

Natural Language Processing
SoSe 2016



Introduction to Natural Language Processing

Dr. Mariana Neves

April 11th, 2016

Outline

- Introduction to Language
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- Challenges
- NLP course

Outline

- Introduction to Language
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- Challenges
- NLP course

Natural Language



(<http://expertenough.com/2392/german-language-hacks>)

日本語で

ふゆ せかいかくち いわ おこな じき
冬は世界各地でさまざまなお祝いが 行 われる時期で
す。ほんのいくつか例を挙げるだけでも、ハナカ、クリス
マス、クワンザ、新年などさまざまなお祝いがあります。
かくぶんか いわ かた
各文化によってその祝い方はさまざまですが、ほとん
どのお祝いにはごちそうが欠かせません。

(http://www.transparent.com/learn-japanese/articles/dec_99.html)

Artificial Language

```

try {
    cMessage = messageQueue.take();
    for (AsyncContext ac : queue) {
        try {
            PrintWriter acWriter = ac.get
            acWriter.println(cMessage);
            acWriter.flush();
        } catch (IOException e) {
            System.out.append(char c)
            queue.append(CharSequence s);
        }
    }
} catch (InterruptedException e) {
    printf(String form

```

(<https://netbeans.org/features/java/>)

```

def add5(x):
    return x+5

def dotwrite(ast):
    nodename = getNodeName()
    label=symbol.sym_name.get(int(ast[0]),ast[0])
    print '%s [%s]' % (nodename,label),
    if isinstance(ast[1], str):
        if ast[1].strip():
            print '= %s';' % ast[1]
        else:
            print ''
    else:
        print '['
        children = []
        for n, child in enumerate(ast[1:]):
            children.append(dotwrite(child))
        print '%s -> (%s)' % (nodename,
        for name in children:
            print '%s' % name,

```

(<http://noobite.com/learn-programming-start-with-python/>)

Language

A **vocabulary** consists of a set of **words** (w_i)



(<http://learnenglish.britishcouncil.org/en/vocabulary-games>)

A **text** is composed of a sequence of **words** from a **vocabulary**

THIS WEEK

EDITORIALS | PUBLISHING Nature journals to offer double-blind peer review [#129](#)

WORLD NEWS Trouble ahead for the European Space Agency [#130](#)

APPLIED SCIENCE Theory trade offers fresh strategy to combat invasion [#129](#)

Beyond the genome

Studies of the epigenomic signatures of many healthy and diseased human tissues could provide crucial information to link genetic variation and disease.

The Greek prefix *e*- can signify open, on, over, near, at, before, or outside. In genetics — particularly the last of them, it is some 14 years, almost to the day, that *Nature* published the draft sequence of the human genome. Since then, the field has moved on to a subsequent study on the non-genetic modifications to the genome — epigenetics — that have been shown to determine the function expressed by which cell type, and when.

It is hard to think of any branch of biology for which that has not been done more rapidly than for epigenetics. Its legacy has perhaps been

most notable in advances in our appreciation of the fact that genes are not the only determinants of what happens in the individual cells in our bodies.

The first major advance came in: Upon the genome, on the genome, over the genome — take your pick — epigenetics emerges to describe changes in the regulation of gene expression that can be passed on without changing the DNA sequence itself.

Since the initial genome sequence had been completed, it became clear that an epigenome — a map of the genome-wide

modifications made to DNA — had to be built that supports

— I mean — life.

The problem is that epigenetics is like, say, not trivial. Every cell in the body carries the same genome (with the exception of the reproductive cells), but they are not all the same.

Epigenetics is still an emerging science, but researchers are now

building tools to study epigenetic changes in the genome in a systematic and genome-wide way. In 2012, *Nature* celebrated the publication of the first genome-wide epigenetic map of the human genome, and since then because most of its results come from a small number of laboratory cell lines. Clinically useful epigenetic information must be found to be useful in the context of the many different types that make up the human body.

The first epigenomic information has now been gathered,

by the Roadmap Epigenomics Project directed by the US National Institutes of Health. It is intended to generate a genome-wide epigenomic data from stem cells, from skin cells from a variety of different tissues from healthy people, and from patients with diseases such as cancer, and neurodegenerative and metabolic disorders.

The main results of this vast project are published in this issue

starting on page 313, as well as in several other Publishing

Insights articles this week.

Insights into three fundamental aspects of epigenetics emerge:

how the epigenome affects gene expression; how the epigenetic changes that occur during development affect the adult organism; and how it changes during disease.

This week's emphasis is on the role of epigenetic information

in understanding these processes. Crucially,

what emerges is that it is not just one or two

epigenetic marks that are important to disease

rarely that simple. Instead, combinations of

epigenetic marks are important in many ways

that a single type of modification does not.

A causal link between epigenetic changes and disease is also being explored.

Identifying such changes is necessary,

and the new tools for mapping epigenetic changes are helping

disease mechanisms and design targeted

treatments. With the new wealth of data, consistent alterations in the epigenetic landscape associated with specific diseases may help

researchers to understand how epigenetic changes have a role in disease progression, or only in its onset.

One lesson that it has been difficult to relate some diseases

to epigenetic changes is that many epigenetic changes occur

in poorly understood regions of the genome, usually outside those that are well characterized. This is where the maps published today should help scientists to navigate this poorly charted landscape. By overlaying these maps, made in relevant tissue types, researchers will be able to see if epigenetic changes associated with a given disease lie in a region of the genome that is well characterized, or in one that is not. This will provide a possible lead to be explored.

Cancer is often called a disease of the genome, but the genome

is also mutated, in spliced locations. Of all diseases, cancer

can have been linked most unambiguously to epigenetic alterations.

Scientists have now mapped the epigenetic changes that affect the genomic location of the mutations that provide cancer. The new findings suggest that this is true, and they go further. They show that the epigenetic changes that are associated with a given disease can be traced to their original cell type.

In other words, epigenetics and epigenetics operate together.

Using disease using information on the genome alone has been

like trying to work with one hand tied behind the back. The new tools

of epigenomics are here. In many ways, they are the answers.

But it could help researchers decide which questions to ask. ■

A **language** is constructed of a set of all possible **texts**



(<http://www.old-engl.sh/language.php>)

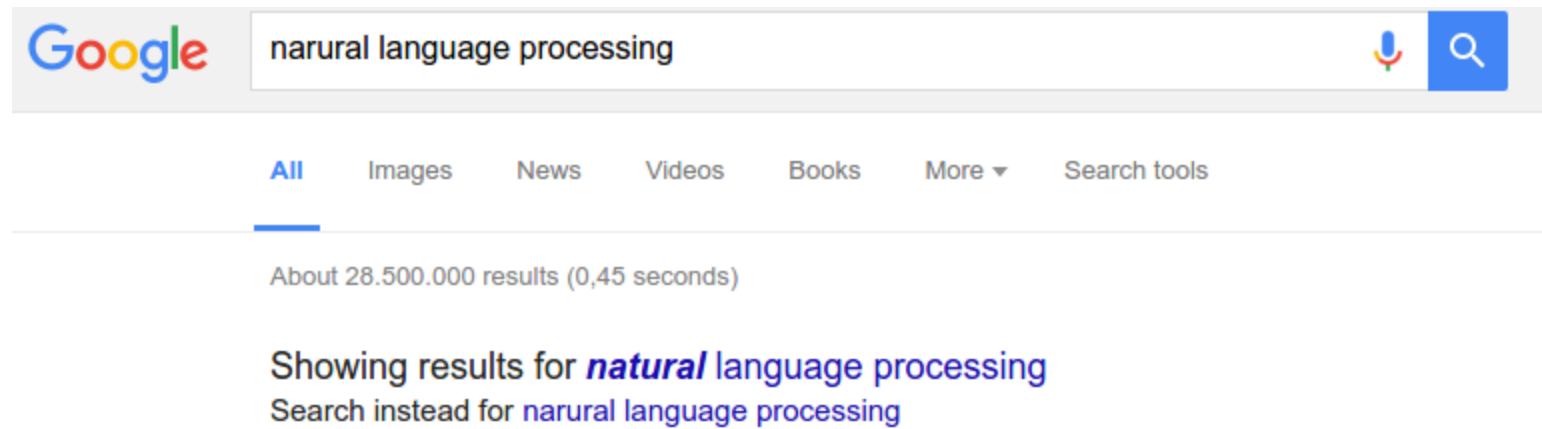
(http://www.nature.com/polopoly_fs/1.16929!/menu/main/topColumns/topLeftColumn/pdf/518273a.pdf)

Outline

- Introduction to Language
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- Challenges
- NLP course

