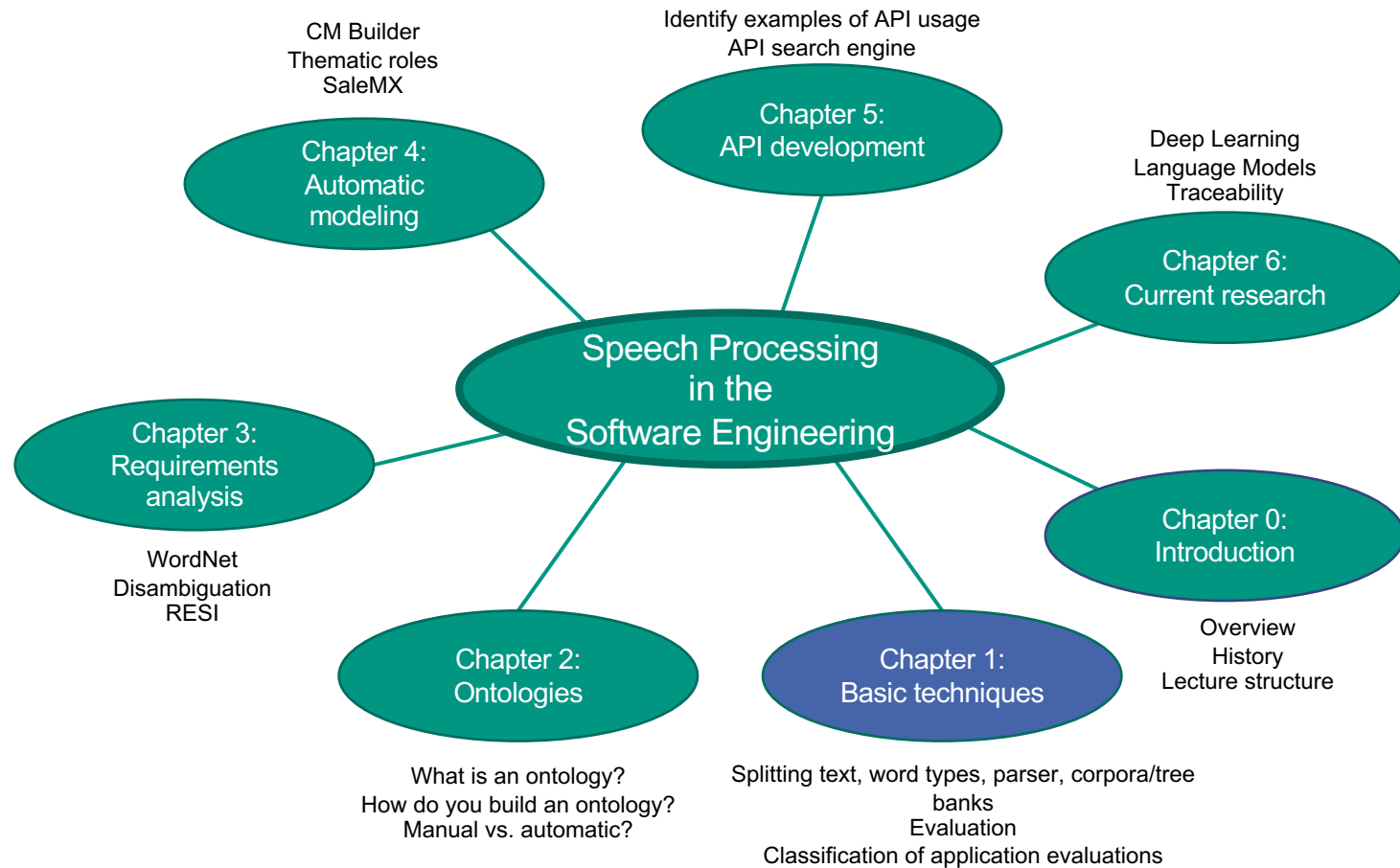


# Language processing in software engineering

## 1.1 Corpora

Walter Tichy





# Idea of corpus linguistics

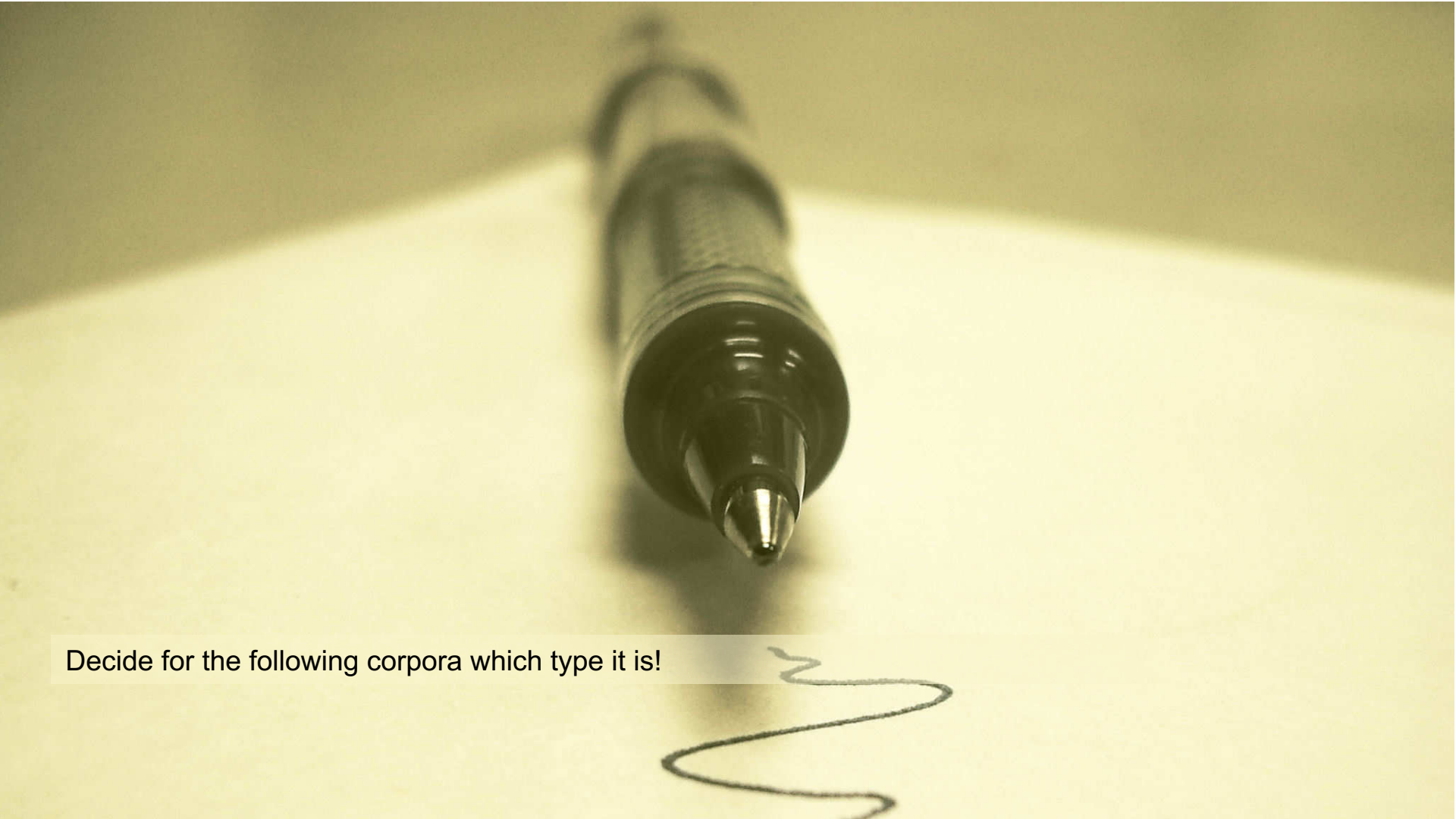
- How can the material of a language be studied empirically?
- Corpus - word origin
  - From Latin corpus, corporis, neuter. Plural: corpora.
  - Therefore: the corpus / corpora
- The corpus of a language is
  - the totality of all linguistic phenomena,
  - in the **statistical sense, the population**
- Since the population is too large, a selection must be made
- A selection corpus is the **empirical basis of corpus linguistics**
  - Corpora of spoken linguistic phenomena (speech corpus)
  - corpora of written linguistic phenomena (**text corpus**)

# The representative corpus - The selection corpus

- What does "representative" mean?
- One needs criteria for the **representativeness of** a selection from the population
  - Study of the language of an author
  - Study of a historical language
  - Study of the language or conversational behavior of a social group
  - Study of the language at a given time
- The probability of occurrence of new word forms decreases with text length
- As the text length increases, new words/meanings appear more and more

# Typology of text corpora

- Total corpus of a language (e.g. Thesaurus Linguae Graecae)
- Author corpus (e.g. Kant corpus)
- **Selection corpus** (e.g. Brown corpus)
- **Monitor corpus** are used to detect new phenomena (words, structures) and to optimize statistical analysis procedures (see e.g. Teubert1998).



