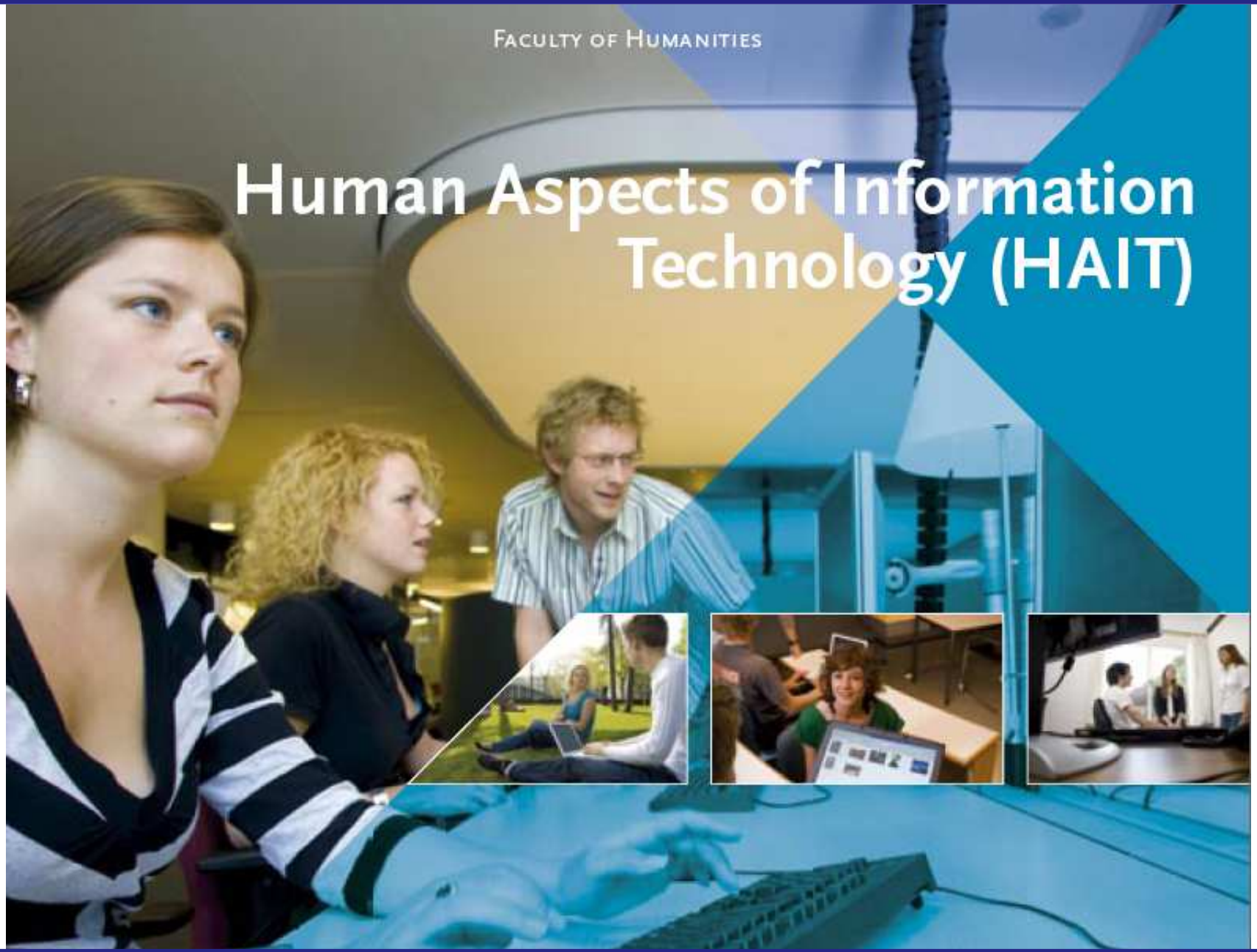

Faculty of Humanities
Human Aspects of Information Technology (HAIT)

Human Aspects of Information Technology (HAIT) 1/24

- Website description
  - Specialist in language and speech technology
  - Developer of IT-services
  - Academic researcher...

Human Aspects of Information Technology (HAIT) 1/24

- Website description
  - Specialist in language and speech technology
  - Developer of IT-services
  - Academic researcher...
- In practice (imagine a birtday party)

them ...and what do you do...?

Human Aspects of Information Technology (HAIT) 1/24

- Website description
  - Specialist in language and speech technology
  - Developer of IT-services
  - Academic researcher...
- In practice (imagine a birtday party)

them ...and what do you do...?

you I'm a language and speech technology specialist...