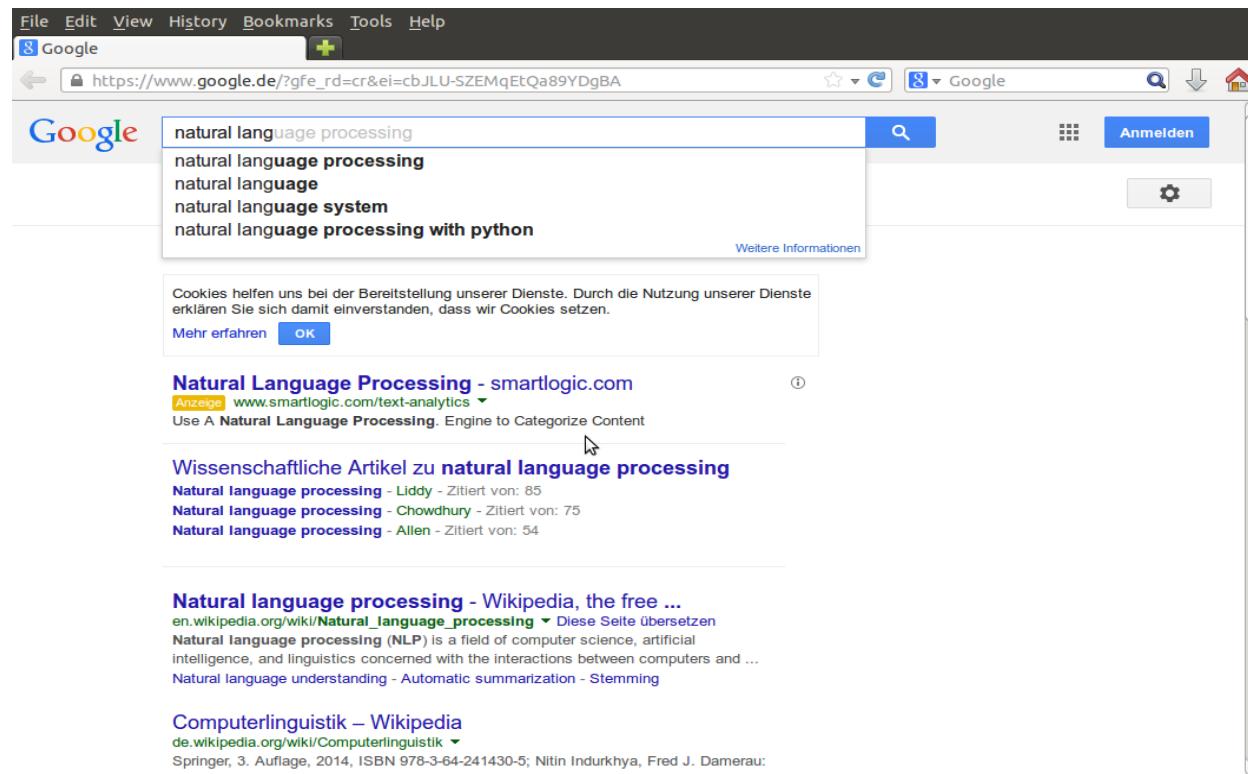
Spell and Grammar Checking

- Checking spelling and grammar
- Suggesting alternatives for the errors



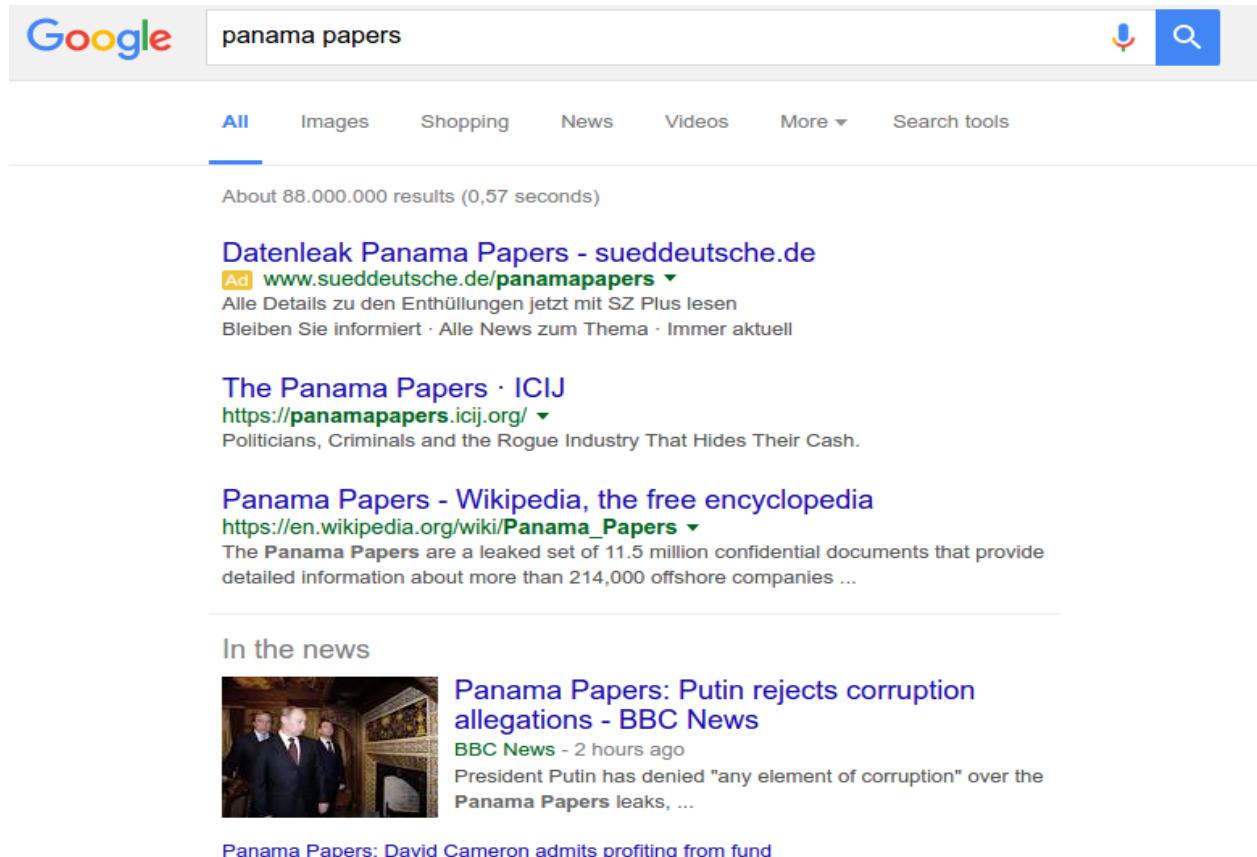
Word Prediction

- Predicting the next word that is highly probable to be typed by the user



Information Retrieval

- Finding relevant information to the user's query



Google search results for "panama papers".

Search bar: panama papers

Filter: All

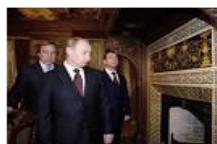
Results count: About 88.000.000 results (0,57 seconds)

1. **Datenleak Panama Papers - sueddeutsche.de**
Ad www.sueddeutsche.de/panamapapers ▾
Alle Details zu den Enthüllungen jetzt mit SZ Plus lesen
Bleiben Sie informiert · Alle News zum Thema · Immer aktuell

2. **The Panama Papers · ICIJ**
<https://panamapapers.icij.org/> ▾
Politicians, Criminals and the Rogue Industry That Hides Their Cash.

3. **Panama Papers - Wikipedia, the free encyclopedia**
https://en.wikipedia.org/wiki/Panama_Papers ▾
The Panama Papers are a leaked set of 11.5 million confidential documents that provide detailed information about more than 214,000 offshore companies ...

In the news

 **Panama Papers: Putin rejects corruption allegations - BBC News**
BBC News - 2 hours ago
President Putin has denied "any element of corruption" over the Panama Papers leaks, ...

Panama Papers: David Cameron admits profiting from fund
11 hours ago

Text Categorization

- Assigning one (or more) pre-defined category to a text

 PubMed US National Library of Medicine National Institutes of Health

PubMed Advanced

Display Settings: Abstract Send to:

[Nature](#), 2014 Mar 20;507(7492):323-8. doi: 10.1038/nature13145. Epub 2014 Mar 12.

Coupling of angiogenesis and osteogenesis by a specific vessel subtype in bone.

Kusumbe AP¹, Ramasamy SK¹, Adams RH².

[Author information](#)

Abstract
 The mammalian skeletal system harbours a hierarchical system of mesenchymal stem cells, osteoprogenitors and osteoblasts sustaining lifelong bone formation. Osteogenesis is indispensable for the homeostatic renewal of bone as well as regenerative fracture healing, but these processes frequently decline in ageing organisms, leading to loss of bone mass and increased fracture incidence. Evidence indicates that the growth of blood vessels in bone and osteogenesis are coupled, but relatively little is known about the underlying cellular and molecular mechanisms. Here we identify a new capillary subtype in the murine skeletal system with distinct morphological, molecular and functional properties. These vessels are found in specific locations, mediate growth of the bone vasculature, generate distinct metabolic and molecular microenvironments, maintain perivascular osteoprogenitors and couple angiogenesis to osteogenesis. The abundance of these vessels and associated osteoprogenitors was strongly reduced in bone from aged animals, and pharmacological reversal of this decline allowed the restoration of bone mass.

Comment in
 Bone biology: Vessels of rejuvenation. [Nature. 2014]

PMID: 24646994 [PubMed - indexed for MEDLINE]

MeSH Terms

[Aging/metabolism](#)
[Aging/pathology](#)
[Animals](#)
[Blood Vessels/anatomy & histology](#)
[Blood Vessels/cytology](#)
[Blood Vessels/growth & development](#)
[Blood Vessels/physiology*](#)
[Bone and Bones/blood supply*](#)
[Bone and Bones/cytology](#)
[Endothelial Cells/metabolism](#)
[Hypoxia-Inducible Factor 1, alpha Subunit/metabolism](#)
[Male](#)
[Mice](#)
[Mice, Inbred C57BL](#)
[Neovascularization, Physiologic/physiology*](#)
[Osteoblasts/cytology](#)
[Osteoblasts/metabolism](#)
[Osteogenesis/physiology*](#)
[Oxygen/metabolism](#)
[Stem Cells/cytology](#)
[Stem Cells/metabolism](#)

Text Categorization



Classify

Classify method: text url

Enter url to download and classify with:

<http://edition.cnn.com/2015/02/18/football/cl>

uClassify!

