

<p>university of groningen</p> <p>Introduction Collocations Unknown Words Similar Words Applications Conclusions</p> <h2 style="text-align: center;">Corpus linguistics</h2> <h3 style="text-align: center;">Word Usage and Word Meaning</h3> <p style="text-align: center;">Gosse Bouma en Erik Tjong Kim Sang</p> <p style="text-align: center;">Information Science University of Groningen</p> <p style="text-align: center;">Week 5</p> <p style="text-align: right;">1/32</p>	<p>university of groningen</p> <p>Introduction Collocations Unknown Words Similar Words Applications Conclusions</p> <h2 style="text-align: center;">Overview</h2> <ol style="list-style-type: none"> 1 Introduction 2 Collocations 3 Unknown Words 4 Similar Words 5 Applications 6 Conclusions <p style="text-align: right;">2/32</p>																								
<p>university of groningen</p> <p>Introduction Collocations Unknown Words Similar Words Applications Conclusions</p> <h2 style="text-align: center;">Collocations</h2> <p>Collocations are fixed expressions of two or more words with a special, fixed, meaning</p> <p>Examples of collocations of words</p> <p>lingua franca , dementia praecox , habeas corpus , instant messaging , gangsta rap , ad hoc , first-person shooter , alcoholic beverages , critically acclaimed , hedge fund , worth noting , mentally ill , black hole</p> <p>Examples of collocations of names</p> <p>Suu Kyi , Phnom Penh , Foo Fighters , Lib Dems , Irian Jaya , Yom Kippur , Dalai Lama</p> <p style="text-align: right;">3/32</p>	<p>university of groningen</p> <p>Introduction Collocations Unknown Words Similar Words Applications Conclusions</p> <h2 style="text-align: center;">Fixed Expressions</h2> <p>Can we extract such fixed expressions from text automatically?</p> <ul style="list-style-type: none"> Fixed expressions consist of words that co-occur frequently However, not all frequent word-combinations are fixed expressions. The most frequent word pairs in a text are not (all) fixed expressions. <p>of the, in the, to the, and the, on the, by the, for the, from the, with the, as a, of a, to be, is a, as the, is the, at the, that the, such as, in a, ...</p> <p style="text-align: right;">4/32</p>																								
<p>university of groningen</p> <p>Introduction Collocations Unknown Words Similar Words Applications Conclusions</p> <h2 style="text-align: center;">Fixed Expressions</h2> <p>Finding fixed expressions automatically</p> <ul style="list-style-type: none"> We want to find arbitrary fixed expressions automatically. Can we do better than just most frequent combinations? The word <i>stainless</i> in a text is very likely to be followed by <i>steel</i>. We are looking for word pairs for which the combination Word1+Word2 occurs (almost) as often as Word1 or Word2 by itself. <p style="text-align: right;">5/32</p>	<p>university of groningen</p> <p>Introduction Collocations Unknown Words Similar Words Applications Conclusions</p> <h2 style="text-align: center;">Word Frequencies</h2> <p>Counts and Frequencies</p> <ul style="list-style-type: none"> Sometimes we want to know whether word <i>W</i> (say, <i>president</i>) occurs more often in text A (1,000 words) than in text B (10,000 words). Count the number of occurrences of <i>W</i> in A and B ? (Relative) Frequency takes text size into account <ul style="list-style-type: none"> If <i>president</i> occurs 20 times in A, the (relative) frequency of <i>president</i> in A is $20/1,000 = 0.02$. If <i>president</i> occurs 40 times in a text B, the (relative) frequency of <i>president</i> in B is $40/10,000 = 0.004$. Relative Frequencies can be seen as probabilities: <ul style="list-style-type: none"> The probability that an arbitrary word from text A is <i>president</i> is 0.02. <p style="text-align: right;">6/32</p>																								
<p>university of groningen</p> <p>Introduction Collocations Unknown Words Similar Words Applications Conclusions</p> <h2 style="text-align: center;">Fixed Expressions</h2> <p>What is the probability of seeing a word pair W_1+W_2 in a text?</p> <p>Observed and Expected Frequency</p> <ul style="list-style-type: none"> Answer 1 (Observed Frequency): the frequency of the pair W_1+W_2 in the text Answer 2 (Expected Frequency) : the frequency of W_1 multiplied with the frequency of W_2. <ul style="list-style-type: none"> As with dice: the probability of throwing $2 \times 6 = 1/6 \times 1/6 = 1/36$ <p>Fixed Expressions vs. other word pairs</p> <ul style="list-style-type: none"> Unrelated pairs: Observed and Expected Frequency are similar Fixed Expressions: Observed Frequency much higher than Expected Frequency <p style="text-align: right;">7/32</p>	<p>university of groningen</p> <p>Introduction Collocations Unknown Words Similar Words Applications Conclusions</p> <h2 style="text-align: center;">Examples</h2> <p>Wikipedia fragment (10M words)</p> <table border="1"> <thead> <tr> <th>W1</th> <th>W2</th> <th>f(W1)</th> <th>f(W2)</th> <th>Observed</th> <th>Expected</th> </tr> </thead> <tbody> <tr> <td>of</td> <td>the</td> <td>0.0462</td> <td>0.0785</td> <td>0.0147</td> <td>0.0036</td> </tr> <tr> <td>United States</td> <td>stainless steel</td> <td>0.0008</td> <td>0.0006</td> <td>0.0005</td> <td>0.00000048</td> </tr> <tr> <td>stainless</td> <td>steel</td> <td>0.000002</td> <td>0.00004</td> <td>0.000002</td> <td>0.000000008</td> </tr> </tbody> </table> <p>of the occurs 4 times as often as expected (0.0147/0.0028)</p> <p>United States occurs 1,000 times as often as expected</p> <p>stainless steel occurs 25,000 times as often as expected</p> <p style="text-align: right;">8/32</p>	W1	W2	f(W1)	f(W2)	Observed	Expected	of	the	0.0462	0.0785	0.0147	0.0036	United States	stainless steel	0.0008	0.0006	0.0005	0.00000048	stainless	steel	0.000002	0.00004	0.000002	0.000000008
W1	W2	f(W1)	f(W2)	Observed	Expected																				
of	the	0.0462	0.0785	0.0147	0.0036																				
United States	stainless steel	0.0008	0.0006	0.0005	0.00000048																				
stainless	steel	0.000002	0.00004	0.000002	0.000000008																				