Human Aspects of Information Technology (HAIT) 1/24

- Website description
  - Specialist in language and speech technology
  - Developer of IT-services
  - Academic researcher...
- In practice (imagine a birtday party)

them ...and what do you do...?
you I'm a language and speech technology specialist...
them A *what?*...

Human Aspects of Information Technology (HAIT) 1/24

- Website description
  - Specialist in language and speech technology
  - Developer of IT-services
  - Academic researcher...
- In practice (imagine a birtday party)

them ...and what do you do...?
you I'm a language and speech technology specialist...
them A *what?*...
you A language and speech technology specialist
you I try to make the computer understand language...

Human Aspects of Information Technology (HAIT) 1/24

- Website description
  - Specialist in language and speech technology
  - Developer of IT-services
  - Academic researcher...
- In practice (imagine a birtday party)

them ...and what do you do...?
you I'm a language and speech technology specialist...
them A *what?*...
you A language and speech technology specialist
you I try to make the computer understand language...
them Ah, computer languages...

Human Aspects of Information Technology (HAIT) 1/24

- Website description
  - Specialist in language and speech technology
  - Developer of IT-services
  - Academic researcher...
- In practice (imagine a birtday party)

them ...and what do you do...?
you I'm a language and speech technology specialist...
them A *what?*...
you A language and speech technology specialist
you I try to make the computer understand language...
them Ah, computer languages...
you No, natural language, you know, what we speak...

Human Aspects of Information Technology (HAIT) 1/24

- Website description
  - Specialist in language and speech technology
  - Developer of IT-services
  - Academic researcher...
- In practice (imagine a birtday party)

them ...and what do you do...?
you I'm a language and speech technology specialist...
them A *what?*...
you A language and speech technology specialist
you I try to make the computer understand language...
them Ah, computer languages...
you No, natural language, you know, what we speak...
them That is so *cool*!

Contents 2/24

To give you an idea what you will find in Tilburg, we will show you three examples of research and teaching in HAIT

- Searching the Internet: e.g. Google
- Language and meaning: Chronological expressions
- Language and numbers: Zipf

Example 1: searching the web 3/24

The first example shows the basics of (internet) search engines.

Example 1: searching the Web 4/24

To search the web, you need a search engine. But what is a search engine?

Answer:

- A search engine is a system that, given a search query (generally keywords), provides:
  - a set or list of documents
  - that are relevant to the query.
- How do they work...

Example 1: searching the Web 4/24

To search the web, you need a search engine. But what is a search engine?

Answer:

- A search engine is a system that, given a search query (generally keywords), provides:
  - a set or list of documents
  - that are relevant to the query.
- How do they work...
  - Build an index (preferably weighted)
  - Compare documents and queries

Example 1: searching the Web 5/25

Building - and weighting - the index:

- Words are not created equal
  - Consider the words 'the', 'dog' and 'IBM'
  - What are the differences...?

Example 1: searching the Web 5/25

Building - and weighting - the index:

- Words are not created equal
  - Consider the words 'the', 'dog' and 'IBM'
  - What are the differences...?
- Frequency-based criteria, for example...:
- tf.idf...
- ...where the weight increases with the frequency in the document but decreases with the document frequency (think about that for a moment)...

Example 1: searching the Web 5/25

Building - and weighting - the index:

- Words are not created equal
  - Consider the words 'the', 'dog' and 'IBM'
  - What are the differences...?
- Frequency-based criteria, for example...:
- tf.idf...
- ...where the weight increases with the frequency in the document but decreases with the document frequency (think about that for a moment)...
- 'the'and 'ibm' have both a frequency of 10 in document A.
- But 'the' occurs in *all* of the 100 documents and 'ibm' in only two.
- The tf.idf of 'and' is 10/100; that of 'ibm' is 10/2.

Example 1: searching the Web 6/25

Comparing the documents (and queries)

- Pages are not created equal either
- However, many pages may match a query...

Example 1: searching the Web 6/25

Comparing the documents (and queries)

- Pages are not created equal either
- However, many pages may match a query...
- Which pages are best? How to rank them?
  - As we have seen, keywords may have different weights
  - Text in titles/headings is important
  - Text earlier in the page is important
  - Text of links *to this page* is important
  - Important pages link to *other* important pages

Example 1: searching the Web 7/24

A good model to compare documents is the Vector Space Model.

