Chapter NLP:II

II. Corpus Linguistics

- Empirical Research
- Text Corpora
- Text Statistics
- Text Statistics in IR
- Annotation

Empirical Research

1. Quantitative research based on numbers and statistics.
2. Studies phenomena and research qestions by analyzing data.
3. Derives knowledge from experience rather than from theory or belief.

Empirical Research

1. Quantitative research based on numbers and statistics.
2. Studies phenomena and research qestions by analyzing data.
3. Derives knowledge from experience rather than from theory or belief.

Quantitative versus qualitative research:

- Quantitative. Characterized by objective measurements.
- Qualitative. Emphasizes the understanding of human experience.

Empirical Research

1. Quantitative research based on numbers and statistics.
2. Studies phenomena and research qestions by analyzing data.
3. Derives knowledge from experience rather than from theory or belief.

Quantitative versus qualitative research:

- Quantitative. Characterized by objective measurements.
- Qualitative. Emphasizes the understanding of human experience.

Descriptive versus inferential statistics:

- Descriptive. Procedures for summarizing and comprehending a sample or distribution of values. Used to describe phenomena.

  1 2 2 2 $\rightarrow$ mean M = 1.75

- Inferential. Procedures that help draw conclusions based on values. Used to generalize inferences beyond a given sample.

  The average number is significantly greater than 1.

Empirical Research

Research Questions

What is a good research question? [Bartos 1992]

- Asks about the relationship between two or more variables.
- Is testable (i.e., it is possible to collect data to answer the question).
- Is stated clearly and in the form of a question.
- Does not pose an ethical or moral problem for implementation.
- Is specific and restricted in scope.
- Identifies exactly what is to be solved.

Empirical Research

Research Questions

What is a good research question? [Bartos 1992]

- Asks about the relationship between two or more variables.
- Is testable (i.e., it is possible to collect data to answer the question).
- Is stated clearly and in the form of a question.
- Does not pose an ethical or moral problem for implementation.
- Is specific and restricted in scope.
- Identifies exactly what is to be solved.

Example of a poorly formulated question:

> *"What is the effectiveness of parent education when given problem children?"*

Empirical Research

Research Questions

What is a good research question? [Bartos 1992]

- Asks about the relationship between two or more variables.
- Is testable (i.e., it is possible to collect data to answer the question).
- Is stated clearly and in the form of a question.
- Does not pose an ethical or moral problem for implementation.
- Is specific and restricted in scope.
- Identifies exactly what is to be solved.

Example of a poorly formulated question:

> *"What is the effectiveness of parent education when given problem children?"*

Example of a well-formulated question:

> *"What is the effect of the STEP parenting program on the ability of parents to use natural, logical consequences (as opposed to punishment) with their child who has been diagnosed with bipolar disorder?"*

Empirical Research

Empirical Research in NLP

- Corpus linguistics.
  NLP is studied in a corpus-linguistics manner; i.e., approaches are developed and evaluated on collections of text.
- Evaluation measures.
  An evaluation of the quality of an approach is important, especially of its effectiveness.
- Experiments.
  The quality of an approach is empirically evaluated on test corpora and compared to alternative approaches.
- Hypothesis testing.
  Methods which verify whether results of an experiment are meaningful/valid by estimating the odds that the results happened by chance.

Empirical Research

Empirical Research in NLP

- Corpus linguistics.
  NLP is studied in a corpus-linguistics manner; i.e., approaches are developed and evaluated on collections of text.

- Evaluation measures.
  An evaluation of the quality of an approach is important, especially of its effectiveness.

- Experiments.
  The quality of an approach is empirically evaluated on test corpora and compared to alternative approaches.

- Hypothesis testing.
  Methods which verify whether results of an experiment are meaningful/valid by estimating the odds that the results happened by chance.

Text Corpora

Corpus Linguistics

- The study of language as expressed in principled collections of natural language texts, called text corpora.
- Aims to derive knowledge and rules from real-world text.
- Covers both manual and automatic analysis of text.

