

Statistics/Probability Theory for Computational Linguistics

Dietrich Klakow



The notion "probability of a sentence" is an
entirely useless one...

Noam Chomsky, 1969



Basics Paradigms in Computational Linguistics

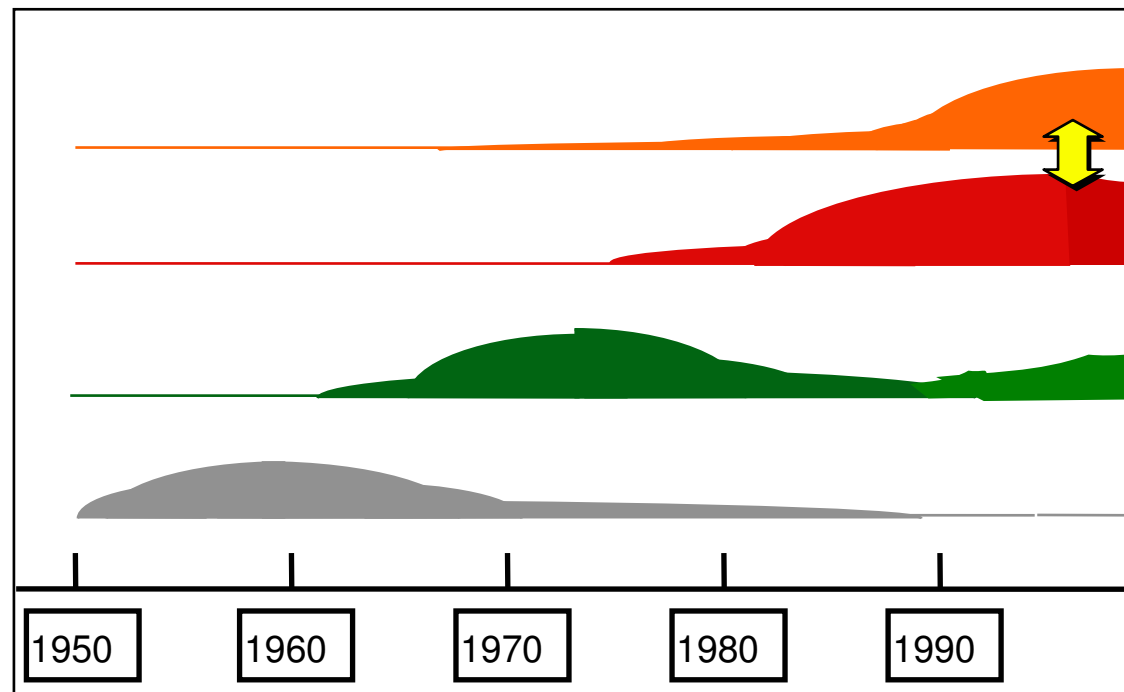


statistical methods

Declarative linguistic formalisms in CL

special methods

direct programming,
no separation of
description and
processing



Based on a slide by Hans Uszkoreit



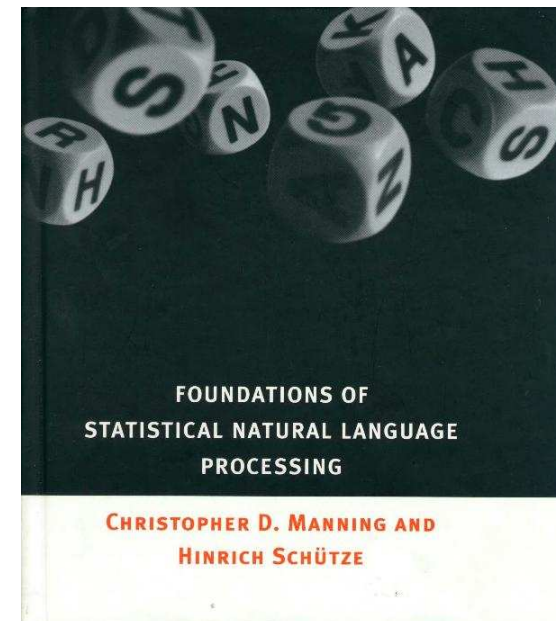
Foundations of Statistical Natural Language Processing

by Christopher D. Manning, Hinrich Schütze

Publisher: The MIT Press;
1st edition (June 18, 1999)