Decide for the following corpora which type it is!

	Selection	Total	Authors	Monitor
Taylor-Swift-Corpus			✗	
Corpus for experimentation (during your Master's thesis 😊 )				✗
Google N-Gram-Coprus	✗			✗
Wallstreet Journal Corpus	✗			

Decide for the following corpora which type it is!

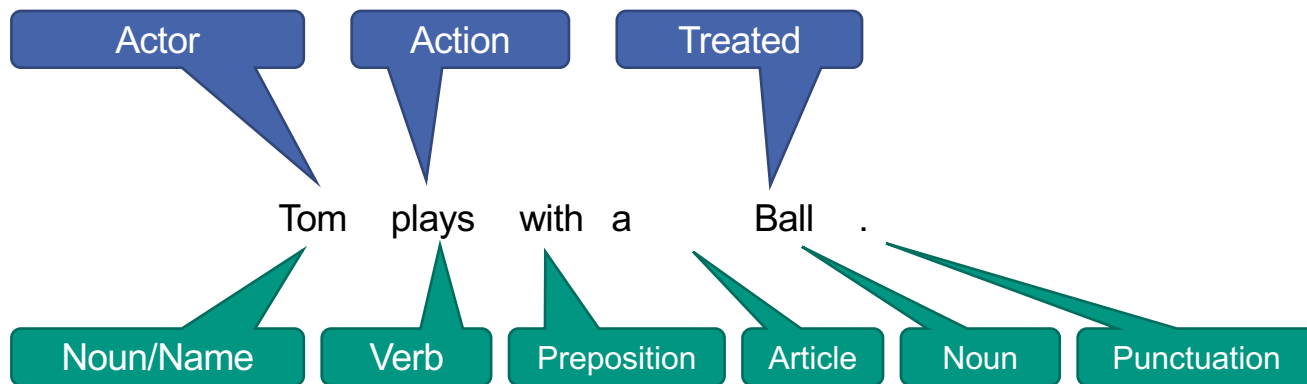
# Typology of corpora after editing

- Unedited corpora: Pure text form
- Edited corpora: texts with labels
  - Phonetic, phonological
  - Morphological
  - Word types (part-of-speech labels)
  - Syntactic (Treebanks)
  - Semantic
  - Text-critical



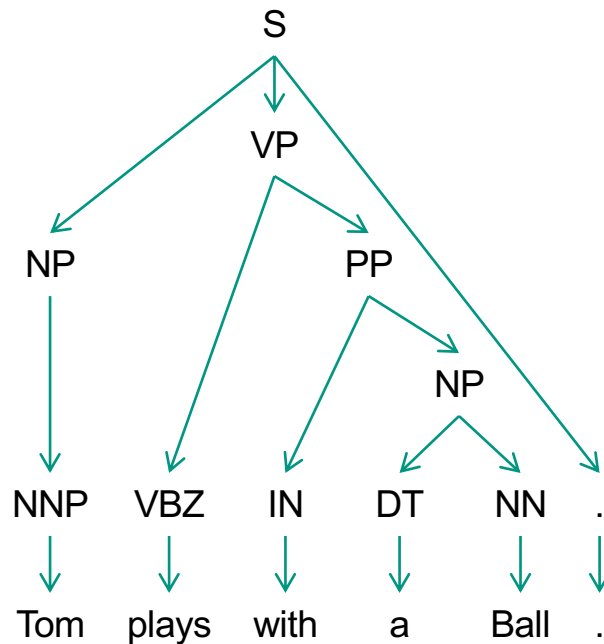
# Example:

## Sentence with part-of-speech labels and semantics



# Example: sentence with syntax tree

S	sentence
NP	noun phrase
NNP	Noun
VP	verb phrase
VBZ	Verb
PP	prepositional phrase
IN	preposition
DT	article/determiner
NN	Noun
.	point



# Use of corpora

- Descriptive phonology  
e.g. determination of the phoneme system of dialects
- Descriptive morphology  
e.g. determination of morphemes and word formation rules
- Lexicography  
e.g. determination and documentation of vocabulary = creation of dictionaries
- Descriptive syntax  
e.g. identifying the types of sentence patterns, phrase patterns
- Descriptive semantics  
e.g. determination of word sets (synonyms with same meaning, *synsets*)
- **Linguistic Technology**  
Development and optimization of stochastic analysis methods

See chapter 3

# Language processing in software engineering

## 1.1.1 Overview of available corpora

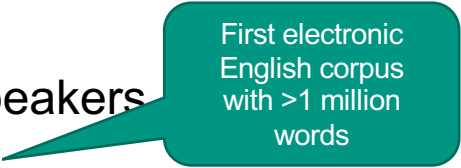
Walter Tichy



# The Brown Corpus

## Overview

- Language: US English
- Medium: written
- Type: representative selection corpus
- Time: 1961
- Language level: Standard language, native US speakers
- Total volume: approx. 1 million words
- Number of words/text: approx. 2000
- Number of texts: 500
- Ling. Description: categorized (15 genres), word types labeled.
- Uses: lexicography, syntax research, statistics, vocabulary research.