- The next slide shows a very simple vector space with only three keywords.
  We also have drawn in three documents (the red lines)...

Example 1: searching the Web 7/24

A good model to compare documents is the Vector Space Model.

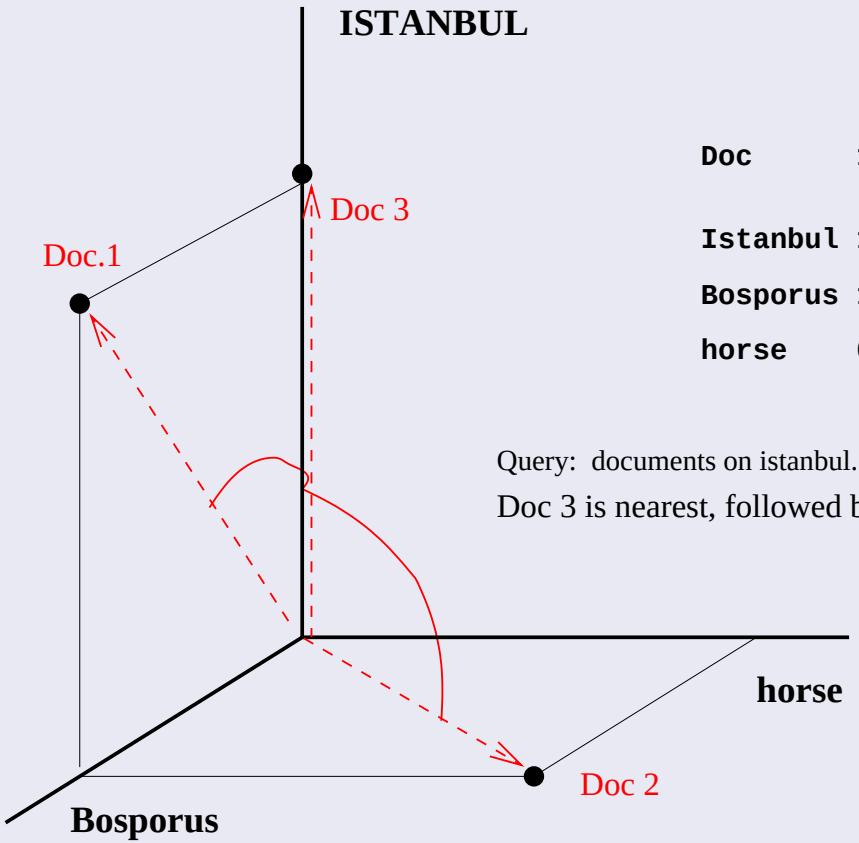
- The next slide shows a very simple vector space with only three keywords.
  We also have drawn in three documents (the red lines)...
- ...the words have binary weights
  (equal distances from the origin)...

Example 1: searching the Web 7/24

A good model to compare documents is the Vector Space Model.

- The next slide shows a very simple vector space with only three keywords.
  We also have drawn in three documents (the red lines)...
- ...the words have binary weights
  (equal distances from the origin)...
- ...and there are no semantic differences
  (the axes stand under an angle of 90 degrees).