Introduction Collocations Unknown Words Similar Words Applications Conclusions

Pointwise Mutual Information

In order to avoid very large and very small scores, we will use the pointwise mutual information score:

$$PMI(W1 + W2) = \log\left(\frac{\text{Observed}}{\text{Expected}}\right) = \log\left(\frac{f(W1 + W2)}{f(W1) \times f(W2)}\right)$$

Wikipedia fragment (10M words)

W1	W2	f(W1)	f(W2)	Observed	Expected	PMI
of	the	0.0462	0.0785	0.0147	0.0036	2.09
United	States	0.0008	0.0006	0.0005	0.00000048	10.26
stainless	steel	0.000002	0.00004	0.000002	0.0000000008	14.50

Gosse Bouma en Erik Tjong Kim Sang 9/32

Introduction Collocations Unknown Words Similar Words Applications Conclusions

Pointwise Mutual Information

Wikipedia fragment (10M words)

W1+W2	PMI	W1+W2	PMI
lingua franca	18.41	Suu Kyi	18.32
dementia praecox	17.51	Foo Fighters	18.25
habeas corpus	16.63	Mao Zedong	17.82
right-handed batsman	16.32	Alcoholics Anonymous	17.74
spinal cord	16.39	Leonhard Euler	17.57
assassination attempt	9.80	Public Library	9.32
social welfare	9.79	Christmas Island	9.32
cable car	9.75	Cornell University	9.24
almost certainly	9.65	National Assembly	9.20
admiration for	5.74	National Council	5.58
sets out	5.74	In 1946	5.54
if they	5.74	The Simpsons	5.53
his career	5.74	The Doors	5.49

Gosse Bouma en Erik Tjong Kim Sang 10/32

Introduction Collocations Unknown Words Similar Words Applications Conclusions

Finding the meaning of unknown words

Can we use techniques from corpus linguistics to find out the meaning of unknown words?

Example: what is a cobza?

Once very widespread in Moldavia and Muntenia, **cobzas** have become increasingly rare

The **Cobza** has become less widespread in usage than previously

Gosse Bouma en Erik Tjong Kim Sang 11/32

Introduction Collocations Unknown Words Similar Words Applications Conclusions

Word Meaning and Corpora

Meaning and Usage

You shall know a word by the company it keeps – J.R. Firth (1957)

- Words don't occur randomly in text
- Meaning of words determines in which texts they will be used
- Conversely: If you know in which texts and contexts a word is used (frequently), you can learn a lot about the meaning of the word.