ISBN: 0262133601

List Price: \$77.00





Motivation



Motivation 1

- Not everything that could happen will happen
- E.g. not all readings of a sentence are equally likely



Example for Ambiguous Readings

*„Früher stellten die Frauen der Inseln am
Wochenende Kopftücher mit
Blumenmotiven her, die ihre Männer an den
folgenden Montagen auf dem
Markt im Zentrum der Hauptinsel verkauften.“*

How many different readings are there? **258.048**

Most of those readings are unlikely
a probabilistic context free grammars



Motivation 2

- Using probability theory, quite powerful systems can be developed



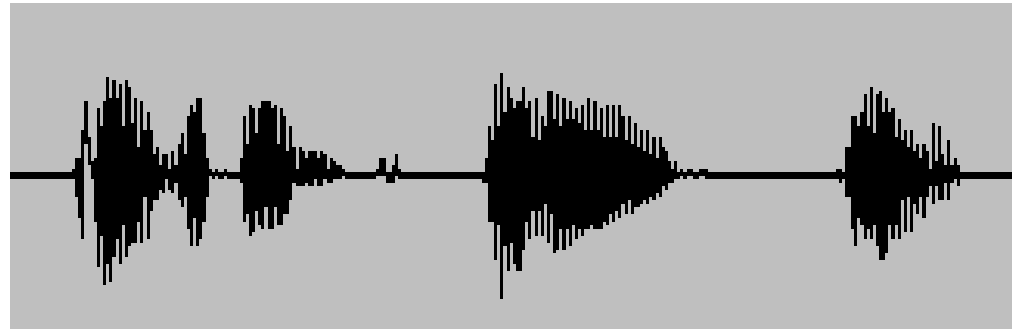
Language coding and compression

Space	Accent Mark	Ind.	Ind.	Caps.	Italics	Letter Sign	Ind.
a	c	e	d	ch	sh	wh	th
,	i	:	j	en	ow	.	w
b	f	h	g	gh	ed	ou	er
'	st	in	ar	-	ing	"	Number Sign (#)
k	m	o	n	u	x	z	y
;	s	!	t	? or "	the	(or)	with
l	p	r	q	v	and	of	for

Frequent letter combinations get their own symbol



Speech Recognition



Model probability of word sequence given
speech signal

⇒ find most likely word sequence



Example for ambiguous translations (from Leos dictionary)