Example 1: searching the Web 8/24



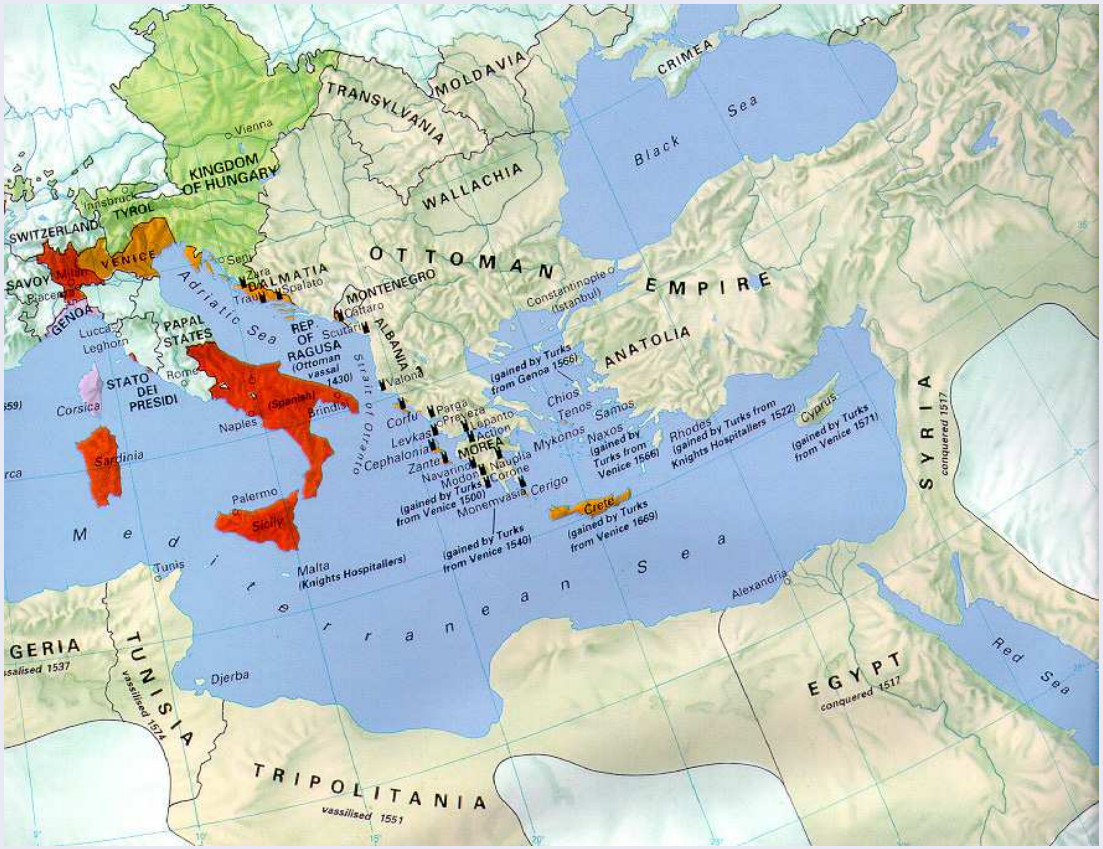
Doc	1	2	3
Istanbul	1	0	1
Bosporus	1	1	0
horse	0	1	0

Query: documents on istanbul.

Doc 3 is nearest, followed by Doc 1

Example 2: Chronological Expressions 9/24

The next example shows an application of Memory Based learning.



Example 2: Chronological Expressions 10/24

Let us return to Google: imagine you want to find pages about Turkey in the time of the Ottoman empire.

- You type "Turkey ottoman"...
- and you get all pages with these words. BUT...

Example 2: Chronological Expressions 10/24

Let us return to Google: imagine you want to find pages about Turkey in the time of the Ottoman empire.

- You type "Turkey ottoman"...
- and you get all pages with these words. BUT...
- ...if a page is about Turkey between 1500 and 1950 and the word 'ottoman' does not occur on that page...

Example 2: Chronological Expressions 10/24

Let us return to Google: imagine you want to find pages about Turkey in the time of the Ottoman empire.

- You type "Turkey ottoman"...
- and you get all pages with these words. BUT...
- ...if a page is about Turkey between 1500 and 1950 and the word 'ottoman' does not occur on that page...
- ... you will not find that page.

Example 2: Chronological Expressions 11/24

The Ottoman empire existed for 550 years. If you want to be certain to find all pages that combine 'turkey' with any or all of those years, you have to type all queries:

- "turkey 1500', 'turkey 1501', 'turkey 1502' etc..."

Example 2: Chronological Expressions 11/24

The Ottoman empire existed for 550 years. If you want to be certain to find all pages that combine 'turkey' with any or all of those years, you have to type all queries:

- "turkey 1500', 'turkey 1501', 'turkey 1502' etc..."
- ... which is absurd...

Example 2: Chronological Expressions 11/24

The Ottoman empire existed for 550 years. If you want to be certain to find all pages that combine 'turkey' with any or all of those years, you have to type all queries:

- "turkey 1500', 'turkey 1501', 'turkey 1502' etc..."
- ... which is absurd...
- ... and you will nevertheless miss 'seventeenth century, 17th century, XVIIth century and all such terms...

Example 2: Chronological Expressions 11/24

The Ottoman empire existed for 550 years. If you want to be certain to find all pages that combine 'turkey' with any or all of those years, you have to type all queries:

- "turkey 1500', 'turkey 1501', 'turkey 1502' etc..."
- ... which is absurd...
- ... and you will nevertheless miss 'seventeenth century, 17th century, XVIIth century and all such terms...
- ... and you will find a *lot* of telephone numbers, prices and inventory numbers!

Example 2: Chronological Expressions 12/24

The solution here may be found in Memory based learning (MBL).
In Memory Based Learning, you first

- collect a large database with descriptions of examples,
- and assign to each example the correct class...