Gosse Bouma en Erik Tjong Kim Sang 12/32

Introduction Collocations Unknown Words Similar Words Applications Conclusions

Word Meaning and Usage

What is a cobza?

Once very widespread in Moldavia and Muntenia, **cobzas** have become increasingly rare

The **Cobza** has become less widespread in usage than previously

A good **cobza** functions like a drum machine

The problem is, there really aren't many good **cobza** players left


The **cobza** is an instrument that is dying out fast

Recorded examples of Moldavian Csango players of **cobza** were made during the 1950s and early 70s

Gosse Bouma en Erik Tjong Kim Sang 13/32

Introduction Collocations Unknown Words Similar Words Applications Conclusions

Word Meaning and Usage



Cobza

The **cobza** (koboz in Hungarian) is a Romanian folk version of the ud, the lute found across the Islamic world from North Africa to Central Asia. The word is Turkish; "kobuz" is the more common form of it in Turkic languages, but it can be applied to many different lute-type instruments. In Europe, the **cobza/koboz** is now restricted to northern and eastern Romania, played by by Hungarian and Romanian communities, and is also played in Hungary by enthusiasts for Transylvanian Hungarian music.

Gosse Bouma en Erik Tjong Kim Sang 14/32

Introduction Collocations Unknown Words Similar Words Applications Conclusions

Contexts

Brandenburg concerto, for solo violin, two solo flutes, strings
 many of the violin and harpsichord concertos
 play the rapid solo violin passages.
 even arranged several violin concertos
 six sonatas and partitas for violin
 sound designer, and electric violin player

KWIC

A program which displays search results in this format is called *keyword in context* (KWIC).

Gosse Bouma en Erik Tjong Kim Sang 15/32

Introduction Collocations Unknown Words Similar Words Applications Conclusions

Word Meaning

Word Meaning and Corpora

We cannot learn word definitions automatically from text yet but text corpora have already been used to learn:

- ISA-relations** (a piano *is a* musical instrument, a grand piano *is a* piano, a violin *is a* musical instrument)
- Synonyms** (two words with the same meaning: *laptop, notebook*)
- Similar words**: words which belong to the same category (*violin, piano, trumpet, guitar, ...*)

This information can be used to extend dictionaries automatically.

Gosse Bouma en Erik Tjong Kim Sang 16/32

Similar Words

We will now extract similar words from a text

Our working assumption is that similar words occur in similar contexts

As contexts we will use the three words before and after the word, for example, for *violin*:

Brandenburg **concerto, for solo violin, two solo flutes**, strings ... **many of the violin and harpsichord concertos** ...

Context Vectors

A context vector contains frequencies of the words in the neighborhood of a specific word:

	solo	the	of	concerto	arranged	electric	player	sonata
violin	100	200	50	50	10	10	30	50
piano	150	400	40	100	5	0	100	160
computer	3	600	500	3	0	300	2	0

The similarity of two context vectors can be used as indication for the similarity between the two words

There are several ways for computing the similarity between two vectors

Comparing Context Vectors: Dice's coefficient

	solo	the	of	concerto	arranged	electric	player	sonata
violin	100	200	50	50	10	10	30	50
piano	150	400	40	100	5	0	100	160
computer	3	600	500	3	0	300	2	0

Dice Score

$$Dice(W1, W2) = 2 \times \frac{\text{Sum of the minimum of each column}}{\text{Sum of row W1} + \text{Sum of row W2}}$$

$$dice(violin, piano) = 2 \times \frac{100 + 200 + 40 + 50 + 5 + 0 + 30 + 50}{500 + 955} = 2 \times \frac{375}{1455} = 0.488$$

$$dice(violin, computer) = 2 \times \frac{3 + 200 + 50 + 3 + 0 + 10 + 2 + 0}{500 + 1408} = 2 \times \frac{268}{1908} = 0.280$$

Improving Context Vectors (2)

Choose Different Contexts

- choose a larger context than 3 surrounding words: 5, 10, ...
- or use all words in a sentence as context or all words in a document
- or only words in a specific syntactic relation to the keyword (adjectives, verbs, words in conjunctions, ...)
- ignore low frequent context words (words that occur less than 5 times)