First electronic  
English corpus  
with >1 million  
words

# The Brown corpus

## Examples of the genres

ID	File	Genre	Description
A16	ca16	news	Chicago Tribune: Society Reportage
B02	cb02	editorial	Christian Science Monitor: Editorials
C17	cc17	reviews	Time Magazine: Reviews
D12	cd12	religion	Underwood: Probing the Ethics of Realtors
E36	ce36	hobbies	Norling: Renting a Car in Europe
F25	cf25	lore	Boroff: Jewish Teenage Culture
G22	cg22	belles_lettres	Reiner: Coping with Runaway Technology
H15	ch15	government	US Office of Civil and Defense Mobilization: The Family Fallout Shelter
J17	cj19	learned	Mosteller: Probability with Statistical Applications
K04	ck04	fiction	W.E.B. Du Bois: Worlds of Color
L13	cl13	mystery	Hitchens: Footsteps in the Night
M01	cm01	science_fiction	Heinlein: Stranger in a Strange Land
N14	cn15	adventure	Field: Rattlesnake Ridge
P12	cp12	romance	Callaghan: A Passion in Rome
R06	cr06	humor	Thurber: The Future, If Any, of Comedy

# The Brown Corpus

## Text origin

The Corpus consists of 500 samples, distributed **across 15 genres** in rough proportion to the amount published in 1961 in each of those genres. All works sampled were **published in 1961**; as far as could be determined they were *first* published then, and were **written by native speakers of American English**.

Each sample began at a random sentence-boundary in the article or other unit chosen, and continued up to the first sentence boundary after 2,000 words. In a very few cases miscounts led to samples being just under 2,000 words.

The original data entry was done on upper-case only keypunch machines; capitals were indicated by a preceding asterisk, and various special items such as formulae also had special codes.

The corpus originally (1961) contained 1,014,312 words sampled from 15 text categories:

- A. PRESS: Reportage (*44 texts*): Political, Sports, Society, Spot News, Financial, Cultural
- B. PRESS: Editorial (*27 texts*): Institutional Daily, Personal, Letters to the Editor
- C. PRESS: Reviews (*17 texts*): *theatre, books, music, dance*
- D. RELIGION (*17 texts*): Books, Periodicals, Tracts
- E. SKILL AND HOBBIES (*36 texts*): Books, Periodicals
- F. POPULAR LORE (*48 texts*): Books, Periodicals
- G. BELLES-LETTRES - Biography, Memoirs, etc. (*75 texts*): Books, Periodicals
- H. MISCELLANEOUS: US Government & House Organs (*30 texts*): Government Documents, Foundation Reports, Industry Reports, College Catalog, Industry House organ.
- J. LEARNED (*80 texts*): Natural Sciences, Medicine, Mathematics, Social and Behavioral Sciences, Political Science, Law, Education, Humanities, Technology and Engineering
- K. FICTION: General (*29 texts*): Novels, Short Stories
- L. FICTION: Mystery and Detective Fiction (*24 texts*): Novels, Short Stories
- M. FICTION: Science (*6 texts*): Novels, Short Stories
- N. FICTION: Adventure and Western (*29 texts*): Novels, Short Stories
- P. FICTION: Romance and Love Story (*29 texts*): Novels, Short Stories
- R. HUMOR (*9 texts*): Novels, Essays, etc.