Text Corpora

Corpus Linguistics

- The study of language as expressed in principled collections of natural language texts, called text corpora.
- Aims to derive knowledge and rules from real-world text.
- Covers both manual and automatic analysis of text.

Three main techniques:

1. Analysis. Developing and evaluating methods based on a corpus.
2. Annotation. Coding data with categories to facilitate data-driven research.
3. Abstraction. Mapping of annotated texts to a theory-based model.

➜ Need for text corpora: Without a corpus, it's hard to develop a strong approach—and impossible to reliably evaluate it.

"It's often not the one who has the best algorithm that wins.
It's who has the most data."

Text Corpora

Definition 1 (Text Corpus [Butler 2004]**)**

A text corpus is (an electronically stored) collection of data designed with according to specific corpus design criteria to be maximally representative of (a particular variety of) language or other semiotic systems.

The basic unit for representing text is typically a word (captures meaning).

Examples:

- 200,000 product reviews for sentiment analysis
- 1,000 news articles for part-of-speech tagging

Corpora in NLP:

- NLP approaches are developed and evaluated on text corpora.
- Usually, the corpora contain annotations of the output information type to be inferred.

Text Corpora

On Representativeness

- *"extent to which a sample includes the full range of variability in a population"* [Biber 1993]

  Here: Sample is our corpus, population is all of the language variety.

- *"A corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its contents can be generalized to the said language variety."* [Leech 1991]

  Question: If we find certain features in the corpus, are we likely to find the same features in further data of that type?

- But—what is representative to the users of language?

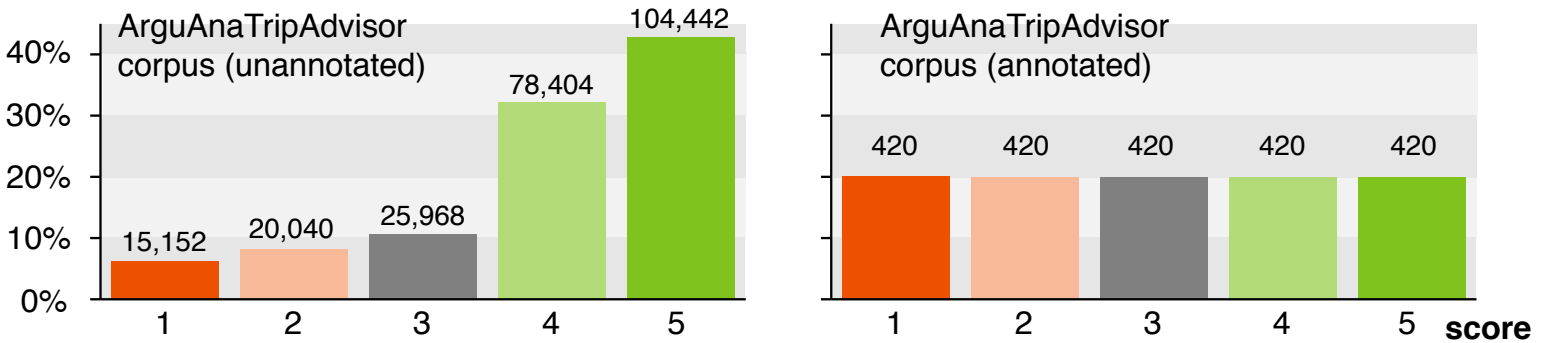
  "According to claims, the most likely document that an ordinary English citizen will cast his or her eyes over is The Sun newspaper" [Sinclair 2005]

  Keyword: reception versus production

- Corpus representativeness is important for generalization, since the corpus governs what can be learned about a given domain.