Applications

- **Advertisements**
 - Google Sponsored links
 - Which search terms are associated with which products?
- **Text Keywords and Tag Clouds**
 - What are the most important words in a text?
 - Used to give an impression of a text, blog, website

Techniques for finding collocations and identifying similar words can be applied for finding answers to these questions

Improving Context Vectors (1)

Replacing Counts by Mutual Information

- Some context words are more informative than others
- Words like *the, of, and, is, ...* will occur frequently with most words
- Words like *sonata, concerto, ...* appear only with relatively few words
- We could ignore frequent words (stop words) in the context
- If we fill our vectors with **PMI scores** instead of counts, we give more importance to words that occur relatively often with the given word.

$$PMI(word_1, word_2) = \log\left(\frac{f(word_1 \text{ in context } word_2)}{f(word_1) * f(word_2)}\right)$$

Building A Dutch Similar Words Demo

A lot of computation was required

- **Creating context vectors**
 - Processing a large corpus and extracting all information
 - For Dutch syntactically annotated corpus (500M words) : 15 hrs
- **Cleaning up the vectors**
 - Replace counts by mutual information scores
 - Remove low frequency words from vectors
- **Finding Similar Words**
 - Dice-score between each word and all other words needs to be computed
 - For 10,000 most frequent Dutch words (nouns and names) 49,500,000 comparisons are necessary

Google Ads



university of groningen
Introduction Collocations Unknown Words Similar Words **Applications** Conclusions

Search terms for a Product

Google Ads

Search term	Advertisement
centrino	dell.nl
racefiets (<i>road bicycle</i>)	gazelle.nl
wielrennen (<i>cycling</i>)	google.nl (<i>sportnieuws</i>)
wielrennen (<i>cycling</i>)	vermageren.com
la marmotte (<i>French cycling event</i>)	wielerhotel Franse Alpen
la marmotte (<i>French cycling event</i>)	inspanningstest (<i>physical test</i>)
lance armstrong	quadrand.com (<i>wrist bands</i>)

Gosse Bouma en Erik Tjong Kim Sang 25/32

university of groningen
Introduction Collocations Unknown Words Similar Words **Applications** Conclusions

Ad-Words and Products

Products

Companies want to find good keywords for their products

- ipod → ???
- hotels.nl → ???
- Heineken → ???

Finding keywords

- Is it possible to find **keywords** for a given **product** automatically?
- Which words do co-occur a lot with the name of the product?

Gosse Bouma en Erik Tjong Kim Sang 26/32

university of groningen
Introduction Collocations Unknown Words Similar Words **Applications** Conclusions

Counting Products and Keywords

What are the search words most often used for a product?

We would like to know what our potential customers are searching for

This information is hard to obtain

Alternatively we can examine webpages to find words related to products: for a given **Product** and **Keyword**:

- What is the number of pages containing the word **Product**?
- What is the number of pages containing both **Product** and **Keyword**?

If $result_2$ is at least 50% of $result_1$, then **Keyword** might be a good keyword for **Product**

Gosse Bouma en Erik Tjong Kim Sang 27/32

university of groningen
Introduction Collocations Unknown Words Similar Words **Applications** Conclusions

Keywords

- **Tags** are keywords that are added by users
- Can we find **keywords** in a text (blog) automatically?
 - Good keywords are words that occur frequently in a text but rarely in other text
- Challenges:
 - General words (**and, a, the, of, by, on, is, are, have, ...**) occur frequently, but are not good keywords
 - Some word-combinations (**Barack Obama, open source**) should be treated as a single keyword

Make your own tag clouds at www.tagcrowd.com!

Gosse Bouma en Erik Tjong Kim Sang 28/32

university of groningen
Introduction Collocations Unknown Words Similar Words Applications **Conclusions**

Conclusions

- **Words do not occur at random in text**
 - Usage (context words) and meaning are related
 - Some word combinations co-occur frequently (collocations)
 - Mutual Information helps to find collocations
- **Similar words**
 - Words with similar meaning occur in similar contexts
 - Use dice-score to compare context vectors
- **Blogosphere:** *"De url (<http://www.let.rug.nl/~gosse/SetsTwNC/>) is niet erg vlot, maar sla op, bewaar, koester en hoop met mij dat de Rijksuniversiteit Groningen de site de rest van onze dagen online houdt."* (http://literairvertalen.org/van_interesse_komt_tijdverlies/index.php)

Gosse Bouma en Erik Tjong Kim Sang 29/32