band das Band

band die Band - Musikgruppe

band [tech.] das Band

band die Bandbreite

band [chem.] die Bande - im Spektrum **band** das Beffchen **band** der Bereich **band** der Bund **band** der Frequenzbereich **band** die Gruppe **band** der Gurt **band** die Kapelle **band** die Leiste **band** die Musikkapelle **band** das Orchester **band** die Schar **band** die Schnur **band** [mus.] der Spielmanszug **band** der Streifen **band** die Truppe narrowband also: narrow-band *adj.* engbandig narrowband also: narrow-band *adj.* schmalbandig sideband also: side band [elec.] [telecom.] das Seitenband **Verben und Verbzusammensetzungen** to **band** together sich verbinden to **band** together sich vereinigen to **band** together sich zusammenrotten to **band** together sich zusammentun to **band** together zu einer Gruppe vereinigen to beat the **band** nie da gewesen sein to cross-**band** [tech.] absperren [Holzverarbeitung] **Zusammengesetzte Einträge** abrasive band - cloth [tech.] das Bandschleifleinen abrasive band - paper [tech.] das Bandschleifpapier adhesive band [tech.] das Klischeeklebeband attenuating band [aviat.] der Dämpfungsbereich audio band [phys.] der Hörbereich **band** aerial die Bandantenne **band-aid** das Heftpflaster **band-aid** [Amer.] [med.] das Pflaster **band-aid** [Amer.] [med.] das Wundpflaster **band** box die Hutschachtel **band** ceramics die Bandkeramik **band** collar der Stehkragen **band-conveyor** das Fließband **band** conveyor [tech.] der Gurtförderer **band-conveyor** das Transportband **band** edge die Bandkante **band** emission [autom.] die Bandemission **band** emission [autom.] die Bandenemission **band** gap [phys.] die Bandlücke **band** gate [tech.] der Bandausschnitt - Spritzgusswerkzeug [Kunststoffe] **band** grinder [tech.] die Bandschleifmaschine **band** matrix [math.] die Bandmatrix **band** of barrel das Fassband **band** of barrel der Fassreifen **band** of radiation [phys.] der Strahlungsbereich **band** of robbers die Räuberbande **band** overlap [tech.] die Bandüberlappung **band** printer [print.] der Banddrucker **band** radiation [autom.] die Bandenstrahlung **band** resaw [tech.] die Trennbandsäge **band** saw [tech.] die Bandsäge **band-saw** die Bandsäge **band** spectrum [tech.] das Bandenspektrum **band-spread** die Bandspreizung **band-stand** der Musikpavillon **band** structure [phys.] die Bandstruktur **band-switch** der Bereichsschalter **band-switch** der Bereichsumschalter **band** width die Bandbreite **base band** [tech.] das Basisband **brake band** [tech.] das Bremsband **brass band** [mus.] die Blaskapelle **brass band** [mus.] die Blechmusik **brass band** [mus.] der Spielmanszug **broad band** [tech.] das Breitband **carrier band** [tech.] das Trägerfrequenzband **clay band** [geol.] das Salband **clincher band** [autom.] das Wulstband [Reifen] **conveyer band** das Förderband **cover band** [tech.] das Deckband **currency band** [bank.] die Währungsbandbreite **dance band** die Tanzkapelle **dead band** [metr.] die Totzone **edge band** [tech.] der Umleimer [Tischlerei] **elastic band** [tech.] das Gummiband **elastic band** der Gummistrumpf **error band** der Zufallsstreuereich **filter band** [tech.] das Siebband **flexible band** die Randzeit - Arbeitszeit **glassy band** [tech.] glasiger Streifen **guard band** [elec.] der Rasen - Abstand zwischen den Schrägspuren, den Videospuren, der benutzt wird, um eine gegenseitige Beeinflussung der Spuren zu vermeiden **guard band** [elec.] der Schutzabstand - Abstand zwischen den Schrägspuren, den Videospuren, der benutzt wird, um eine gegenseitige Beeinflussung der Spuren zu vermeiden **guard band** [elec.] [telecom.] der Schutzbereich - zwischen zwei Kanälen zur Vermeidung von Interferenzen **guard band** [telecom.] der

⇒ Word Sense Disambiguation



Part-Of-Speech Tagging

Xinhua News Agency , Guangzhou , March 16 (Reporter Chen Ji) The latest statistics show that from January through February this year , the export of high-tech products in Guangdong Province reached 3.76 billion US dollars , up 34.8% over the same period last year and accounted for 25.5% of the total export in the province .



Part-Of-Speech Tagging

Xinhua/NNP News/NNP Agency/NNP ,/,
Guangzhou/NNP ,/, March/NNP 16/CD (/
Reporter/NNP Chen/NNP Ji/NNP)/SYM
The/DT latest/JJS statistics/NNS show/VBP
that/IN from/IN January/NNP through/IN
February/NNP this/DT year/NN ,/, the/DT
export/NN of/IN high-tech/JJ products/NNS
in/IN Guangdong/NNP Province/NNP
reached/VBD 3.76/CD billion/CD US/PRP
dollars/NNS ,/, up/IN 34.8%/CD over/IN the/DT
same/JJ period/NN last/JJ year/NN and/CC
accounted/VBD for/IN 25.5%/CD of/IN the/DT
total/JJ export/NN in/IN the/DT province/NN ./.