Text Corpora

Representative Data versus Balanced Data



- A corpus is representative for some output information type C, if it includes the full range of variability of texts with respect to C.
- The distribution of texts over the values of C should be representative for the real distribution.
- Balance with respect to a feature means that no value/level of the feature dominates; equally distributed with respect to a feature (e.g. genre, category of linguistic phenomena).
- A balanced distribution, where all values are evenly represented, may be favorable (particularly for machine learning).

Text Corpora

Character Encoding [detailed in WT:III-163 ff.]

- Character encoding is a mapping between bits and *code points*, where each code point is associated with a glyph.
  - Getting from bits in a file to characters on a screen.
  - Can be a major source of incompatibility.
- Charset for English: ASCII
  - Encodes 128 letters, numbers, special characters, and control characters in 7 bits, extended with an extra bit for storage in bytes.
- Charset for other European: Latin-1 (ISO-8859-1)

Text Corpora

Character Encoding [detailed in WT:III-163 ff.]

- Character encoding is a mapping between bits and *code points*, where each code point is associated with a glyph.
  - Getting from bits in a file to characters on a screen.
  - Can be a major source of incompatibility.
- Charset for English: ASCII
  - Encodes 128 letters, numbers, special characters, and control characters in 7 bits, extended with an extra bit for storage in bytes.
- Charset for other European: Latin-1 (ISO-8859-1)

"Even when documents say they are in ASCII or ISO 8859-1, you have to assume that they are lying, because it's extremely common for such documents to be actually encoded in Windows-1252."

[David Hawking]

Text Corpora

Character Encoding (continued) [detailed in WT:III-163 ff.]

- Other languages can have many more glyphs:
  - Chinese has more than 40,000 characters, with over 3,000 in common use.
- Many languages have multiple encoding schemes:
  - the CJK (Chinese-Japanese-Korean) family of East Asian languages, Hindi, Arabic
  - must specify encoding, cannot have multiple languages in one file

➜ Solution: Unicode

Text Corpora

Character Encoding (continued) [detailed in WT:III-163 ff.]

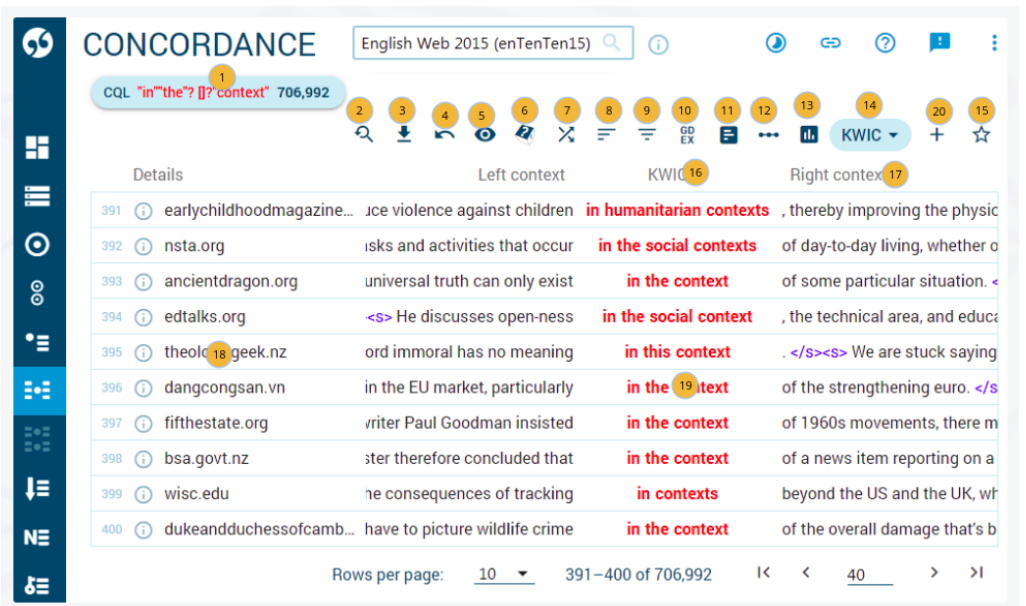
Unicode:

- All-encompassing charset and encoding for most writing systems.
- Allows for multiple languages in one file.
- Tailored encoding schemes to translate code points to a byte representation:
  - UTF-8 uses one byte for English (ASCII), and as many as 4 bytes for some traditional Chinese characters (variable length encoding).
  - UTF-16 uses 2 or 4 bytes for every character.
  - UTF-32 uses 4 bytes for every character.
- Applications may use UTF-32 for internal encoding (fast random lookup) and UTF-8 for disk storage (less space).

Text Corpora

Research in Language Use

Concordance: (alphabetical) list of principal words (or phrases) used in a book (nowadays: corpus) listing every instance of each with immediate context

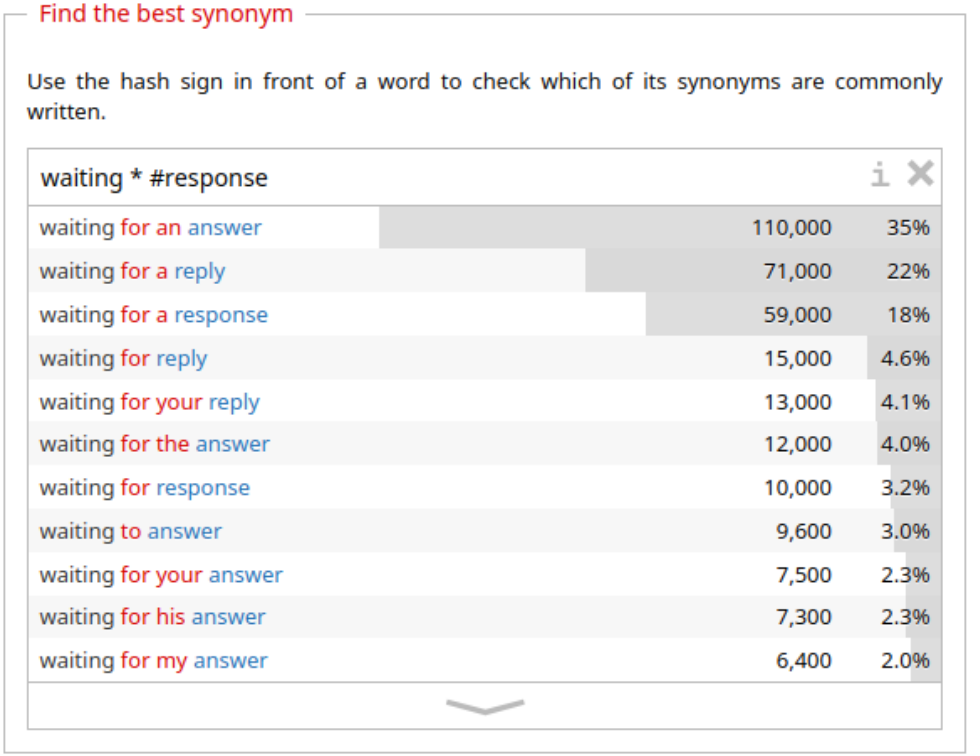


[www.sketchengine.eu]

Text Corpora

Research in Language Use (continued)

Compare usages of a word, analyse keywords, analyse frequencies, find phrases, idioms, etc.



Find the best synonym

Use the hash sign in front of a word to check which of its synonyms are commonly written.

waiting * #response

waiting for an answer	110,000	35%
waiting for a reply	71,000	22%
waiting for a response	59,000	18%
waiting for reply	15,000	4.6%
waiting for your reply	13,000	4.1%
waiting for the answer	12,000	4.0%
waiting for response	10,000	3.2%
waiting to answer	9,600	3.0%
waiting for your answer	7,500	2.3%
waiting for his answer	7,300	2.3%
waiting for my answer	6,400	2.0%

[netspeak.org]