Excerpt from the English Wikipedia article [Brown Corpus](#)

# The Brown corpus in the NLTK

```
>>> import nltk
>>> from nltk.corpus import brown
>>> cfd = nltk.ConditionalFreqDist(
    (genre, word)
    for genre in brown.categories()
    for word in brown.words(categories=genre))
>>> genres = ['news', 'religion', 'hobbies',
    'science_fiction', 'romance', 'humor']
>>> days = ['Monday', 'Tuesday', 'Wednesday', 'Thursday',
    'Friday', 'Saturday', 'Sunday']
```

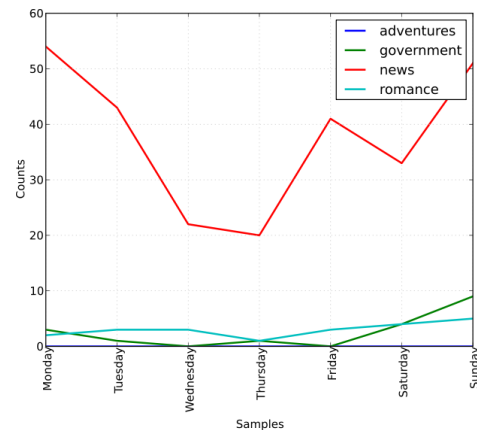


# The Brown corpus in the NLTK

```
>>> cfd.tabulate(samples=days, conditions=['adventures',  
      'government', 'news', 'romance'])
```

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
adventures	1	0	0	0	0	0	0
government	3	1	0	1	0	4	9
news	54	43	22	20	41	33	51
romance	2	3	3	1	3	4	5

```
>>> cfd.plot(samples=days,  
      conditions=['adventures',  
      'government', 'news', 'romance'])
```



# The Penn Treebank

- Language: English
- Medium: written
- Type: representative selection corpus
- Time: 1989-1996
- Language level: Standard language
- Ling. Description: part-of-speech labeled, syntax trees, sem. Predicate-argument structures, transcriptions
- Total volume: approx. 7 million words
  - 7 million with part-of-speech labels
  - 3 million parsed (syntax trees)
  - 1.6 million transcribed speech inputs
- Number of words/text: approx. 2000
- Number of texts: 500
- Encoding: SGML (TEI)
- Uses: lexicography, statistics, speech recognition, CL applications.

(Taylor2003)

# The Penn Treebank - POS Tag Set

CC	Coordinating conj.	TO	infinitival <i>to</i>
CD	Cardinal number	UH	Interjection
DT	Determiner	VB	Verb, base form
EX	Existential there	VBD	Verb, past tense
FW	Foreign word	VBG	Verb, gerund/present pple
IN	Preposition	VBN	Verb, past participle
JJ	Adjective	VBP	Verb, non-3rd ps. sg. present
JJR	Adjective, comparative	VBZ	Verb, 3rd ps. sg. present
JJS	Adjective, superlative	WDT	Wh-determiner
LS	List item marker	WP	Wh-pronoun
MD	Modal	WP\$	Possessive <i>wh</i> -pronoun
NN	Noun, singular or mass	WRB	Wh-adverb
NNS	Noun, plural	#	Pound sign
NNP	Proper noun, singular	\$	Dollar sign
NNPS	Proper noun, plural	.	Sentence-final punctuation
PDT	Predeterminer	,	Comma
POS	Possessive ending	:	Colon, semi-colon
PRP	Personal pronoun	(	Left bracket character
PP\$	Possessive pronoun	)	Right bracket character
RB	Adverb	"	Straight double quote
RBR	Adverb, comparative	'	Left open single quote
RBS	Adverb, superlative	“	Left open double quote
RP	Particle	,	Right close single quote
SYM	Symbol	”	Right close double quote

(Taylor2003)

# The Penn Treebank

## POS tags (manual annotations)

### *Output of tagger*

Battle-tested/NNP Japanese/NNP industrial/JJ managers/NNS  
here/RB always/RB buck/VB up/IN nervous/JJ newcomers/NNS  
with/IN the/DT tale/NN of/IN the/DT first/JJ of/IN  
their/PP\$ countrymen/NNS to/TO visit/VB Mexico/NNP ,/,  
a/DT boatload/NN of/IN samurai/NNS warriors/NNS blown/VBN  
ashore/RB 375/CD years/NNS ago/RB ./.

### *Hand-corrected by annotator*

Battle-tested/NNP\*/JJ Japanese/NNP\*/JJ industrial/JJ  
managers/NNS here/RB always/RB buck/VB\*/VBP up/IN\*/RP  
nervous/JJ newcomers/NNS with/IN the/DT tale/NN of/IN  
the/DT first/JJ of/IN their/PP\$ countrymen/NNS to/TO  
visit/VB Mexico/NNP ,/, a/DT boatload/NN of/IN  
samurai/NNS\*/FW warriors/NNS blown/VBN ashore/RB 375/CD  
years/NNS ago/RB ./.

### *Final version*

Battle-tested/JJ Japanese/JJ industrial/JJ managers/NNS  
here/RB always/RB buck/VBP up/RP nervous/JJ newcomers/NNS  
with/IN the/DT tale/NN of/IN the/DT first/JJ of/IN  
their/PP\$ countrymen/NNS to/TO visit/VB Mexico/NNP ,/,  
a/DT boatload/NN of/IN samurai/FW warriors/NNS blown/VBN  
ashore/RB 375/CD years/NNS ago/RB ./.