Remove html

1. Sports (92.8 %)
2. Entertainment (4.8 %)
3. Men (0.7 %)

[Show all classifications >>](#)

<http://www.uclassify.com/browse/mvazquez/News-Classifier>

Summarization

- Generating a short summary from one or more documents, sometimes based on a given query



This is a 7 sentence summary of <http://hpi.de/en/news/jahrgaenge/2015/des...>

Summary processing at low priority, [upgrade to BOOST](#)

Design Thinking Week: Students Improve the Daily Life Experience for People with Illiteracies

On the occasion of the World Literacy Day on September 8 more than 40 young innovators applied their Design Thinking skills in order to make life easier for these people.

Here, the focus was especially on the possibilities of using digital technologies and computers to better the daily obstacles in life of the people concerned.

Under the guidance of the D-School's coaches the teams researched, developed and prototyped - and could present many versatile solutions in the end: e.g. one of the groups came up with an idea for a software program that lets internet browsers read texts, functions and links out loud so that people with reading problems can still use news sites or social networks like Facebook.

<http://smmry.com/>

Summarization

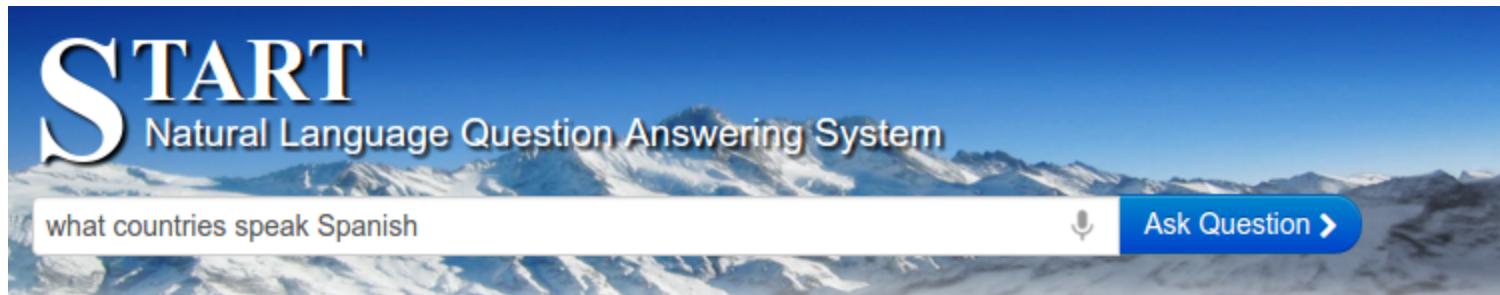


General annotation (Comments)

Function	Acts as a tumor suppressor in many tumor types; induces growth arrest or apoptosis depending on the physiological circumstances and cell type. Involved in cell cycle regulation as a trans-activator that acts to negatively regulate cell division by controlling a set of genes required for this process. One of the activated genes is an inhibitor of cyclin-dependent kinases. Apoptosis induction seems to be mediated either by stimulation of BAX and FAS antigen expression, or by repression of Bcl-2 expression. In cooperation with mitochondrial PPIF is involved in activating oxidative stress-induced necrosis; the function is largely independent of transcription. Induces the transcription of long intergenic non-coding RNA p21 (lincRNA-p21) and lincRNA-Mklm1. LincRNA-p21 participates in TP53-dependent transcriptional repression leading to apoptosis and seem to have to effect on cell-cycle regulation. Implicated in Notch signaling cross-over. Prevents CDK7 kinase activity when associated to CAK complex in response to DNA damage, thus stopping cell cycle progression. Isoform 2 enhances the transactivation activity of isoform 1 from some but not all TP53-inducible promoters. Isoform 4 suppresses transactivation activity and impairs growth suppression mediated by isoform 1 . Isoform 7 inhibits isoform 1-mediated apoptosis. Ref.70 Ref.93 Ref.95 Ref.107 Ref.110 Ref.122 Ref.125
Cofactor	Binds 1 zinc ion per subunit.
Subunit structure	Interacts with AXIN1. Probably part of a complex consisting of TP53, HIPK2 and AXIN1 By similarity . Binds DNA as a homotetramer. Interacts with histone acetyltransferases EP300 and methyltransferases HRMT1L2 and CARM1, and recruits them to promoters. In vitro, the interaction of TP53 with cancer-associated/HPV (E6) viral proteins leads to ubiquitination and degradation of TP53 giving a possible model for cell growth regulation. This complex formation requires an additional factor, E6-AP, which stably associates with TP53 in the presence of E6. Interacts (via C-terminus) with TAF1; when TAF1 is part of the TFIID complex. Interacts with ING4; this interaction may be indirect. Found in a complex with CABLES1 and TP73. Interacts with HIPK1, HIPK2, and TP53INP1. Interacts with WWOX. May interact with HCV core protein. Interacts with USP7 and SYVN1. Interacts with HSP90AB1. Interacts with CHD8; leading to recruit histone H1 and prevent transactivation activity By similarity . Interacts with ARMC10, BANP, CDKN2AIP, NUAK1, STK11/LKB1, UHRF2 and E4F1. Interacts with YWHAZ; the interaction enhances TP53 transcriptional activity. Phosphorylation of YWHAZ on 'Ser-58' inhibits this interaction. Interacts (via DNA-binding domain) with MAML1 (via N-terminus). Interacts with MKRN1. Interacts with PML (via C-terminus). Interacts with MDM2; leading to ubiquitination and proteasomal degradation of TP53. Directly interacts with FBXO42; leading to ubiquitination and degradation of TP53. Interacts (phosphorylated at Ser-15 by ATM) with the phosphatase PPP2R2C holoenzyme; regulates stress-induced TP53-dependent inhibition of cell proliferation. Interacts with PPP2R2A. Interacts with AURKA, DAXX, BRD7 and TRIM24. Interacts (when monomethylated at Lys-382) with L3MBTL1. Isoform 1 interacts with isoform 2 and with isoform 4 . Interacts with GRK5. Binds to the CAK complex (CDK7, cyclin H and MAT1) in response to DNA damage. Interacts with CDK5 in neurons. Interacts with AURKB, SETD2, UHRF2 and NOC2L. Interacts (via N-terminus) with PTK2/FAK1; this promotes ubiquitination by MDM2. Interacts with PTK2B/PYK2; this promotes ubiquitination by MDM2. Interacts with PRKCG. Interacts with PPIF; the association implicates preferentially tetrameric TP53, is induced by oxidative stress and is impaired by cyclosporin A (CsA). Interacts with human cytomegalovirus/HHV-5 protein UL123. Interacts with SNAI1; the interaction induces SNAI1 degradation via MDM2-mediated ubiquitination and inhibits SNAI1-induced cell invasion. Interacts with KAT6A. Interacts with UBC9. Interacts with ZNF385B; the interaction is direct. Interacts (via DNA-binding domain) with ZNF385A; the interaction is direct and enhances p53/TP53 transactivation functions on cell-cycle arrest target genes, resulting in growth arrest. Interacts with ANKRD2. Interacts with RFFL (via RING-type zinc finger); involved in p53/TP53 ubiquitination. Ref.8 Ref.34 Ref.38 Ref.42 Ref.43 Ref.54 Ref.55 Ref.56 Ref.57 Ref.58 Ref.59 Ref.61 Ref.62 Ref.64 Ref.65 Ref.66 Ref.67 Ref.68 Ref.72 Ref.73 Ref.74 Ref.75 Ref.76 Ref.78 Ref.80 Ref.81 Ref.83 Ref.86 Ref.87 Ref.88 Ref.89 Ref.92 Ref.93 Ref.94 Ref.99 Ref.101 Ref.103 Ref.105 Ref.106 Ref.107 Ref.112 Ref.113 Ref.116 Ref.117 Ref.119 Ref.121 Ref.122 Ref.124 Ref.125 Ref.126 Ref.127 Ref.129 Ref.137 Ref.138 Ref.139 Ref.140 Ref.141 Ref.151

Question answering

- Answering questions with a short answer



==> what countries speak Spanish

The language Spanish is spoken in Argentina, Aruba, Belize, Bolivia, Brazil, Canada, Cayman Islands, Chile, Colombia, Costa Rica, Cuba, Curacao, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Falkland Islands (Islas Malvinas), Gibraltar, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Saint Martin, Sint Maarten, Spain, Switzerland, Trinidad and Tobago, United States, Uruguay, Venezuela, and Virgin Islands.

The language Castilian Spanish is spoken in Spain.

Question Answering & Summarization

BioMedical Question Answering System
VM (166,133 documents)

What do you want to know?
which drugs can be used to treat lung cancer?

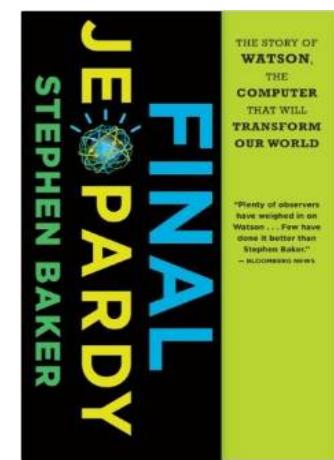
Show analysis details ASK

Amifostine (50.00%)
INJECTION, AMIFOSTINE 500 MG ADMINISTERED (50.00%)

Subsequently, qRT PCR of miR U2 1 using serum from 62 lung cancer patients and 96 various controls demonstrated that its expression levels identify lung cancer patients with 79% sensitivity and 80% specificity. miR U2 1 expression correlated with the presence or absence of lung cancer in patients with chronic obstructive pulmonary disease (COPD), other diseases of the lung - not cancer , and in healthy controls . Epidermal growth factor receptor inhibitors are used to treat advanced lung cancer patients for almost a decade. . We evaluated whether advanced LCNEC should be treated similarly to small cell lung cancer (SCLC) or non small cell lung cancer (NSCLC). INTRODUCTION : Drugs directed toward the epidermal growth factor receptor (EGFR), such as erlotinib (Tarceva) and gefitinib (Iressa), are used for the treatment of patients with advanced non small cell lung cancer (NSCLC) , including patients with brain metastases. . OBJECTIVE : To investigate the clinical significance of the expression of MHC class I chain related gene A (MICA) in patients with advanced non small cell lung cancer and explore the relationship between MICA expression and the efficacy of cytokine induced killer cell (CIK) therapy for treating advanced non small cell lung cancer. .