Named Entity Tagging

Task:

Identify names of people, organizations, locations, ... in text

```
President <ENAMEX id="9"  
type="PERSON">Richard  
Nixon</ENAMEX> in <ENAMEX id="10"  
type="LOCATION">Moscow.</ENAMEX  
>
```



Information Retrieval

The screenshot shows a Microsoft Internet Explorer browser window with the title "Information Retrieval - Google Search - Microsoft Internet Explorer". The address bar contains the URL "http://www.google.de/search?hl=en&q=Information+Retrieval". The search results page displays the following information:

Web Results 1 - 10 of about 115,000,000 for **Information Retrieval**. (0.23 seconds)

Information Retrieval
An online book by CJ van Rijsbergen, University of Glasgow.
www.dcs.gla.ac.uk/Keith/Preface.html - 7k - [Cached](#) - [Similar pages](#)

Information Retrieval
Online text of a book by Dr. CJ van Rijsbergen of the University of Glasgow covering advanced topics in **information retrieval**.
www.dcs.gla.ac.uk/~iain/keith/ - 5k - [Cached](#) - [Similar pages](#)

Modern Information Retrieval
A recent IR book, covering algorithms, implementation, query languages, user interfaces, and multimedia and web **retrieval**.
www.sims.berkeley.edu/~heerst/irbook/ - 9k - [Cached](#) - [Similar pages](#)

UMASS Amherst. Center for Intelligent Information Retrieval
University of Massachusetts research lab focused on efficient access to large, heterogeneous, distributed, text and multimedia databases.
ciir.cs.umass.edu/ - 6k - [Cached](#) - [Similar pages](#)

Information Retrieval Research - SearchTools Topics
An up-to-date overview of research in the field of **information retrieval**.
www.searchtools.com/info/info-retrieval.html - 22k - [Cached](#) - [Similar pages](#)

SIGIR: Information Retrieval
"Addresses issues ranging from theory to user demands in the application of computers to the acquisition, organization, storage, **retrieval**, and distribution ..."
www.acm.org/sigir/ - [Similar pages](#)

Sponsored Links

- Text Retrieval Software**
Text search engine for PC, networks intranets & websites. Free trial.
www.isys-search.com
- Database Search Software**
Accurately retrieve records in your database with error-tolerant search
www.Netrics.com
- Volltextsuche spart Zeit**
Finden mit hilfreicher Suchfunktion in File-Servern, Intranet, Internet
www.knowledger.net
- Downloadable Papers**
Established site features papers on **Information retrieval**
<http://www.1millionpapers.com>
- Information Retrieval**
Always Find What You Need On Your Intranet. Sign Up For Info!
www.google.de/appliance
- Information Retrieval**
Find most information on...

The taskbar at the bottom shows the Start button and several open applications: Kalender - Micro..., Dokument1 - Micro..., SNLP_06_Chap0, Foundations_0506..., and Information Retri... The system clock shows 14:49.



Text Classification



e.g. Spam Mail Classification

V / a g r a \$ 3 , 3 1

A m b / e n

M e r / d i a

C / a l i s \$ 3 , 7 5

V a l / u m \$ 1 , 2 1

X & n a x

S o m &

<http://www.Chanatanxte.scriptmania.com/>



Statistical Machine Translation

মানব পরিবারের সকল সদস্যের সমান ও অবিচ্ছেদ্য অধিকারসমূহ
এবং সহজাত মর্যাদার স্বীকৃতিই হচ্ছে বিশ্বে শান্তি, স্বাধীনতা এবং
ন্যায়বিচারের ভিত্তি



Whereas recognition of the inherent
dignity and of the equal and inalienable
rights of all members of the human family
is the foundation of freedom, justice and
peace in the world



History of Probability Theory



History of Probability Theory



- Altertum
 - Search for ideal dice
 - Gambling, oracles
 - Insurances
 - Babylon, China
 - Pensions
 - Rom
- No formal approaches known





History of Probability Theory



- Medieval times
 - Research mostly done in cloisters
 - No significant progress in probability theory



History of Probability Theory



- Blaise Pascal (1623-1662)
 - Dice problems like this:
 - What is the chance that there is at least one six if you throw four dice at the same time
 - First approaches to combinatorics





History of Probability Theory



- Jakob Bernoulli (1655-1705)
 - Binomial distribution
 - Draw balls from an urn with returning them
 - Bernoulli chains
 - Law of large numbers:
 - The relative frequency of a random event will approach its theoretically expected fraction the more often the random experiment is repeated.