Example 2: Chronological Expressions 12/24

The solution here may be found in Memory based learning (MBL).
In Memory Based Learning, you first

- collect a large database with descriptions of examples,
- and assign to each example the correct class...
- New cases are compared with the examples in the database
- and given the class of the examples that are most similar...

Example 2: Chronological Expressions 12/24

The solution here may be found in Memory based learning (MBL).
In Memory Based Learning, you first

- collect a large database with descriptions of examples,
- and assign to each example the correct class...
- New cases are compared with the examples in the database
- and given the class of the examples that are most similar...
- Of course this is not so easy as it looks...

Example 2: Chronological Expressions 12/24

The solution here may be found in Memory based learning (MBL).
In Memory Based Learning, you first

- collect a large database with descriptions of examples,
- and assign to each example the correct class...
- New cases are compared with the examples in the database
- and given the class of the examples that are most similar...
- Of course this is not so easy as it looks...
  - Which features are the important ones?
  - How do you weigh them?
  - How do you measure the differences between the examples...?

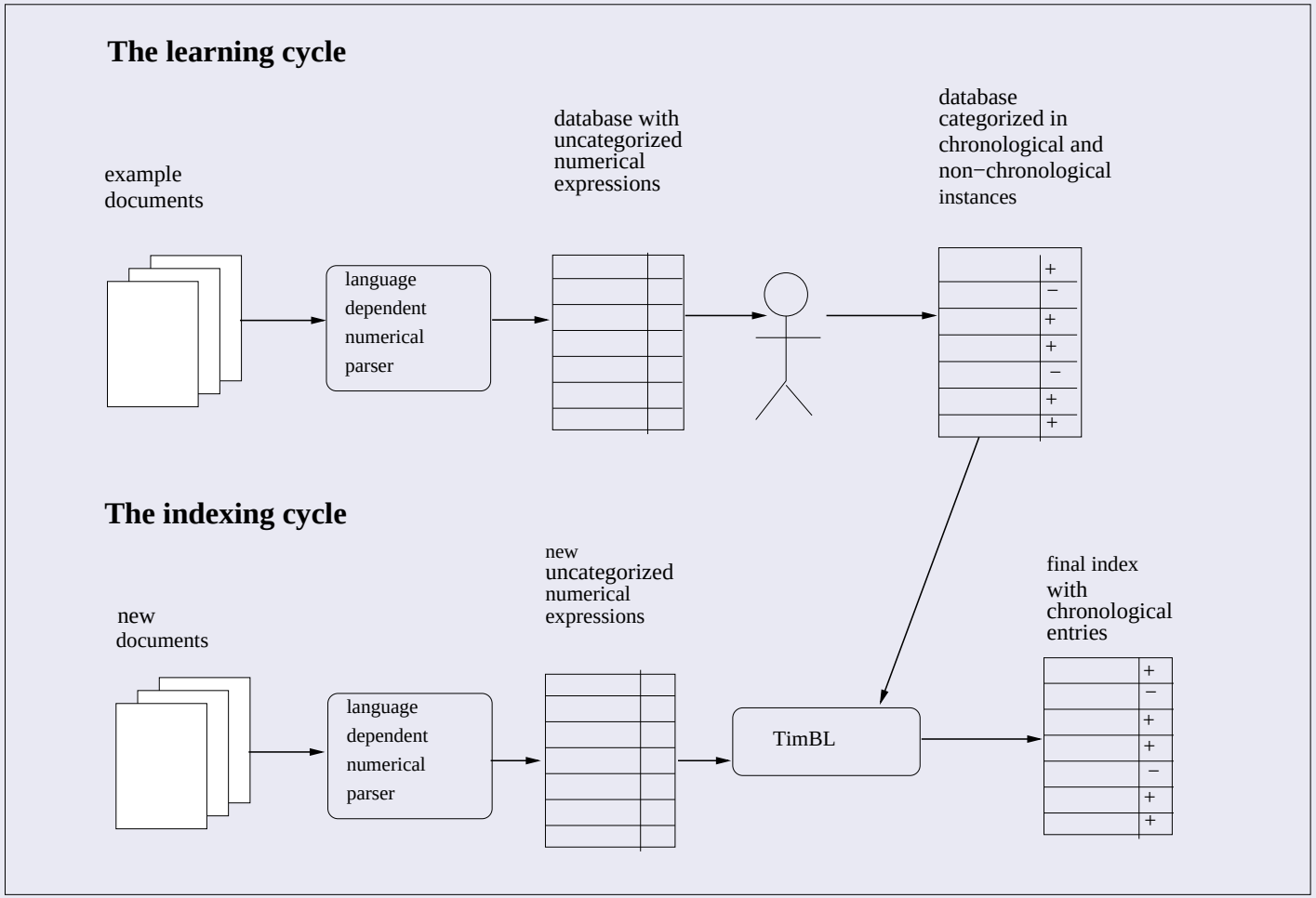
Example 2: Chronological Expressions 12/24