(Taylor2003)

# The Penn Treebank - syntactic tagset

---

ADJP	Adjective phrase
ADVP	Adverb phrase
NP	Noun phrase
PP	Prepositional phrase
S	Simple declarative clause
SBAR	Subordinate clause
SBARQ	Direct question introduced by <i>wh</i> -element
SINV	Declarative sentence with subject-aux inversion
SQ	Yes/no questions and subconstituent of SBARQ excluding <i>wh</i> -element
VP	Verb phrase
WHADVP	Wh-adverb phrase
WHNP	Wh-noun phrase
WHPP	Wh-prepositional phrase
X	Constituent of unknown or uncertain category
*	“Understood” subject of infinitive or imperative
0	Zero variant of <i>that</i> in subordinate clauses
T	Trace of <i>wh</i> -Constituent

---

(Taylor2003)

# The Penn Treebank

## syntactic tags (example)

```
( (S
  (NP Martin Marietta Corp.)
  was
  (VP given
    (NP a
      $ 29.9
      million Air Force contract
      (PP for
        (NP low-altitude navigation
          and
          targeting equipment))))))
.)
```

(Taylor2003)

# The Wallstreet Journal (WSJ) Corpus

- Language: US English
- Medium: written
- Type: representative selection corpus
- Time: 1987-89
- Language level: Standard language
- Total volume: approx. 30 million words
- Ling. Description: POS labeled, syntax trees
- Usage: lexicography, **statistics**, CL applications

# The British National Corpus (BNC)

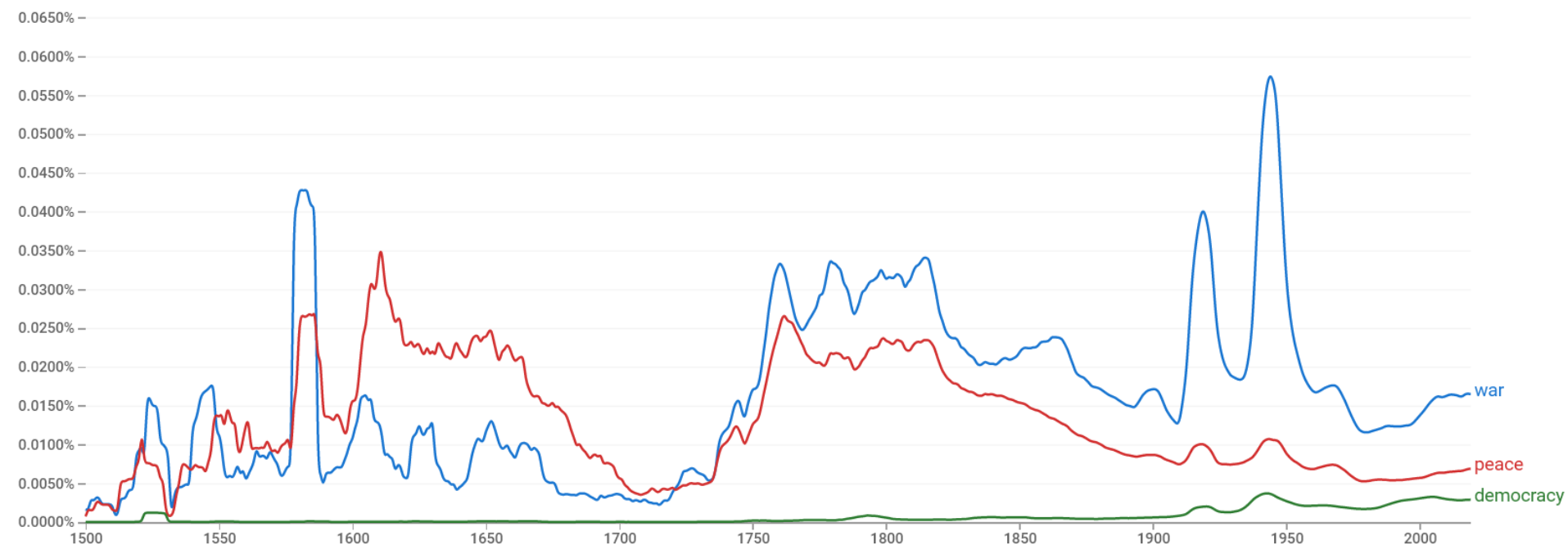
- Language: GB-English
- Medium: written and spoken
- Type: representative selection corpus
- Time: 1991-1994
- Language level: Standard language
- Total words: 1 million
- Number of words/text: 2000
- Number of texts: 500
- Encoding: SGML (TEI)
- Ling. Description: POS labeled
- Usage: lexicography, statistics, speech recognition