Question answering

- IBM Watson in Jeopardy



https://www.youtube.com/watch?v=WFR3lOm_xhE

Information Extraction

- Extracting important concepts from texts and assigning them to slot in a certain template



WIKIPEDIA
 The Free Encyclopedia

Angela Merkel



Merkel at the EPP Summit, March 2016

Chancellor of Germany	
Incumbent	
Assumed office	22 November 2005
President	Horst Köhler Christian Wulff Joachim Gauck
Deputy	Franz Müntefering Frank-Walter Steinmeier Guido Westerwelle Philipp Rösler Sigmar Gabriel
Preceded by	Gerhard Schröder
Leader of the Christian Democratic Union	
Incumbent	
Assumed office	10 April 2000
Preceded by	Wolfgang Schäuble
Minister for the Environment	

In office

17 November 1994 – 26 October 1998

Chancellor Helmut Kohl

Preceded by Klaus Töpfer

Succeeded by Jürgen Trittin

Minister for Women and Youth

In office

18 January 1991 – 17 November 1994

Chancellor Helmut Kohl

Preceded by Ursula Lehr

Succeeded by Claudia Nolte

Personal details

Born Angela Dorothea Kasner
17 July 1954 (age 61)
Hamburg, West Germany

Political party Democratic Awakening (1989–1990)
Christian Democratic Union (1990–present)

Spouse(s) Ulrich Merkel (1977–1982)
Joachim Sauer (1998–present)

Alma mater Leipzig University

Religion Lutheranism (within Evangelical Church)

Signature 

Information Extraction

- Includes named-entity recognition

Helicopters will patrol the temporary no-fly zone around [New Jersey's MetLife Stadium](#) Sunday, with [F-16s](#) based in [Atlantic City](#) **ready** to be scrambled if an unauthorized aircraft does enter the restricted airspace.

Down below, **bomb-sniffing** dogs will patrol the trains and buses that are expected to take approximately 30,000 of the [80,000-plus](#) spectators to Sunday's [Super Bowl](#) between [the Denver Broncos](#) and [Seattle Seahawks](#).

The [Transportation Security Administration](#) said it has added about two dozen dogs to monitor passengers coming in and out of the airport around the Super Bowl.

Information Extraction

 **lancet**
a Medication Event Extraction System for Clinical Text

[Project Home](#) [Downloads](#) [Wiki](#) [Issues](#) [Source](#)

[Summary](#) [People](#)

Project Information

Starred by 1 user [Project feeds](#)

Code license
[GNU GPL v2](#)

Labels
medication, extractor, lancet, discharge, summary, i2b2, NLP, challenge, 2009

Members
lijuf...@gmail.com

Lancet is a supervised machine-learning system that automatically extracts medication events consisting of medication names and information pertaining to their prescribed use (dosage, mode, frequency, duration and reason) from lists or narrative text in medical discharge summaries.

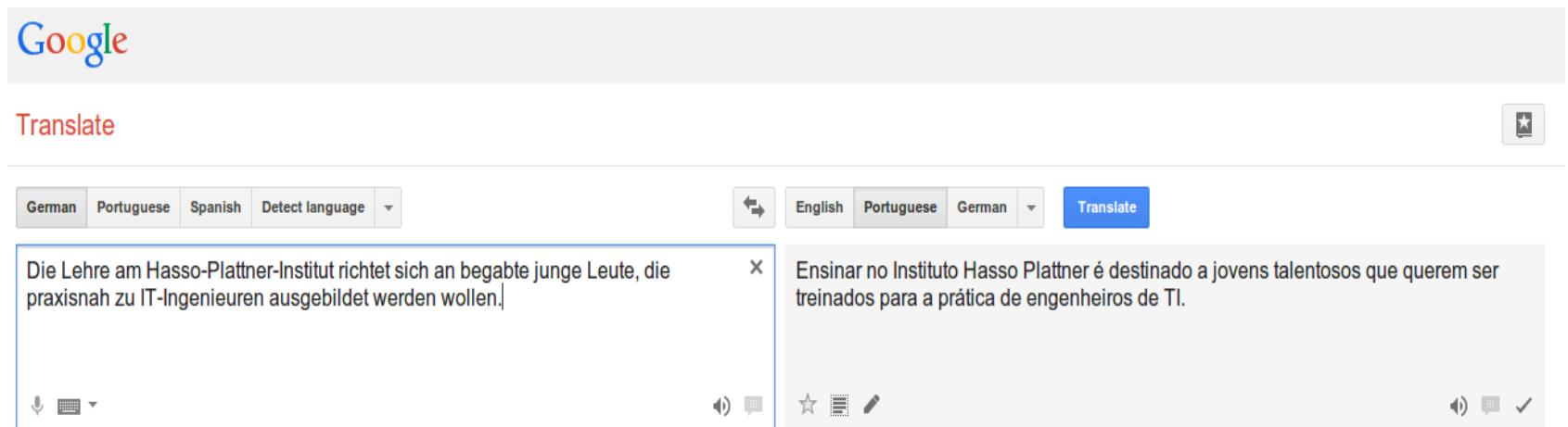
Thus, she was transitioned over to a ciprofloxacin 700 mg p.o. b.i.d. regime for a total of 12 days for a presumed urinary tract infection.

narrative

■ = medication ■ = dosage ■ = manner ■ = frequency ■ = duration ■ = reason

Machine Translation

- Translating a text from one language to another



Sentiment Analysis

- Identifying sentiments and opinions stated in a text

Customer Reviews

Speech and Language Processing, 2nd Edition



The most helpful favorable review

4 of 4 people found the following review helpful

Great introductions and reference book
I read the first edition of that book and it is terrific. The second edition is much more adapted to current research. Statistical methods in NLP are more detailed and some syntax-based approaches are presented. My specific interest is in machine translation and dialogue systems. Both chapters are extensively rewritten and much more elaborated. I believe this book is...

[Read the full review >](#)

Published on August 9, 2008 by carheg

› See more [5 star](#), [4 star](#) reviews

37 of 37 people found the following review helpful

Good description of the problems in the field, but look elsewhere for practical solutions
The authors have the challenge of covering a vast area, and they do a good job of highlighting the hard problems within individual sub-fields, such as machine translation. The availability of an accompanying Web site is a strong plus, as is the extensive bibliography, which also includes links to freely available software and resources.

Now for the...

[Read the full review >](#)

Published on April 2, 2009 by P. Nadkarni

› See more [3 star](#), [2 star](#), [1 star](#) reviews

Vs.

Optical Character Recognition

- Recognizing printed or handwritten texts and converting them to computer-readable texts



Speech recognition

- Recognizing a spoken language and transforming it into a text



Siri.
Your wish is
its command.

Siri lets you use your voice to send messages, schedule meetings, place phone calls, and more. Ask Siri to do things just by talking the way you talk. Siri understands what you say, knows what you mean, and even talks back. Siri is so easy to use and does so much, you'll keep finding more and more ways to use it.

Speech synthesis

- Producing a spoken language from a text



Spoken dialog systems

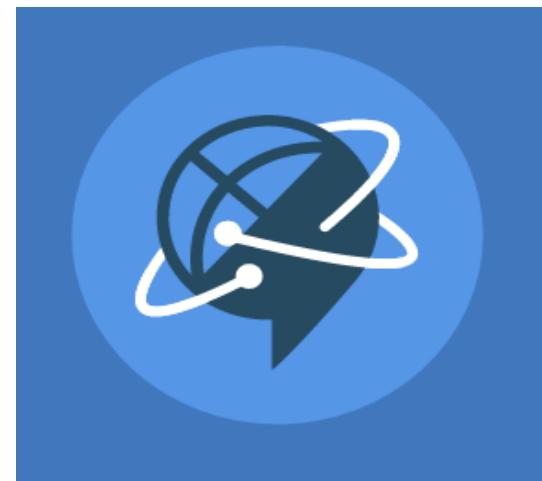
- Running a dialog between the user and the system



Siri.
Your wish is
its command.

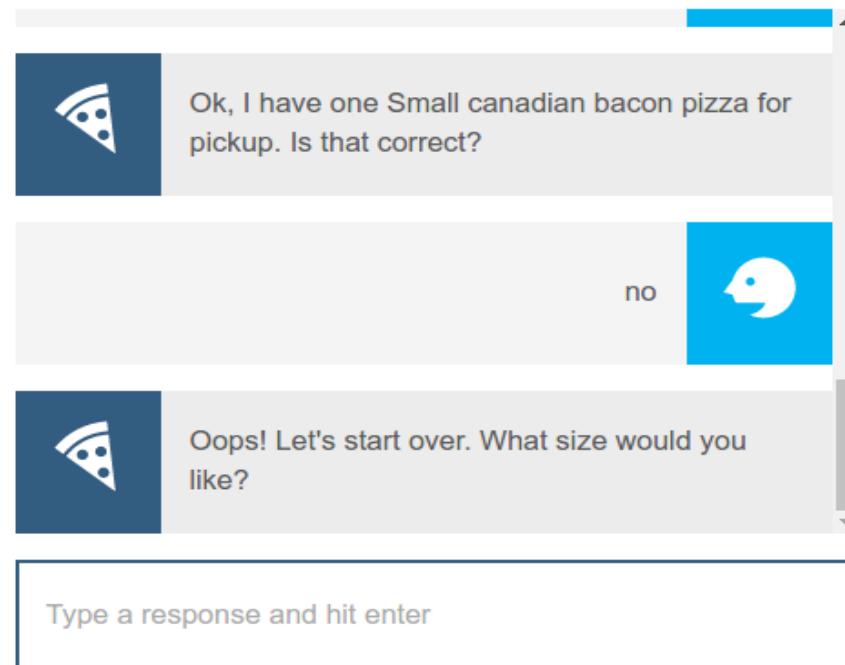
Siri lets you use your voice to send messages, schedule meetings, place phone calls, and more. Ask Siri to do things just by talking the way you talk. Siri understands what you say, knows what you mean, and even talks back. Siri is so easy to use and does so much, you'll keep finding more and more ways to use it.

