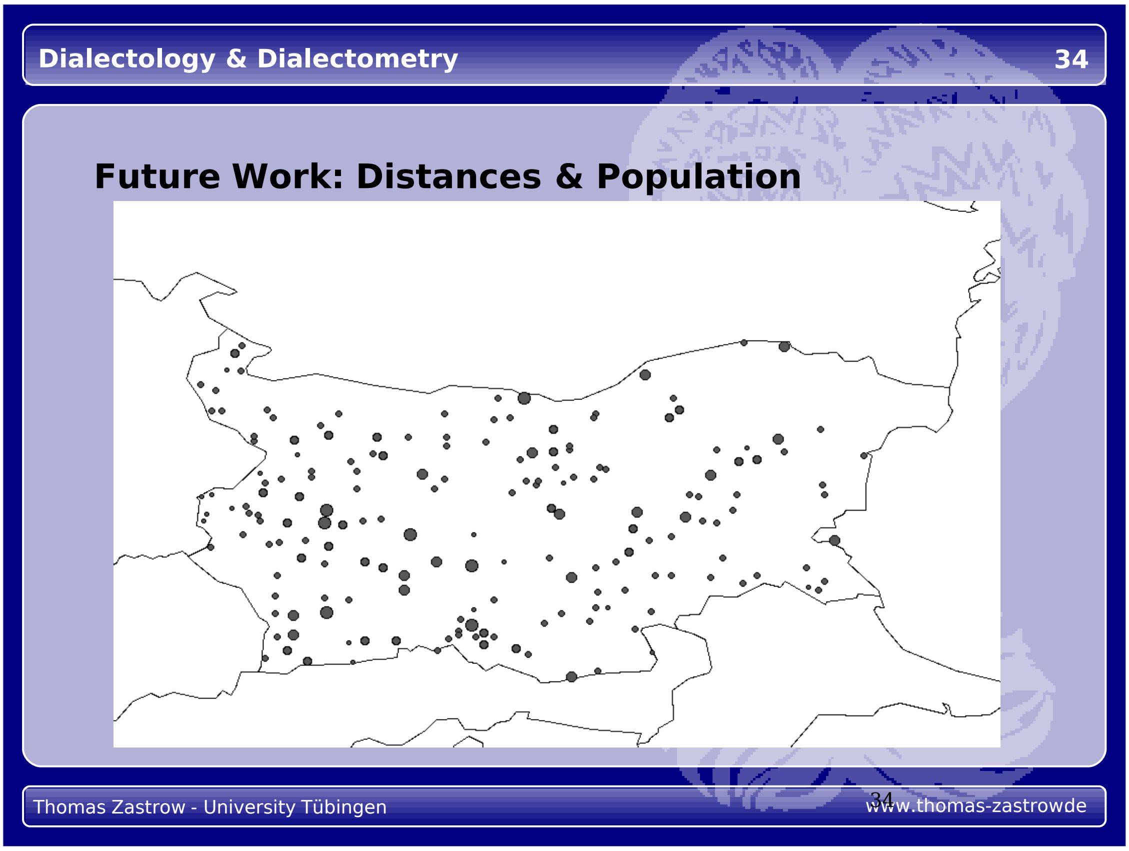
Conclusions

- Both methods are showing similiar, but not identical results
- An isogloss divides Bulgaria in a west and an east part
- These results are conform with the traditional dialectology

So, the conclusion is: Both methods, the VA and the entropy related approaches are producing the similiar results.

These results are not identical - but it is clear that we have a strong isogloss between the east and the west of Bulgaria. Inside the east respective the west part, we have a dialect continuum which breaks at this strong isogloss.

Another important thing is, that this result, which was reached just by the use of mtahematical methods, is conform with the results which were stated by the work of “traditional” dialectologist. So, we can say yes, our methods are useful to divide a language into several dialect regions.