# The Google Books Corpus

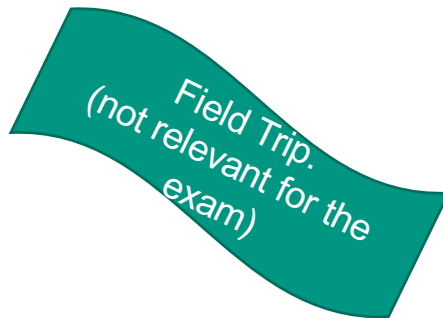
- Language: Various (US, GB, ES, DE, RU, CN, ...)
- Medium: written
- Type: extensive selection corpus
- Time: 1500-2019
- Language level: Standard language
- Total: All: > 500 trillion words, English: > 400 billion.
- Number of books: 15 million
- Processing: OCR (!)
- Ling. Description: automatically (!) annotated word types, etc.
- Usage: lexicography, statistics (*Ngram viewer*)

# Google Books Ngram Viewer



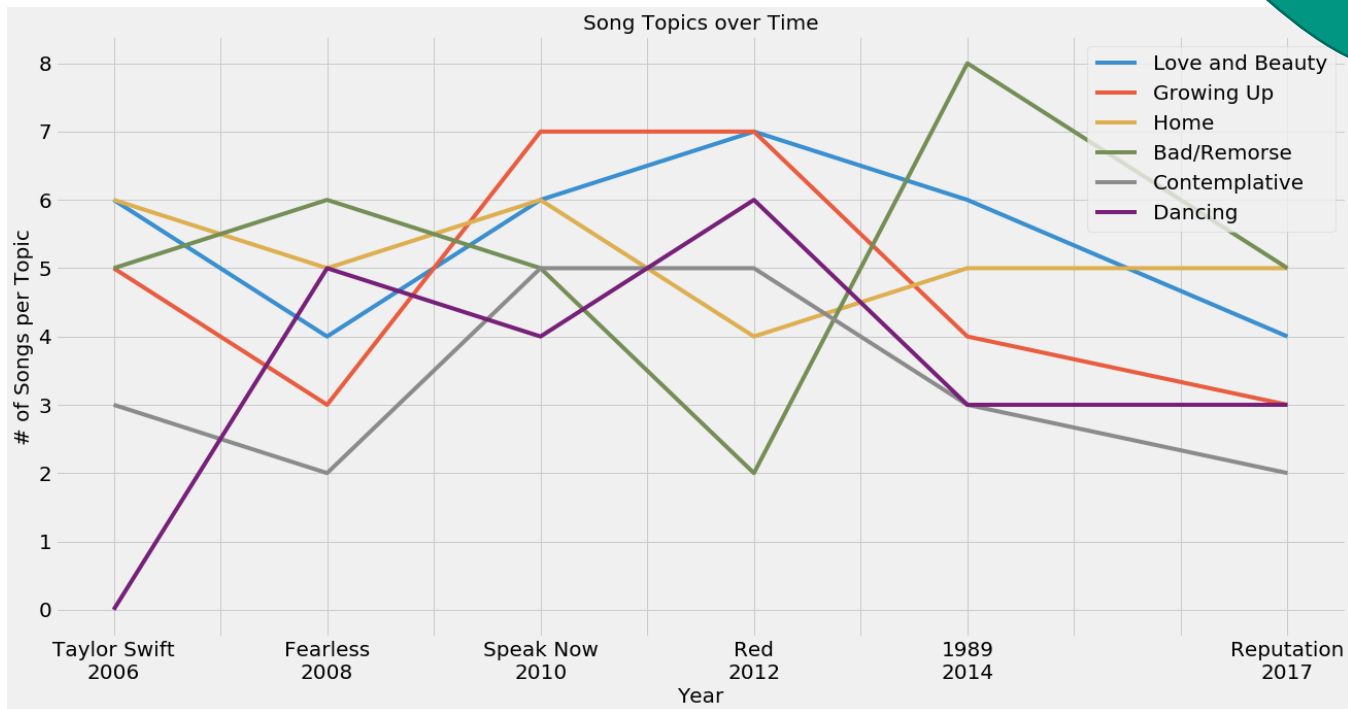
# The Taylor Swift corpus

- Language: US English
- Medium: written
- Type: Author corpus
- Time: 2006 - 2017
- Language level: "Standard language" (poetry)
- Total: 4862 song lines
- Ling. Description: none
- Usage: Tutorial for computational linguistics