The solution here may be found in Memory based learning (MBL). In Memory Based Learning, you first

- collect a large database with descriptions of examples,
- and assign to each example the correct class...
- New cases are compared with the examples in the database
- and given the class of the examples that are most similar...
- Of course this is not so easy as it looks...
  - Which features are the important ones?
  - How do you weigh them?
  - How do you measure the differences between the examples...?
- Actually, it does not differ so much from the first example about search engines.

Example 2: Chronological Expressions 13/24

Machine Based Learning: scheme



Example 2: Chronological Expressions 14/24

Flat	roof	tile	1	0.193	Post	-	roman	34	[other]
Depth	:	l	;	0.2	m	18.8	m	OD	[other]
is	first	seen	in	675	AD	when	it	was	[Timespan]
Pitcher	handle	22	0.443	10th	to	13th	cent	30	[Timespan]
75	human	bone	-	0.095	76	Pottery	1	0.045	[other]
deposit	was	up	to	0.1	m	thick	and	sloped	[other]
occurs	in	charters	from	1128	and	1153	.	The	[Timespan]

The class of the focusword (red) is defined by its context.

Example 2: Chronological expressions 15/24

In this way we can index every expression in a document that indicates a year or an era.
Similar techniques are used to disambiguate geographical names or to extract names of persons and institutions.
Outside IR there are even more important linguistic applications for Memory Based Learning:

- Part of speech tagging
- morphological segmentation
- machine translation
- and so on.

Example 3: Zipf... 16/14

The last example shows just wat you can do when you count words.

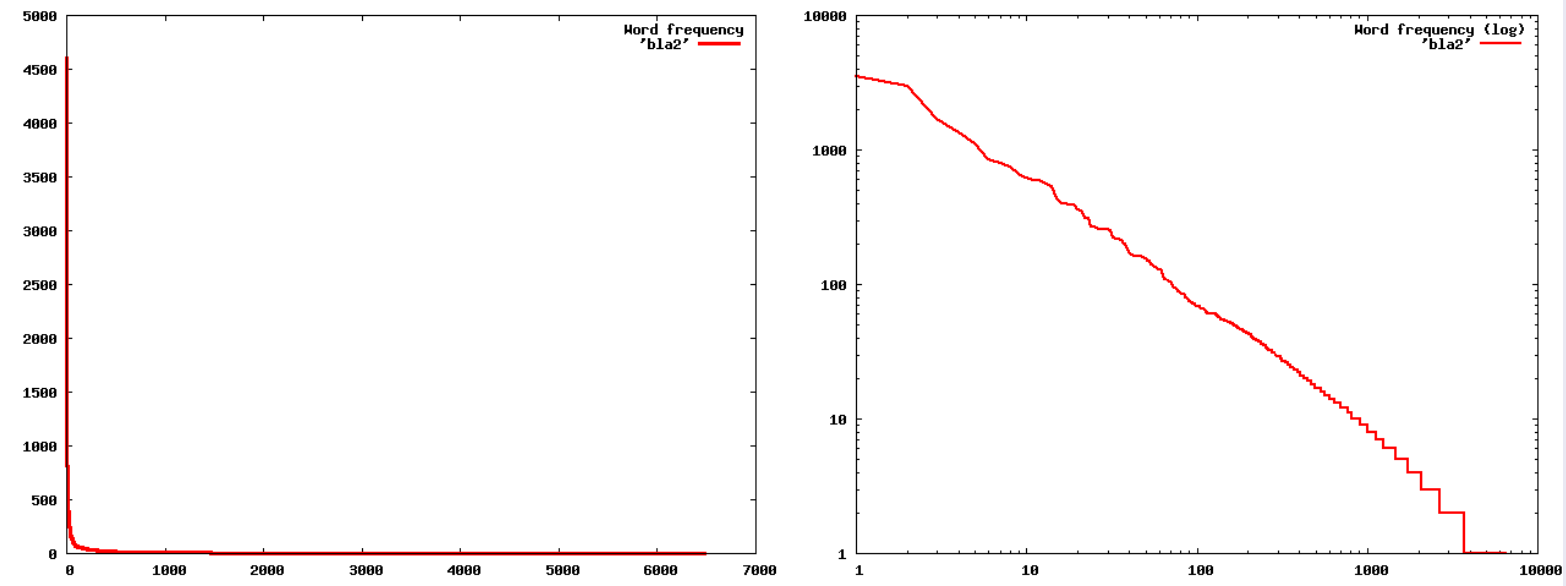
Example 3: Zipf... 17/14

Let us look at a typical frequency list of words in a text.