IBM Watson Developer Cloud



Spoken dialog systems

Try the service



The screenshot shows a mobile-style interface for a spoken dialog system. At the top, there's a blue header bar with a progress bar. Below it, the first message from the system is displayed: "Ok, I have one Small canadian bacon pizza for pickup. Is that correct?". To the left of the text is a blue square icon containing a white pizza slice. In the bottom right corner of this message area, there's a small blue square with a white speech bubble icon. Below this message, a user response is shown: "no". To the left of the word "no" is another blue square with a white pizza slice icon. The next message from the system is: "Oops! Let's start over. What size would you like?". This message also has a blue square with a white pizza slice icon to its left. At the bottom of the screen is a large input field with a dark blue border, containing the placeholder text "Type a response and hit enter".

(<http://dialog-demo.mybluemix.net/>)

Level of difficulties

- Easy (mostly solved)
 - Spell and grammar checking
 - Some text categorization tasks
 - Some named-entity recognition tasks

Level of difficulties

- Intermediate (good progress)
 - Information retrieval
 - Sentiment analysis
 - Machine translation
 - Information extraction

Level of difficulties

- Difficult (still hard)
 - Question answering
 - Summarization
 - Dialog systems

Outline

- Introduction to Language
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- Challenges
- NLP course

Section splitting

- Splitting a text into sections

Eur Radiol
DOI 10.1007/s00330-014-3135-8

BREAST

Correlation between three-dimensional ultrasound features and pathological prognostic factors in breast cancer

Jun Jiang · Ya-qing Chen · Yi-jian Xu · Ming-h Chen ·
Yun-kai Zhu · Wen-bin Guan · Xiao-jin Wang

Received: 13 November 2013 / Revised: 30 January 2014 / Accepted: 17 February 2014
© European Society of Radiology 2014

Abstract

Objectives To investigate the correlation of three-dimensional (3D) ultrasound features with prognostic factors in invasive ductal carcinomas.

Methods Surgical resection specimens of 85 invasive ductal carcinomas of 85 women who had undergone 3D ultrasound were included. Morphology features and vascularization perfusion on 3D ultrasound were evaluated. Pathologic prognostic factors, including tumour size, histological grade, lymph node status, oestrogen and progesterone receptor status (ER, PR), c-erbB-2 and p53 expression, and microvessel density (MVD) were determined. Correlations of 3D ultrasound features and prognostic factors were analyzed.

Results The retraction pattern in the coronal plane had a significant value as an independent predictor of a small tumour size ($P=0.014$), a lower histological grade ($P=0.009$) and positive ER or PR expression status ($P=0.001$, $P=0.041$). The retraction pattern with a hyperechoic ring only existed in low-grade and ER-positive tumours. The presence of the hyperechoic ring strengthened the ability of the retraction pattern to predict a good prognosis of breast cancer. The increased intra-tumour vascularization index (VI) mean

tumour vascularity) reflected a higher histological grade ($P=0.025$) and had a positive correlation with MVD ($t=0.530$, $P=0.001$).

Conclusion The retraction pattern and histogram indices of VI provided by 3D ultrasound may be useful in predicting pathologic prognostic factors to determine whether 3D ultrasound could be used in the non-invasive prognostic evaluation of breast cancer.

Key points

- Three-dimensional ultrasound can potentially provide prognostic evaluation of breast cancer.

- The retraction pattern and hyperechoic ring in the coronal plane suggest good prognosis.

- The increased intra-tumour vascularization index reflects a higher histological grade.

- The intra-tumour vascularization index is positively correlated with microvessel density.

Keywords

Breast · Neoplasms · Ultrasound · Three-dimensional · Prognostic factors

Introduction

The three strongest prognostic factors in invasive breast cancer are widely accepted to be the size of tumour, histological grade and lymph node stage. The larger tumour size (>2 cm), high nuclear grade, and lymph node-positive status usually predict the aggressive biological behavior with a high recurrence rate and a low survival rate. In addition, the tumor size and lymph node status greatly influence the choice of operative procedure and the decision to administer neoadjuvant chemotherapy [1, 2].

Biochemical markers such as oestrogen receptors (ER), progesterone receptors (PR), human epidermal growth factor receptor 2 (c-erbB-2) and the p53 index can also be used for prediction of medical treatment response and patient prognosis. The presence of ER and PR in breast cancer always

determined and the region of interest had been identified, the volume box was superimposed and set so to include the entire display screen so as to cover the lesion and maximum amount of normal surrounding tissue. The sweep angle was adjusted to 15–29° according to the size of the breast lesion. Then the ultrasound probe was held still with enough gently to contact the skin gently. The volume mode was switched on and the 3D ultrasound volume was generated by the automatic rotation of the mechanical transducer. When the first ultrasound examination was finished, the power Doppler mode was added for the second examination and the fixed preselected power Doppler settings used were 0.3 kHz pulse repetition frequency, “low 1” wall motion filter, –2.0 gain and high frequency. The first examination for 3D grayscale imaging took 10–20 s and the second, for 3D power Doppler imaging, took 25–45 s, depending on the size of the tumour. Then the total acquisition time for 3D ultrasound was about 1–2 min. The entire examination was saved in DICOM format and stored on the hard disk for further analysis.

Image analysis

The 3D ultrasound images were reviewed for this analysis by another two radiologists with 8–10 years of experience in breast ultrasound and characterized by consensus. In addition, the radiologists had not performed the data acquisition and were blinded to the patients' clinical and mammographic findings.

The ultrasound image was opened by using the 4D View software. First, the tomographic ultrasound imaging (TUI) was used for a slice by slice documentation in the coronal plane. Then, the volume contract imaging (VCI) and the surface render mode were added for better observation of the lesion and the surrounding tissue. All the slices were carefully observed to identify the presence of the retraction pattern in the surrounding tissue and the margin of the lesion. The retraction pattern was defined as the hyperechoic straight lines that radiated perpendicularly from the surface of the solid nodule, producing a stellar pattern [3, 4] (Fig. 1). The presence of the retraction pattern was further divided into with or without a hyperechoic ring, which was displayed as an echogenic halo ring between the mass and the surrounding tissue in the coronal plane (Fig. 2a).

Materials and methods

Patients

This retrospective study was approved by the ethical standards of the institutional ethics committee, and informed consent was obtained from all patients.

From September 2011 to May 2013, 85 patients with 85 lesions, pathologically proven to be invasive ductal carcinoma, were included in this study. The exclusion criteria were pregnancy or lactation, administration of preoperative chemotherapy or adjuvant chemotherapies. Patients with a breast mass larger than 3.0 cm were also excluded because more than one 3D volume acquisition was necessary to include the whole lesion plus 3 mm surrounding the breast lesion. All patients were female and aged 26–90 years (mean age, 56.3 years).

Ultrasound examination

All ultrasound images were obtained with one type of system (GE Voluson E8 Expert, Zipf, Austria) by two radiologists with 7–12 years of experience in breast ultrasound. An 11-LD linear transducer with a frequency of 5–12 MHz was used for 2D ultrasound, and an RSP6-1D dedicated volume transducer with a frequency of 6–12 MHz was used for 3D ultrasound.

Ultrasound examination was performed with patients in the supine position with elevated arms. Once the breast lesion was

detected and the region of interest had been identified, the volume box was superimposed and set to include the entire display screen so as to cover the lesion and maximum amount of normal surrounding tissue. The sweep angle was adjusted to 15–29° according to the size of the breast lesion. Then the ultrasound probe was held still with enough gently to contact the skin gently. The volume mode was switched on and the 3D ultrasound volume was generated by the automatic rotation of the mechanical transducer. When the first ultrasound examination was finished, the power Doppler mode was added for the second examination and the fixed preselected power Doppler settings used were 0.3 kHz pulse repetition frequency, “low 1” wall motion filter, –2.0 gain and high frequency. The first examination for 3D grayscale imaging took 10–20 s and the second, for 3D power Doppler imaging, took 25–45 s, depending on the size of the tumour. Then the total acquisition time for 3D ultrasound was about 1–2 min. The entire examination was saved in DICOM format and stored on the hard disk for further analysis.

Image analysis

The 3D ultrasound images were reviewed for this analysis by another two radiologists with 8–10 years of experience in breast ultrasound and characterized by consensus. In addition, the radiologists had not performed the data acquisition and were blinded to the patients' clinical and mammographic findings.

The ultrasound image was opened by using the 4D View software. First, the tomographic ultrasound imaging (TUI) was used for a slice by slice documentation in the coronal plane. Then, the volume contract imaging (VCI) and the surface render mode were added for better observation of the lesion and the surrounding tissue. All the slices were carefully observed to identify the presence of the retraction pattern in the surrounding tissue and the margin of the lesion. The retraction pattern was defined as the hyperechoic straight lines that radiated perpendicularly from the surface of the solid nodule, producing a stellar pattern [3, 4] (Fig. 1). The presence of the retraction pattern was further divided into with or without a hyperechoic ring, which was displayed as an echogenic halo ring between the mass and the surrounding tissue in the coronal plane (Fig. 2a).