Law of large numbers

Number of rolls	Number of Heads		Ratio		Absolute difference	Relative difference
	Theoretical	Observed	Theoretical	Observed		
100	50	48	0.500	0.480	2	0.020
1000	500	491	0.500	0.491	9	0.009
10000	5000	4970	0.500	0.497	30	0.003



History of Probability Theory



- Abraham de Moivre (1667-1754)

- Normal distribution

- Central limit theorem

- If the random variable X is the some of infinitely many identically distributed random variables than X is normally distributed

- Simulation from

http://www.statistics4u.com/fundstat_germ/cc_central_limit.html





History of Probability Theory



- Thomas Bayes (1702–1761)
 - Conditional probabilities
 - Bayes rule

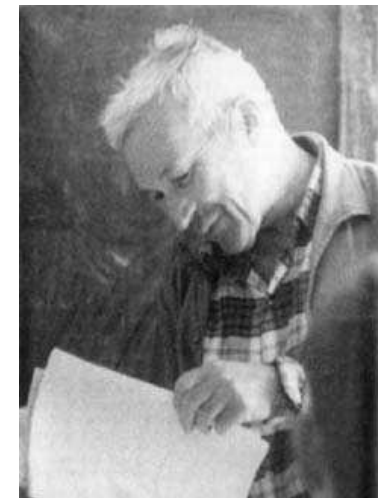




History of Probability Theory



- Andrej Kolmogorov (1903-1987)
- Axiomatic approach:
 - Probabilities are values between 0 and 1
 - Probabilities are normalised
 - Probabilities for “different” events add up





Early applications of statistics to Computational Linguistics



- Part-Of-Speech Tagging
 - Introduction of Hidden Markov Models in the mid 80s
 - In general much better than other methods known at that time
- Speech recognition
 - ca. 1980: Hidden-Markov-Modelle



Introduction to
Probability Theory



Introduction to Probability Theory



-> White board



Simple Experiments



Simple statistical experiments

- Zipf distribution
- -> Perl script



Simple statistical experiments

- Distribution of the length of questions
- -> Perl script



Correlation Function

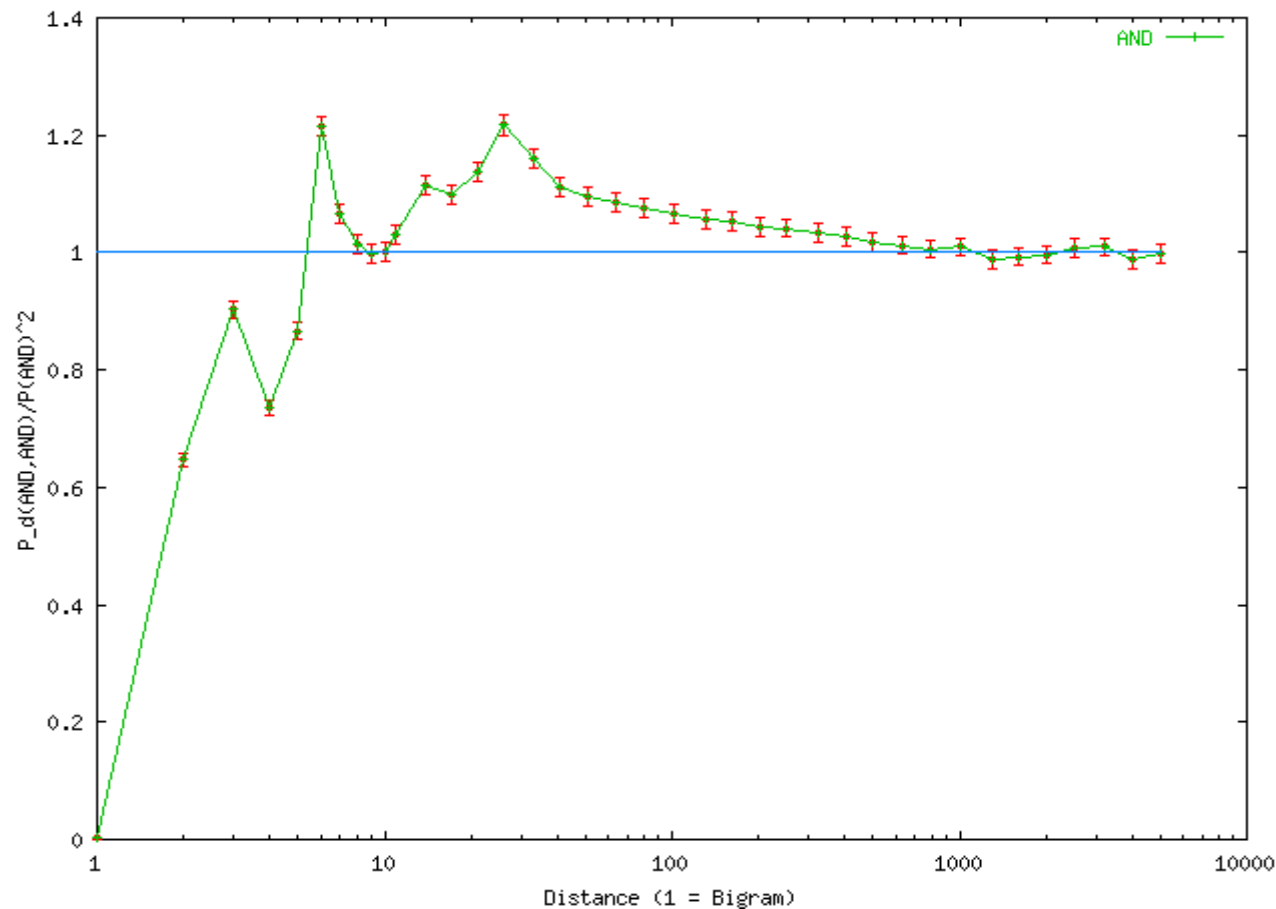
- Definition:

$$c_d(w) = \frac{P_d(w w)}{P(w)^2}$$

- d: distance of two observations of word w
- Statistical independence: $c(w)=1$



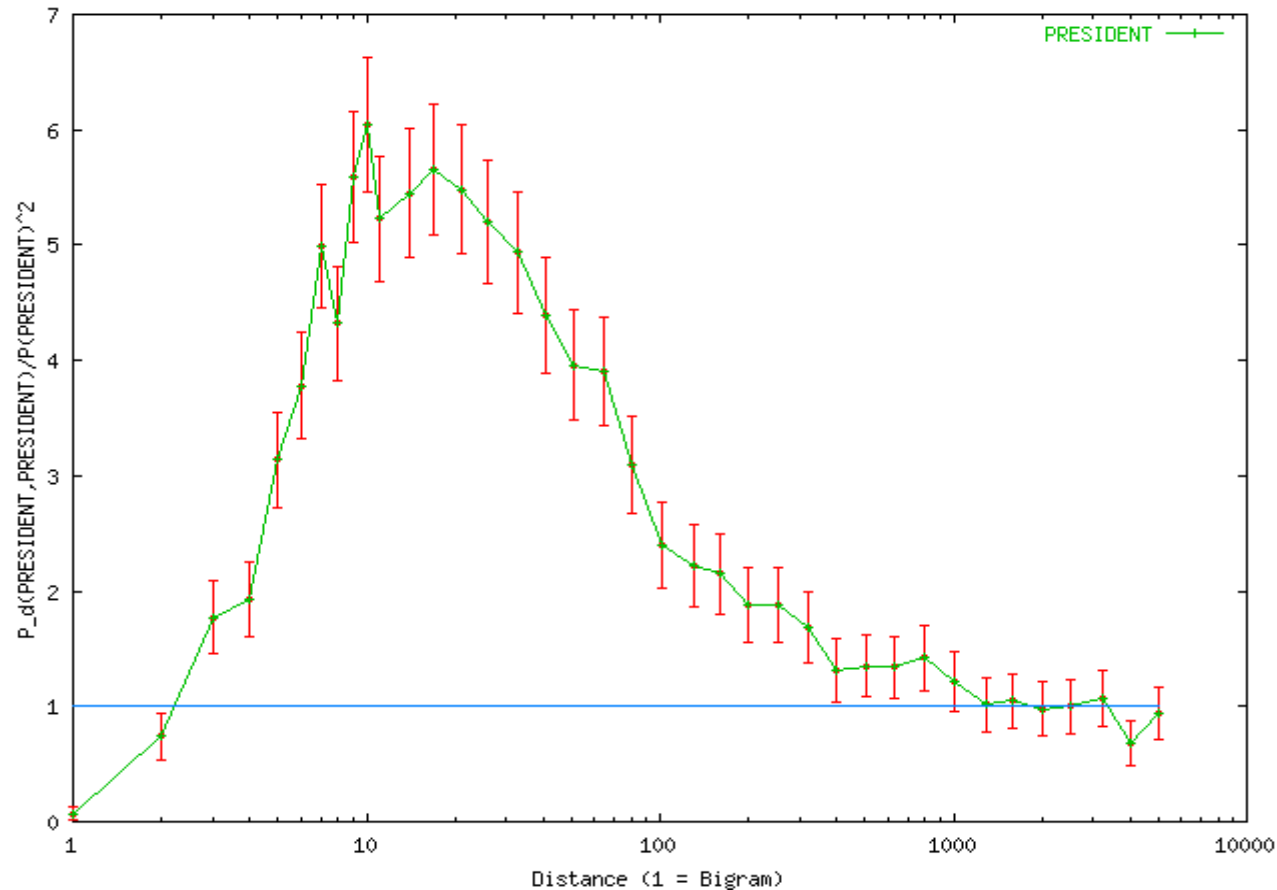
Correlation Function „and”