rank	freq.	word sorted	freq. sorted	word
1	99	A	2387	the
2	1	ACK	1381	of
3	3	AD	814	in
4	34	ALG	787	and
5	4	ANSI	776	a
6	2	ARFF	770	to
7	1	ARPAnet	622	is
8	1	AS	618	small
9	29	ASCII	467	that
10	1	AWK	443	The
11	1	Abort	437	tt
...	...	...	...	...
4547	18	you	1	Abuse
4548	11	your	1	Absicht
4549	14	z	1	About
4550	2	zaag	1	Aborts
4551	1	zealand	1	Abort

The next slide shows graphs of frequency against rank.

Example 3: Zipf... 18/24



An important insight was discovered before the war, when Zipf discovered the relation between rank and frequency of tokens in a language and formulated his 'law of least effort', not unlike Pareto's distribution.

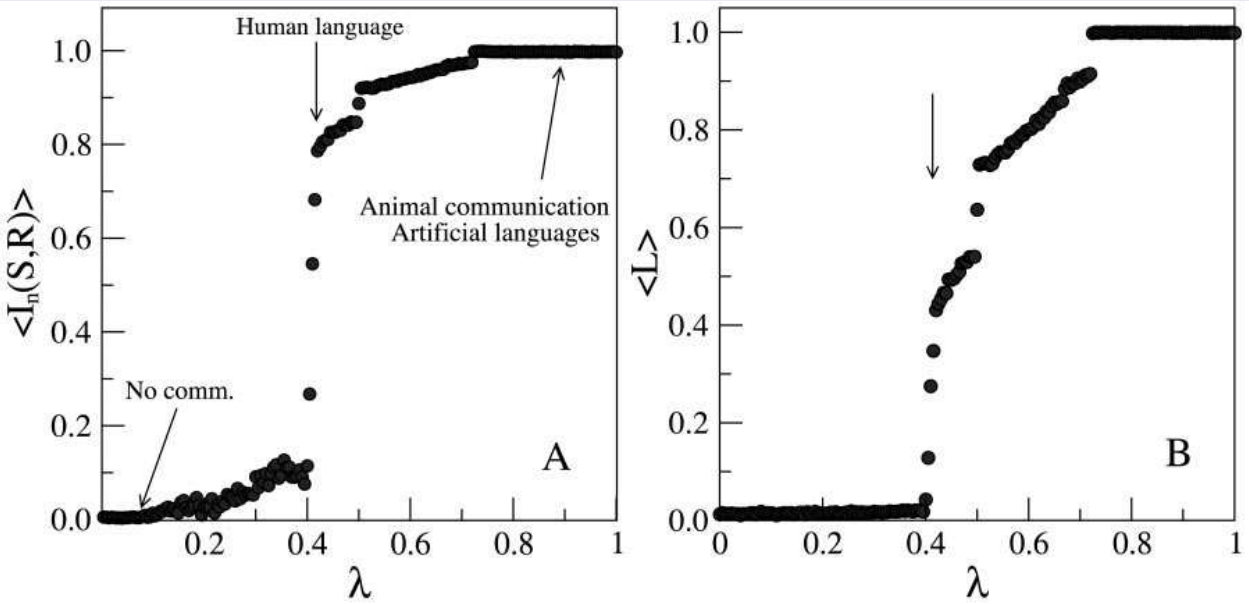
Example 3: Zipf... 19/24

Let us have a closer look at the possible relations between words and meaning in terms of speaker/listener effort:

	word1	word2	word3	word1	word2	word3
meaning1	1	0	0	1	0	0
meaning2	1	0	0	0	1	0
meaning3	1	0	0	0	0	1
	No effort for speaker			No effort for listener		

In 2002 two spanish scientists started from this matrix and computed the invested effort for speaker and listener for all word-meaning combinations.
The next slide shows some very interesting graphs:

Example 3: Zipf... 20/24

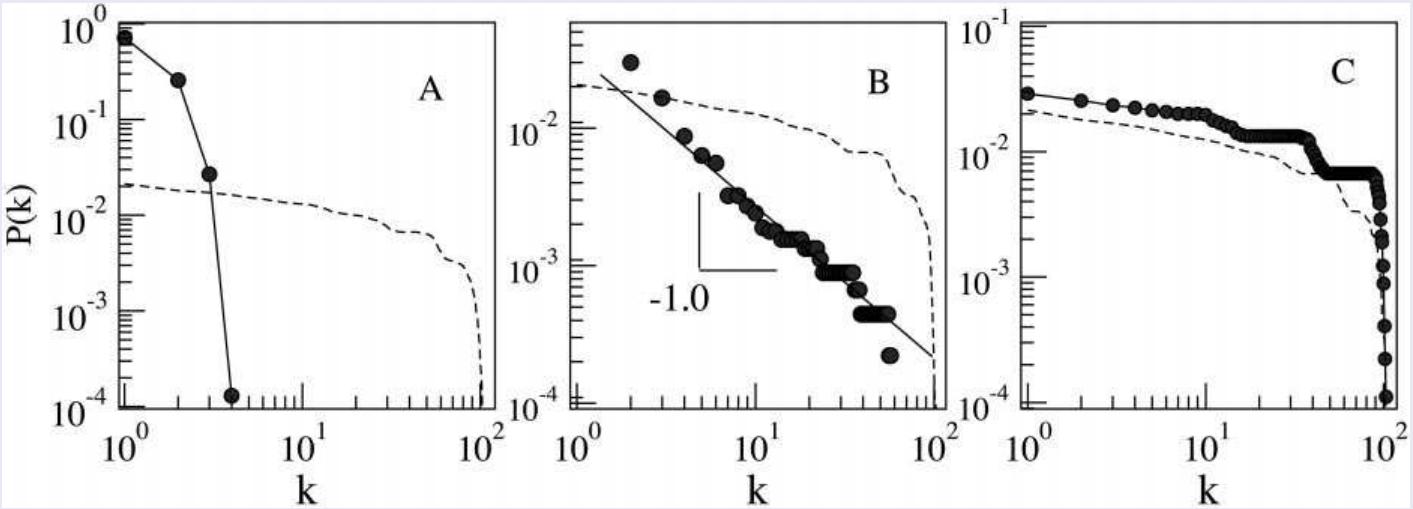


This figure and that of the next slide are from Ferrer/Sole "Least effort and the origins of scaling in human language",2003.

- Horizontally, λ shows the ratio of effort of speaker/listener
- The vertical axis shows the mutual information (A) and the effective lexicon (B)

Example 3: Zipf... 21/24

If we translate the matrices of $\lambda < 41$, $\lambda = 41$ and $\lambda > 41$ into frequency–rank graphs, we see that the matrix of $\lambda = 41$ displays the Zipf curve.



This demonstrates the unique and very narrow position of human language between all other communicative models.

Example 3: Zipf... 22/24

This is all very nice, but what can we do with it?

Example 3: Zipf... 22/24

This is all very nice, but what can we do with it?

- To start with: it gives new insights in the development of language.
- New and better models of human communication can profit from such insights.
- It helps with understanding animal communication.
  In fact, dolphins and some other animals display Zipfian distributions in their communication systems.
- It can be used for decyphering unknown or ancient languages.

Example 3: Zipf... 23/24

Finally, when my students get very bored, I ask them to decypher the so-called Voynich manuscript.

Example 3: Zipf... 23/24

Finally, when my students get very bored, I ask them to decypher the so-called Voynich manuscript.

- It purportedly is a 16-th century manuscript in an unknown language, with pictures of unknown plants, cities and objects...

Example 3: Zipf... 23/24

Finally, when my students get very bored, I ask them to decypher the so-called Voynich manuscript.

- It purportedly is a 16-th century manuscript in an unknown language, with pictures of unknown plants, cities and objects...
- But it may also be a hoax, albeit a very complicated one...

Example 3: Zipf... 23/24

Finally, when my students get very bored, I ask them to decypher the so-called Voynich manuscript.

- It purportedly is a 16-th century manuscript in an unknown language, with pictures of unknown plants, cities and objects...
- But it may also be a hoax, albeit a very complicated one...
- To this moment, linguists have been unable to decide.

Example 3: Zipf... 24/24

So now you can have a go at it!