The 3D power Doppler imaging analyses were performed using a virtual organ computer-aided analysis (VOCAL)-imaging program (GE, Zipf, Austria), which could automatically calculate the histogram indices of vascularization index (VI), flow index (FI) and vascularization flow index (VFI). VI represents the vessels in the total volume by multiplying the number of colour-coded pixels in the region of interest, i.e. the mean tumour vascularity. FI represents the average intensity of flow by measuring the mean colour value in the colour voxels, i.e. the mean blood flow volume; VFI represents both

Eur Radiol

DOI

Eur Radiol

regression modelling techniques to identify the most significant and independent 3D image findings. A P value less than 0.05 was considered statistically significant.

Results

Prognostic factors

Regression modelling

In the current study group, the surgical specimens revealed 75 lesions with invasive ductal carcinoma and the remaining 10 lesions with invasive ductal carcinoma with DCIS components. The mean percentage of the DCIS components in the lesion was 8.10 ± 4.93 % (range, 2–20 %).

The size of 85 lesions ranged from 5 to 30 mm, and the mean size was 19.92 mm (SD ± 7.56 mm). Of the 85 tumours, 47 (55.3 %) were equal to or smaller than 2 cm and 38 (44.7 %) were larger than 2 cm. According to the Elston–Ellis grading system, there were 58 (68.2 %) grade II tumours and 27 (31.8 %) grade III. Lymph node metastasis was present in 30 (35.3 %) patients. There were 58 (68.2 %) ER-positive, 54 (63.5 %) PR-positive, 70 (82.4 %) c-erbB-2-positive and 42 (49.4 %) p53-positive tumours.

Correlation between MVD and prognostic factors

Significantly higher MVD was observed in the larger size group ($P<0.01$) and higher grade group ($P<0.05$). There were no significant associations between MVD and other pathologic factors ($P>0.05$) (Table 1).

Correlation between morphological features and prognostic factors

Of the 85 breast lesions, 57 (67.1 %) showed the retraction pattern in the coronal plane of 3D ultrasound. Of these 57 lesions, 17 (29.8 %) showed the retraction pattern with a hyperechoic ring and 40 (70.2 %) were without the hyperechoic ring (Fig. 2a).

The tumour size, histological grade, ER and PR status all showed significant associations with the presence of the retraction pattern ($P<0.01$) (Table 2). Tumours with the retraction pattern were significantly more likely to be small in size, low grade, ER-positive and PR-positive (Fig. 3). Moreover, the retraction pattern with a hyperechoic ring, which presented as intricately mixed fibrous tissue and infiltrating carcinoma cells on pathologic specimens, only existed in low-grade and ER-positive tumours (Fig. 2b).

The odds ratios of tumour size, tumour grade, ER and PR status for association with the retraction pattern and a hyperechoic ring versus no retraction pattern were all higher than those with the retraction pattern without a hyperechoic ring versus no retraction pattern (Table 3). The presence of the hyperechoic ring strengthened

Table 1 Association between MVD and prognostic factors				
Prognostic factor	N	Mean	SD	P value
Tumour size (cm)				
≤2	47	19.30	5.25	
>2	38	25.60	7.60	0.007
Tumour grade				
I/II	58	19.83	5.55	
III	27	25.83	8.02	0.023
Lymph node				
Negative	55	21.31	6.70	
Positive	30	22.08	7.34	0.946
ER				
Negative	27	23.27	8.36	
Positive	58	20.93	5.14	0.931
PR				
Negative	31	25.00	8.59	
Positive	54	19.82	5.09	0.092
c-erb-B-2				
Negative	15	21.50	9.57	
Positive	70	21.55	6.65	0.788
p53				
Negative	43	23.13	7.04	
Positive	42	19.63	6.20	0.083

the ability of the retraction pattern to predict these good prognoses. However, the lymph node status and the expression of c-erbB-2 and p53 showed no statistically significant correlation with the retraction pattern ($P>0.05$).

As for MVD, however, no significant correlation was found between MVD and the presence of the retraction pattern on 3D ultrasound ($P>0.05$).

Correlation between vascularization perfusion and prognostic factors

For intra-tumoral regions, the mean VI, FI and VFI of 85 lesions were 6.84 (range, 0.02–21.61), 3.772 (range, 21.81–53.32) and 2.61 (range, 0.04–9.11), respectively. For shells with a thickness of 3 mm surrounding the breast lesion, the VI, FI and VFI were 7.31 (range, 0.14–25.13), 3.872 (range, 23.27–56.90) and 2.88 (range, 0.04–11.93), respectively. Compared with the small tumours, the tumours with a diameter greater than 2 cm were more likely to show a higher inVI, inFI, inVFI, out3mmVI and out3mmVFI. The tumours with a high grade or lymph node-positive status had a higher inVI, inFI, out3mmVI and out3mmVFI than the tumours with low grade or lymph node-negative status. ER-negative tumours had a higher inFI than ER-positive tumours and the tumours with negative expression of PR had a higher inVI, inVFI and out3mmVFI than PR-positive tumours (Table 4).

Sentence splitting

- Splitting a text into sentences

11 Sentences (= "T-" or "Terminable" units *only* if independent clauses are punctuated as separate sentences, e.g. "I came and he went"-->"I came. And he went.")

Average 23.55 words (SD=12.10)

OBJECTIVES: To investigate the correlation of three-dimensional (3D) ultrasound features with prognostic factors in invasive ductal carcinoma.

METHODS: Surgical resection specimens of 85 invasive ductal carcinomas of 85 women who had undergone 3D ultrasound were included.

Morphology features and vascularization perfusion on 3D ultrasound were evaluated.

Pathologic prognostic factors, including tumour size, histological grade, lymph node status, oestrogen and progesterone receptor status (ER, PR), c erbB-2 and p53 expression, and microvessel density (MVD) were determined.

Correlations of 3D ultrasound features and prognostic factors were analysed.

RESULTS: The retraction pattern in the coronal plane had a significant value as an independent predictor of a small tumour size ($P \#8201;= 0.014$), a lower histological grade ($P \#8201;= 0.009$) and positive ER or PR expression status ($P \#8201;= 0.001, 0.044$).

The retraction pattern with a hyperechoic ring only existed in low-grade and ER-positive tumours.

The presence of the hyperechoic ring strengthened the ability of the retraction pattern to predict a good prognosis of breast cancer.

The increased intra-tumour vascularization index (VI, the mean tumour vascularity) reflected a higher histological grade ($P \#8201;= 0.025$) and had a positive correlation with MVD ($r \#8201;= 0.530, P \#8201;= 0.001$).

CONCLUSIONS: The retraction pattern and histogram indices of VI provided by 3D ultrasound may be useful in predicting prognostic information about breast cancer.

KEY POINTS: • Three-dimensional ultrasound can potentially provide prognostic evaluation of breast cancer. • The retraction pattern and hyperechoic ring in the coronal plane suggest good prognosis. • The increased intra-tumour vascularization index reflects a higher histological grade. • The intra-tumour vascularization index is positively correlated with microvessel density.

Part-of-speech tagging

- Assigning a syntactic tag to each word in a sentence

Stanford Parser

Please enter a sentence to be parsed:

Surgical resection specimens of 85 invasive ductal carcinomas of 85 women who had undergone 3D ultrasound were included.

Language: English ▾

[Sample Sentence](#)

[Parse](#)

Your query

Surgical resection specimens of 85 invasive ductal carcinomas of 85 women who had undergone 3D ultrasound were included.

Tagging

Surgical/NNP resection/NN specimens/NNS of/IN 85/CD invasive/JJ
ductal/JJ carcinomas/NNS of/IN 85/CD women/NNS who/WP had/VBD
undergone/VBN 3D/CD ultrasound/NN were/VBD included/VBN ./.

Parsing

- Building the syntactic tree of a sentence

Parse

```
(ROOT
  (S
    (NP
      (NP (NNP Surgical) (NN resection) (NNS specimens))
      (PP (IN of)
        (NP
          (NP (CD 85) (JJ invasive) (JJ ductal) (NNS carcinomas))
          (PP (IN of)
            (NP
              (NP (CD 85) (NNS women))
              (SBAR
                (WHNP (WP who))
                (S
                  (VP (VBD had)
                    (VP (VBN undergone)
                      (NP (CD 3D) (NN ultrasound)))))))))))
        (VP (VBD were)
          (VP (VBN included))))
      (. .)))
```

Parsing

- Building the syntactic tree of a sentence

Typed dependencies

```
nn(specimens-3, Surgical-1)
nn(specimens-3, resection-2)
nsubjpass(included-18, specimens-3)
prep(specimens-3, of-4)
num(carcinomas-8, 85-5)
amod(carcinomas-8, invasive-6)
amod(carcinomas-8, ductal-7)
pobj(of-4, carcinomas-8)
prep(carcinomas-8, of-9)
num(women-11, 85-10)
pobj(of-9, women-11)
nsubj(undergone-14, who-12)
aux(undergone-14, had-13)
rcmod(women-11, undergone-14)
num(ultrasound-16, 3D-15)
dobj(undergone-14, ultrasound-16)
auxpass(included-18, were-17)
root(ROOT-0, included-18)
```

Named-entity recognition

- Identifying pre-defined entity types in a sentence

b2cas Annotate Help API Widget About Contact

HIGHLIGHT

All None

✓ Anatomy
✓ Disorders
✓ Chemicals
✓ Genes and Proteins
✓ Cellular Components
✓ Molecular Functions
✓ Biological Processes
✓ Ambiguous

In Duchenne muscular dystrophy (DMD), the infiltration of skeletal muscle by immune cells aggravates disease, yet the precise mechanisms behind these inflammatory responses remain poorly understood. Chemoattractant cytokines, or chemokines, are considered essential recruiters of inflammatory cells to the tissues. We assayed chemokine and chemokine receptor expression in DMD muscle biopsies ($n = 9$, average age 7 years) using immunohistochemistry, immunofluorescence, and *in situ* hybridization. CXCL1, CXCL2, CXCL3, CXCL8, and CXCL11, absent from normal muscle fibers, were induced in DMD myofibers. CXCL11, CXCL12, and the ligand-receptor couple CCL2-CCR2 were upregulated on the blood vessel endothelium of DMD patients. CD68(+) macrophages expressed high levels of CXCL8, CCL2, and CCL5. Our data suggest a possible beneficial role for CXCR1/2/4 ligands in managing muscle fiber damage control and tissue regeneration. Upregulation of endothelial chemokine receptors, and CXCL8, CCL2, and CCL5 expression by cytotoxic macrophages, may regulate myofiber necrosis.