Only weak short
range dependencies



Correlation Function „President”

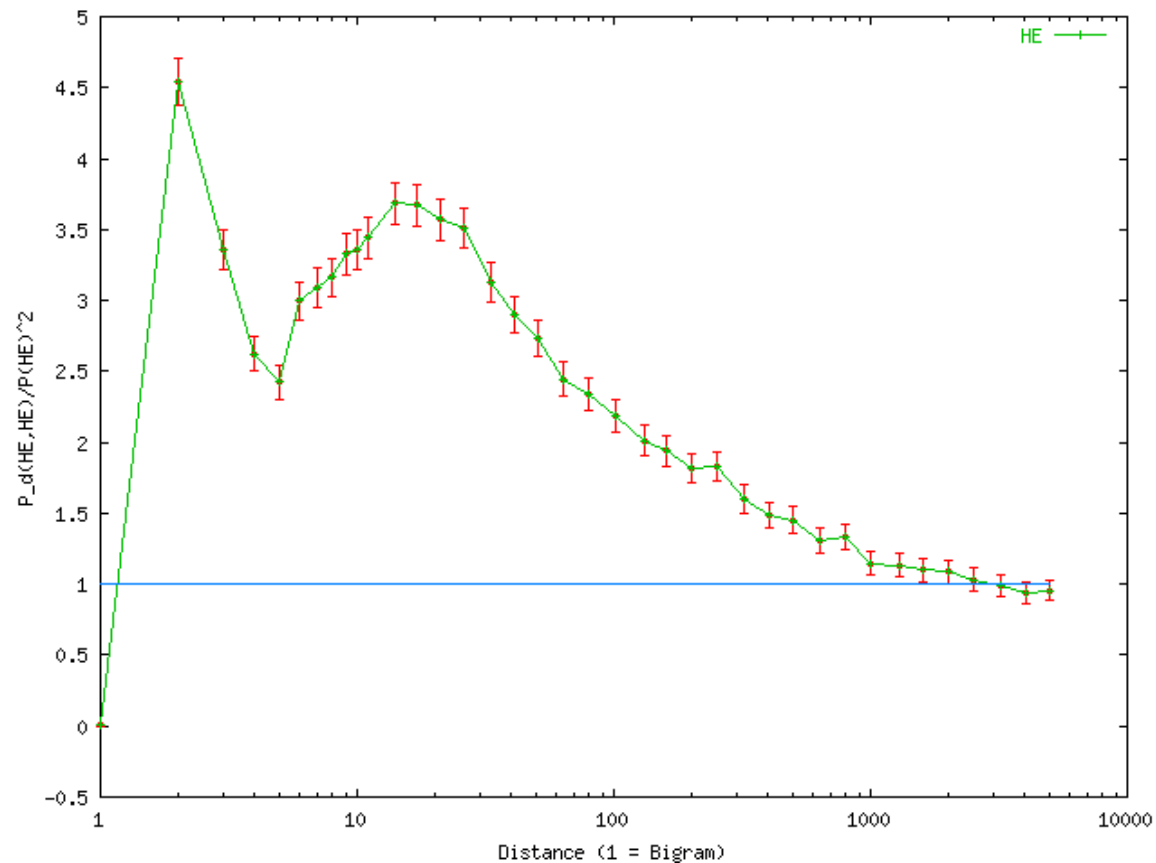


-Long range
(semantic)
dependency

-Decay of
correlations after
about 1000 words



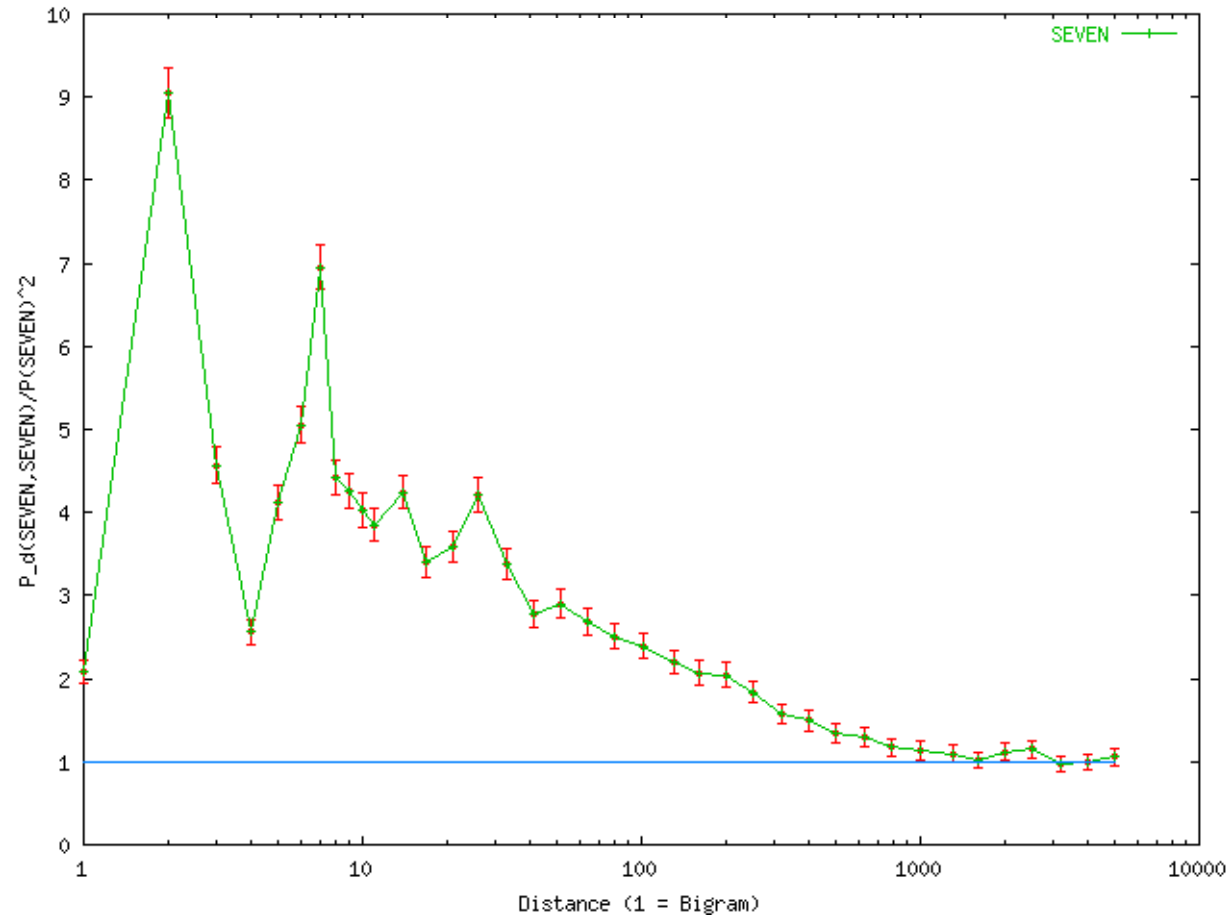
Correlation Function „he“



Short- and
Long Range
Dependencies



Correlation Function „seven”





Summary

- Examples for applications
- History of probability theory
- Introduction to basic notions
- Simple experiments