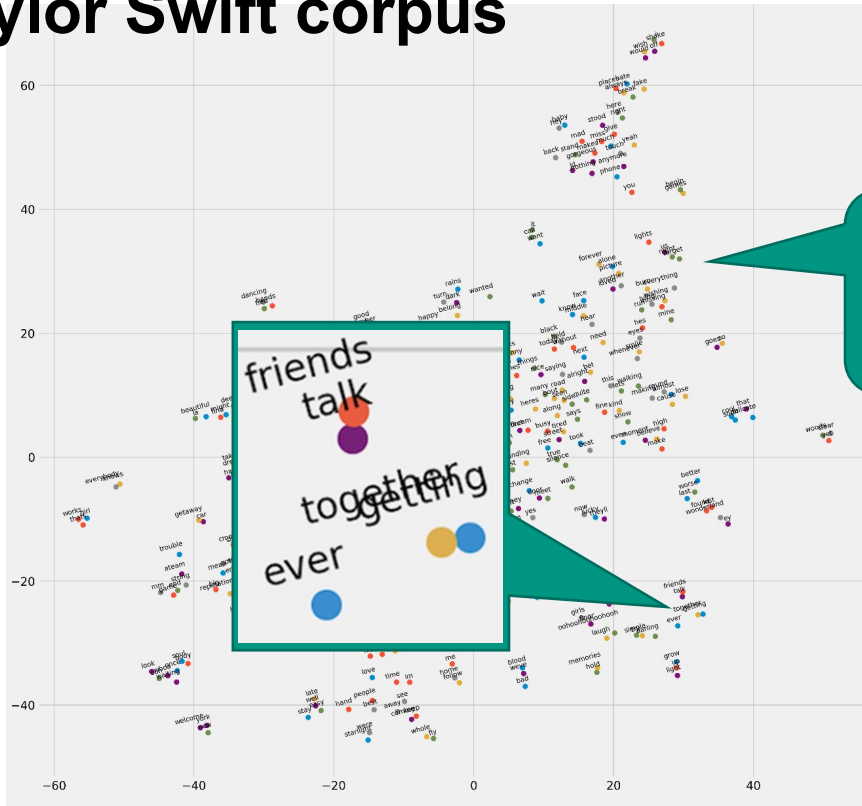


# The Taylor Swift corpus

Field Trip.  
(not relevant for the  
exam)



# The Taylor Swift corpus



Field Trip.  
(not relevant for the  
exam)

Word vectors:  
Describe words  
based on their  
environment

Read more:

<https://news.codecademy.com/taylor-swift-lyrics-machine-learning/>

# References

- **(Bird2009)** Bird, S. ; Klein, E. & Loper E. : Natural language processing with Python : [analyzing text with the natural language toolkit]. O'Reilly, 2009. chapter 1.5
- **(Carstensen2010)** Carstensen, K.-U.; Ebert, C.; Jekat, S. J.; Klabunde, R. & Langer, H. (Eds.) : Computational linguistics and language technology - An introduction. Spektrum Akademischer Verlag, 2010. [10.1007/978-3-8274-2224-8](#). chapter 1.
- **(Jurafsky2009)** Jurafsky, D. & Martin, J. H. : Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, Pearson Education International, 2009. chapter 1.5, 1.6.
- **(Kof2005)** Kof, L. : Montoyo, A.; Muñoz, R. & Métais, E. (Eds.) : Natural language processing: mature enough for requirements documents analysis? NLDB, Springer, 2005, 3513, 91-102. [10.1007/11428817\\_9](#).
- **(Mich2004)** Mich, L.; Franch, M. & Inverardi, P. N. : Market research for requirements analysis using linguistic tools. Requirements Engineering, Springer, 2004, 9, 40-56. [10.1007/s00766-003-0179-8](#).

# Sources

■ These slides are based in part on the lectures, papers, and/or slides of

■ [Natural Language Understanding](#)

by [Dan Jurafsky](#)

Stanford University

Lecture 1

■ [Fundamentals of Computational Linguistics 1+2](#)

by [Winfried Lenders](#)

University of Bonn

Lectures 1 and 2

■ [Natural Language Engineering](#)

by [Mary M<sup>c</sup> Gee Wood](#)

The University of Manchester

Lecture 1.1



D. Jurafsky



W. Lenders



M. M<sup>c</sup> Gee Wood