Load text Annotated 46 concept occurrences in 0.173s. Export ▾

New to b2cas? Take the tour »

Concept Tree

+ Expand All - Collapse All ⌂ Toggle All

- + **Anatomy (12)**
 - **Disorders (4)**
 - ⌚ DMD (1)
 - ⌚ Duchenne muscular dystrophy (1)
 - ⌚ infiltration (1)
 - ⌚ inflammatory responses (1)
- + **Chemicals (2)**
 - **Genes and Proteins (11)**
 - **Cellular Components (3)**
 - **Molecular Functions (1)**
 - **Biological Processes (9)**

Word sense disambiguation

- Figuring out the exact meaning of a word or entity

Noun 1. tie - neckwear consisting of a long narrow piece of material worn (mostly by men) under a collar and tied in knot at the front; "he stood in front of the mirror tightening his necktie"; "he wore a vest and tie"

[necktie](#)

[bola](#), [bola tie](#), [bolo](#), [bolo tie](#) - a cord fastened around the neck with an ornamental clasp and worn as a necktie

[bow tie](#), [bow-tie](#), [bowtie](#) - a man's tie that ties in a bow

[four-in-hand](#) - a long necktie that is tied in a slipknot with one end hanging in front of the other

[neckwear](#) - articles of clothing worn about the neck

[old school tie](#) - necktie indicating the school the wearer attended

[string tie](#) - a very narrow necktie usually tied in a bow

[Windsor tie](#) - a wide necktie worn in a loose bow



2. tie - a social or business relationship; "a valuable financial affiliation"; "he was sorry he had to sever his ties with other members of the team"; "many close associations with England"

[affiliation](#), [tie-up](#), [association](#)

[relationship](#) - a state involving mutual dealings between people or parties or countries



3. tie - equality of score in a contest

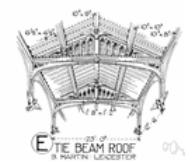
[equivalence](#), [par](#), [equality](#), [equation](#) - a state of being essentially equal or equivalent; equally balanced; "on a par with the best"

[deuce](#) - a tie in tennis or table tennis that requires winning two successive points to win the game

4. tie - a horizontal beam used to prevent two other structural members from spreading apart or separating; "he nailed the rafters together with a tie beam"

[tie beam](#)

[beam](#) - long thick piece of wood or metal or concrete, etc., used in construction



Word sense disambiguation

Analysis with definitions(s)

Bill Gates has developed an interest/[readiness to give attention] in language technology and yesterday aquired a 10 % interest/[a share (in a company, business, etc.)] in Torbjörn Lager 's sense disambiguation technology . Lager will retain a 90 % interest/[a share (in a company, business, etc.)] in the new company , which will be based in Göteborg , Sweden . Last year 's drop in interest/[money paid for the use of money] rates will probably be good for the company . Finally , although all this may sound like an arcane maneuver of little interest/[quality of causing attention to be given] outside Wall Street , it would set off an economical earthquake .

These are the six senses of the noun *interest* according to the LDOCE:

Sense	Definition
1	readiness to give attention
2	quality of causing attention to be given
3	activity, subject, etc., which one gives time and attention to
4	advantage, advancement, or favour
5	a share (in a company, business, etc.)
6	money paid for the use of money

Word sense disambiguation

becas Annotate Help API Widget About Contact

HIGHLIGHT

All None

Anatomy
 Disorders
 Chemicals
 Genes and Proteins
 Cellular Components
 Molecular Functions
 Biological Processes
 Ambiguous

In Duchenne muscular dystrophy (DMD), the infiltration of skeletal muscle by immune cells aggravates disease, yet the precise mechanisms behind these inflammatory responses remain poorly understood. Chemotactic cytokines, or chemokines, are considered essential recruiters of inflammatory cells to the tissues. We assayed chemokine and chemokine receptor expression in DMD muscle biopsies (n = 9, average age 7 years) using immunohistochemistry, immunofluorescence, and *in situ* hybridization. CXCL1, CXCL2, CXCL3, CXCL8, and CXCL11, absent from normal muscle fibers, were induced in DMD myofibers. CXCL11, CXCL12, and the ligand-receptor couple CCL2-CCR2 were upregulated on the blood vessel endothelium of DMD patients. CD68 (+) macrophages expressed high levels of CXCL8, CCL2, and CCL5. Our data suggest a possible beneficial role for CXCR1/2/4 ligands in managing muscle fiber damage control and tissue regeneration. Upregulation of endothelial chemokine receptors and CXCL8, CCL2, and CCL5 expression by cytotoxic macrophages may regulate myofiber necrosis.

[Load text](#) Annotated 46 concept occurrences in 0.179s. [Export ▾](#)

New to becas? [Take the tour »](#)

Concept Tree

+ Expand All - Collapse All

- + **Anatomy (12)**
 - **Disorders (4)**
 - + **DMD (1)**
 - **Muscular Dystrophy, Duchenne (4)**
 - + NCI:C75482
 - + NCI:0013264
 - + SNOMEDCT:76670001
 - + omim.org:302045

Semantic role labeling

- Extracting subject-predicate-object triples from a sentence



Semantic Role Labeling Demo

Input Text:

They had brandy in the library .

[Click For General Explanation of Argument Labels](#)

Output:

<input type="checkbox"/> SRL	<input type="checkbox"/> Nom	<input type="checkbox"/> Preposition	<input type="checkbox"/>
They	owner [A0]		
had	V: have.03		
brandy	possession [A1]		Governor
in			Locationin:1(1)
the	location [AM-LOC]		
library		Object	
.			

http://cogcomp.cs.illinois.edu/page/demo_view/srl

Outline

- Introduction to Language
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- Challenges
- NLP course

Phonetics and phonology

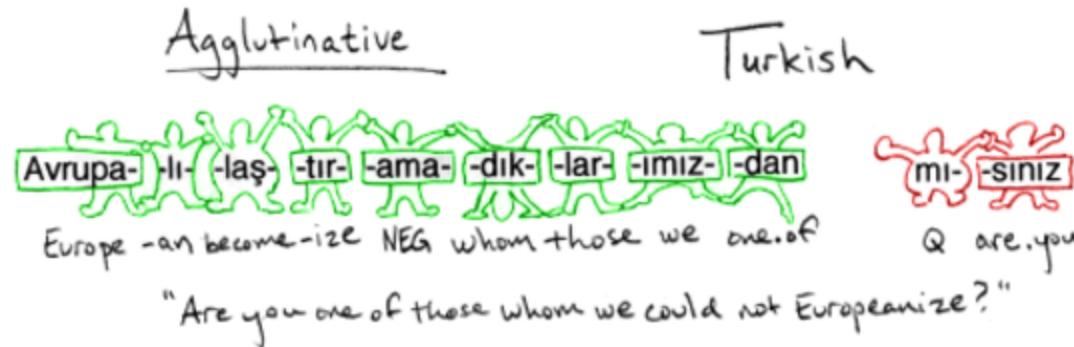
- The study of linguistic sounds and their relations to words

<http://german.about.com/library/blfunkabc.htm>

Das Funkalphabet - German Phonetic Spelling Code compared to the international ICAO/NATO code Listen to AUDIO for this chart! (below)		
Germany*	Phonetic Guide	ICAO/NATO**
A wie Anton	AHN-tone	Alfa/Alpha
Ä wie Ärger	AIR-gehr	(1)
B wie Berta	BARE-tuh	Bravo
C wie Cäsar	SAY-zar	Charlie
Ch wie Charlotte	shar-LOT-tuh	(1)
D wie Dora	DORE-uh	Delta
E wie Emil	ay-MEAL	Echo
F wie Friedrich	FREED-reech	Foxtrot
G wie Gustav	GOOS-tahf	Golf
H wie Heinrich	HINE-reech	Hotel
I wie Ida	EED-uh	India/Indigo
J wie Julius	YUL-ee-oos	Juliet
K wie Kaufmann	KOWF-mann	Kilo
L wie Ludwig	LOOD-vig	Lima
AUDIO 1 > Listen to mp3 for A-L		
M wie Martha	MAR-tuh	Mike
N wie Nordpol	NORT-pole	November
O wie Otto	AHT-toe	Oscar
Ö wie Ökonom (2)	UEH-ko-nome	(1)
P wie Paula	POW-luh	Papa
Q wie Quelle	KVEL-uh	Quebec
R wie Richard	REE-shart	Romeo
S wie Siegfried (3)	SEEG-freed	Sierra
Sch wie Schule	SHOO-luh	(1)
ß (Eszett)	ES-TSET	(1)
T wie Theodor	TAY-oh-dore	Tango
U wie Ulrich	OOL-reech	Uniform
Ü wie Übermut	UEH-ber-moot	(1)
V wie Viktor	VICK-tor	Victor
W wie Wilhelm	VIL-helm	Whiskey
X wie Xanthippe	KSAN-tipp-uh	X-Ray
Y wie Ypsilon	IPP-see-lohn	Yankee
Z wie Zeppelin	TSEP-puh-leen	Zulu

Morphology

- The study of internal structures of words and how they can be modified
- Parsing complex words into their components



Syntax

- The study of the structural relationships between words in a sentence

Parse

```
(ROOT
  (S
    (NP
      (NP (NNP Surgical) (NN resection) (NNS specimens))
      (PP (IN of)
        (NP
          (NP (CD 85) (JJ invasive) (JJ ductal) (NNS carcinomas))
          (PP (IN of)
            (NP
              (NP (CD 85) (NNS women))
              (SBAR
                (WHNP (WP who))
                (S
                  (VP (VBD had)
                  (VP (VBN undergone)
                    (NP (CD 3D) (NN ultrasound)))))))))))
      (VP (VBD were)
        (VP (VBN included))))
    (. .)))
```

Semantics

- The study of the meaning of words, and how these combine to form the meanings of sentences
 - Synonymy: fall & autumn
 - Hypernymy & hyponymy (is a): animal & dog
 - Meronymy (part of): finger & hand
 - Homonymy: fall (verb & season)
 - Antonymy: big & small

Pragmatics

- Social use of language
- The study of how language is used to accomplish goals, and the influence of context on meaning
- Understanding the aspects of a language which depends on situation and world knowledge

Give me the salt!