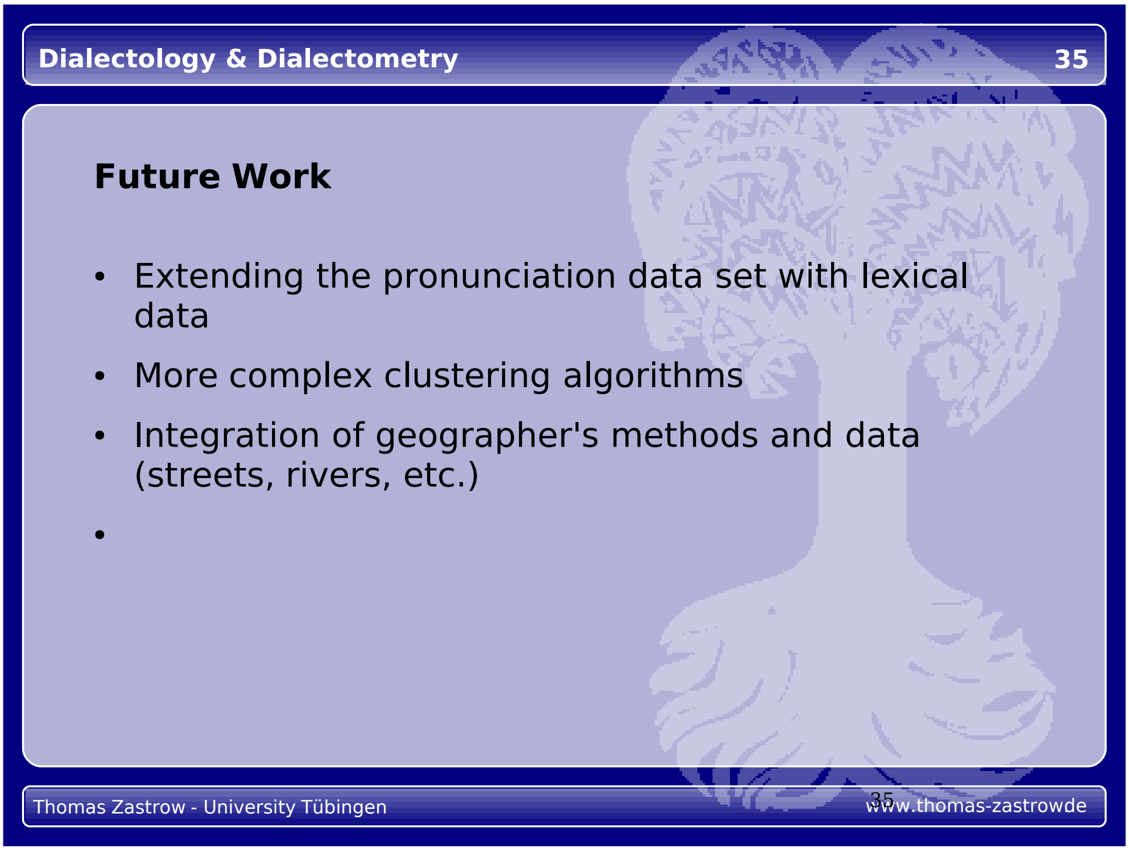
Dialectology & Dialectometry 34

Future Work: Distances & Population

Thomas Zastrow - University Tübingen

www.thomas-zastrowde

As I already said, the next step is to integrate extralinguisti data into our research. I have already collected population data for all our sites and with some algorithms from the geographers, we can bring population, dialect values and distances between the sites into one big relation. Hopefully ;-)



Dialectology & Dialectometry 35

Future Work

- Extending the pronunciation data set with lexical data
- More complex clustering algorithms
- Integration of geographer's methods and data (streets, rivers, etc.)
-

Thomas Zastrow - University Tübingen

www.thomas-zastrowde

Another topic for the future is that we will also get lexical data for Bulgarian. I'm not sure at the moment how we can combine these data with the pronunciation data, if anyone has a good idea, he or she is alwas welcome :-)

Then, I will try some more complex clustering algorithms and last not least, integrate some more of the geographer's data like streets, rivers, mountains and so into the data. Perhaps we can see that dialect variation follows the big traffic routes in a land ....