Could you please give me the salt?

Discourse

- The study of linguistic units larger than a single statement

John reads a book. **He** borrowed **it** from **his** friend.

Berlin (/bərˈlɪn/, German: [b̥eʁˈli:n] (listen)) is the **capital of Germany**, and one of the 16 **states of Germany**. With a population of 3.5 million people,^[4] Berlin is Germany's largest city. It is the second **most populous city proper** and the seventh **most populous urban area in the European Union**.^[5] Located in northeastern Germany on the banks of River **Spree**, it is the center of the **Berlin-Brandenburg Metropolitan Region**, which has about 6 million residents from over 180 nations.^{[6][7][8][9]} Due to its location in the **European Plain**, Berlin is influenced by a **temperate seasonal climate**. Around one third of the city's area is composed of forests, parks, gardens, rivers and lakes.^[10]

(<http://en.wikipedia.org/wiki/Berlin>)

Outline

- Introduction to Language
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- Challenges
- NLP course

Paraphrasing

- Different words/sentences express the same meaning
 - Season of the year
 - Fall
 - Autumn
 - Book delivery time
 - When will my book arrive?
 - When will I receive my book?

Ambiguity

- One word/sentence can have different meanings
 - Fall
 - The third season of the year
 - Moving down towards the ground or towards a lower position
 - The door is open.
 - Expressing a fact
 - A request to close the door

Phonetics and Phonology



Phonological ambiguities or Give peas a chance!

One of my favourite ways to have fun with communication are phonological ambiguities.

Phonological ambiguities are two or more words which sound the same and have different meanings.



intended to be heard.

Language can contain ambiguities - and more than one way to compose a set of sounds into words.

So listen to yourself: It is always good to notice a spoken sentence often contains many words which are (sometimes not)

English examples:

- there – their
- here - hear
- plane – plain
- Hamburger (Citizens of Hamburg) – hamburger (burger, food)
- sea - see
- Friday - fry day
- weekend - weak end
- ice cream - I scream.
- new direction - nude erection
- new day - nude, eh?
- I don't know! - I don't - no!
- but – butt
- Wait - Weight
- psychotherapist - psycho the rapist
- You're unconscious now... - Your unconscious now...
- Your students... - You're students...
- Two - too - to

German examples:

- Du hast Gewehre. (You have got guns.) - Du hasst Gewehre. (You hate guns.)
- Lehrer (teacher) - leerer (emptier)

Syntax and ambiguity

- I saw the man with a telescope.
 - Who had the telescope?



(<http://www.realtytrac.com/landing/2009-year-end-foreclosure-report.html>)

Semantics

- The astronomer loves the **star**.
 - Star in the sky
 - Celebrity



(<http://en.wikipedia.org/wiki/Star#/media/File:Starsinthesky.jpg>)



(<http://www.businessnewsdaily.com/2023-celebrity-hiring.html>)

Discourse analysis

- Alice understands that you like your mother, but **she** ...
 - Does **she** refer to Alice or your mother?

Outline

- Introduction to Language
- NLP Applications
- NLP Techniques
- Linguistic Knowledge
- Challenges
- NLP course

NLP Course

- Home page:
 - <http://hpi.de/plattner/teaching/summer-term-2016/natural-language-processing.html>
- Lecture
 - Monday 13:30-15:00
 - HS3
 - 3 credit points

Grading

- 60% Project
- 40% Final exam (written)

Program

(Program is subject to change)

Week	Date	Topic
1	April 11, 2016	Introduction to Natural Language Processing
2	April 18, 2016	Regular Expressions and Automata
3	April 25, 2016	N-Grams
4	May 2, 2016	Part-of-Speech Tagging
5	May 9, 2016	Syntactic Parsing
6	May 16, 2016	(Pfingstmontag - no lecture)
7	May 23, 2016	Lexical Semantics
8	May 30, 2016	Discourse
9	June 6, 2016	Information Extraction
10	June 13, 2016	Text Classification and Sentiment Analysis
11	June 20, 2016	Information retrieval
12	June 27, 2016	Question Answering and Summarization
13	July 4, 2016	Machine Translation
14	July 11, 2016	(project presentation)
15	July 18, 2016	Final exam (HS 3, 13:15)

Project

- Development of a NLP application
 - Information Retrieval
 - Information Extraction
 - Text Summarization
 - Question Answering
 - Sentiment Analysis
 - Machine Translation
 - Etc..

Project

- The application should include following components:
 - Part-of-speech tagging
 - Syntactic parsing
 - Lexical semantics
 - Discourse analysis
 - Named-entity recognition

Project

- Any NLP or ML libraries
 - Stanford Core NLP
 - NLTK
 - Apache OpenNLP
 - GATE
 - SAP HANA (contact me)
 - R
 - Weka

Project

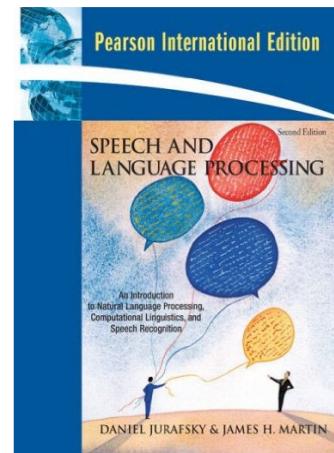
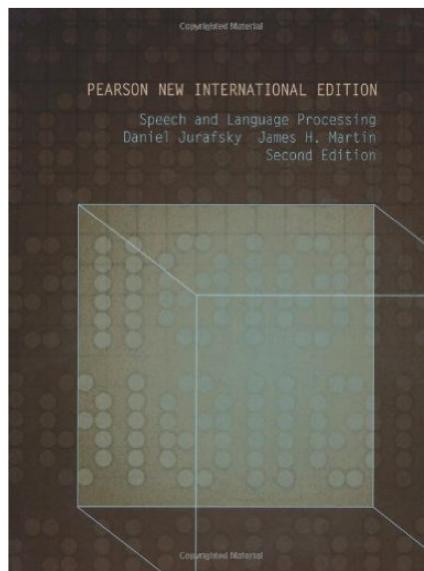
- Any language
 - English, German, etc.
 - Check available NLP tools
- Any text collections
 - Social media, Web pages, publications, Wikipedia, etc.
 - Benchmarks or new collections
- Any domain

Project

- Teams (2-3 students)
- Send me an email with your proposal as soon as possible
- Updates (presentations) on the progress of the project
 - Slots during the lectures
 - Also considered for grading

Course book

- Speech and Language Processing
 - Daniel Jurafsky and James H. Martin



Standort: [Bereichsbibliothek Babelsberg --> Wegweiser](#)

ST 306 JUR

Ausleihbar - LBS

Verfuegbar: BB Babelsberg / LBS.

Standort: [Bereichsbibliothek Babelsberg --> Wegweiser](#)

ST 306 JUR

Ausleihbar - LBS

Verfuegbar: BB Babelsberg / LBS.

Standort: [Bereichsbibliothek Babelsberg --> Wegweiser](#)

ST 306 JUR

Ausleihbar

verfuegbar → [Bestellen](#)

Standort: [Bereichsbibliothek Babelsberg --> Wegweiser](#)

ST 306 JUR

Ausleihbar - LBS

Verfuegbar: BB Babelsberg / LBS.

Standort: [Bereichsbibliothek Babelsberg --> Wegweiser](#)

ST 306 JUR

Ausleihbar - LBS

Verfuegbar: BB Babelsberg / LBS.

Standort: [Bereichsbibliothek Babelsberg --> Wegweiser](#)

ST 306 JUR

Ausleihbar - LBS

Verfuegbar: BB Babelsberg / LBS.

Standort: [Bereichsbibliothek Babelsberg --> Wegweiser](#)

ST 306 JUR

Praesenzbestand

Verfuegbar: BB Babelsberg / Präsenz.

Journal and conferences

- Journal
 - Computational Linguistics
- Conferences
 - ACL: Association for Computational Linguistics (**ACL'16 in Berlin!**)
 - NAACL: North American Chapter
 - EACL: European Chapter
 - HLT: Human Language Technology
 - EMNLP: Empirical Methods on Natural Language Processing
 - CoLing: Computational Linguistics
 - LREC: Language Resources and Evaluation

NLP Course

- Contact
 - Mariana.Neves@hpi.uni-potsdam.de
 - Room 0.01 (Villa), appointment under request
- We have a student position for NLP at the EPIC chair!