NATURAL LANGUAGE PROCESSING

LESSON 4: CORPUS (FEATURES AND ANALYSIS)

OUTLINE

- What Is Corpus?
- Corpus Types
- Approaches In Corpus Design
- Corpus Analysis Softwares
- Corpus Examples
- Corpus Evaluation Algorithms
- Examples of Corpus Tools

WHAT IS CORPUS?

- A corpus (corpora is the plural) is simply a body of text that has been collected for some purpose.
- Corpora prepared for NLP applications provide the data source for many machine-learning approaches.
- Corpora may also be collected for a specific task. For instance, when implementing an email answering application, it is essential to collect samples of representative emails.

WHAT IS CORPUS?

- A balanced corpus contains texts which represent different genres (newspapers, fiction, textbooks, parliamentary reports, cooking recipes, scientific papers etc)
- Corpora are essential for most modern NLP research. Because of the collection difficulties, newspaper texts often used.
- Corpora are often marked and annotated in some way. One of the most important types of annotation is Part of Speech Tagging.
- **Brown** and **Lancaster-Oslo-Bergen** corpus >>1 million words. **British National** corpus >> 100 million words of spoken English.

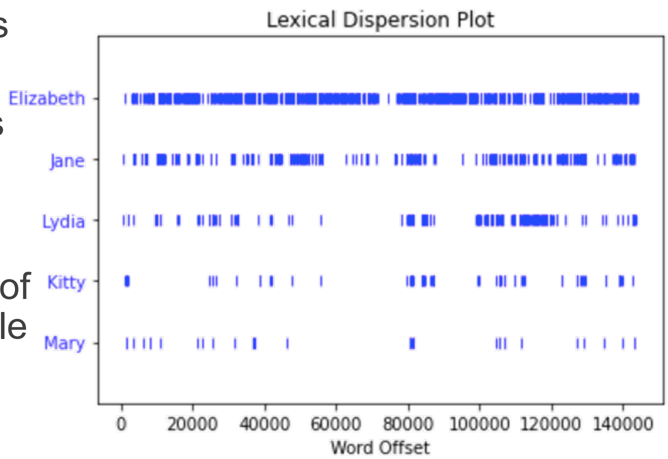
SOME TERMS FOR CORPUS

- **Word Frequency:** It is used to know which words occur most frequently in the texts via a software. This software counts up the occurrences of each word form and list them in descending or ascending order of frequency, or alphabetically.
- **Co-occurrences of words:** It is to become together of the words in the same environment. It can be found by using any concordance tool which can analyze a text.
- **Distribution of words:** It is used to discover how certain words or sets of words are distributed in the various parts of a text.

SOME TERMS FOR CORPUS

Here is an interesting analysis method, **lexical dispersion**.

It shows where specific words appear throughout the entire document. It can be used to visualize the evolution of language or perhaps the role of characters. Here is an example on the right.



CORPUS TYPES

- Corpora may encode a language according to any communication method. For example, while some corpora contain spoken language, most are written language.

- Many corpora cover data from more than one communication methods, such as the British National Corpus. It includes both written and spoken language.

CORPUS TYPES

Written corpora

- Corpora, which represents written language, can include both electronic texts and electronic images.
- Computer-generated electronic texts such as e-pubs are more preferred in NLP studies since they have less technical difficulties.
- Electronic images obtained by scanning printed sources such as books, on the other hand, have problems related to other disciplines such as image processing, as well as traditional NLP problems.

Sec. 15.3.1

CORPUS TYPES

Spoken corpora

- These corpora are typically produced by recording human interactions and then encoding them by humans.
- These encodings are usually attached to the original recording through a process called time alignment so that the encoded texts show the correct position in the audio file.

9

CORPUS TYPES

Comparable corpora

- It contains components in two or more languages that have been collected using the same sampling method.
- They can be the same proportions of the texts of the same genres, in the same domains, in a range of different languages, in the same sampling period.
- The sub-corpora of a comparable corpus are not translations of each other.

10

CORPUS TYPES

Parallel corpora

- By contrast, a parallel corpus contains native language source texts and their translations.
- For a parallel corpus to be useful, an essential step is to align the source texts and their translations, annotating the correspondences between the two at the sentence or word level.
- This type of corpus is great for training machine translation systems.

11

APPROACHES IN CORPORA DESIGN

Two broad approaches to the issue of choosing what data to collect have emerged:

- **the monitor corpus approach**, where the corpus continually expands to include more and more texts over time;
- and **the balanced corpus** or sample corpus approach.

Corpus builders adopt an existing corpus model to suit their purpose when collecting their data.

12

APPROACHES IN CORPORA DESIGN

Monitor corpora

This approach is used if the aim is to track the change of language over time.

- A monitor corpus is a dataset which grows in size over time and contains a variety of materials. The relative proportions of different types of materials may vary over time.
- The Bank of English (BoE), developed at the University of Birmingham. The BoE was started in the 1980s.

APPROACHES IN CORPORA DESIGN

Balanced corpora

- In contrast to monitor corpora, balanced corpora, also known as sample corpora, try to represent the chosen text types for a specific span of time.
- This approach is chosen if the goal is to focus on a specific time period. A balanced corpus contains a wide variety of text categories that are supposed to represent the diversity of languages studied.
- In doing so they seek to be balanced and representative within *a particular sampling frame*.

CORPUS ANNOTATION

- **Linguistic analyses** encoded in the corpus data itself are usually called corpus annotation.
- For example, we may wish to annotate a corpus to show parts of speech, assigning to each word a grammatical category label. So when we see the word **talk** in the sentence **I heard John's talk and it was the same old thing**, we would assign it the category noun in that context.
- There are online systems that will allow with automatic annotation without having to install any software on computer.

CORPUS ANNOTATION

- CLAWS tagger manages grammatical tagging of a small-to-medium text using the web-interface.
- A more complex form of grammatical annotation is parsing. One easy way to try out parsing is to use the **Online Stanford Parser**.
- This program does two different types of parsing: **dependency parsing** and **constituency parsing.**

CORPUS INFORMATION

Corpora typically contain within them three types of information that can help in investigating the data: **metadata, textual markup, and linguistic annotation.**

- **Metadata** is information that tells you something about the text itself: for example, the metadata may tell you who wrote a text and when it was published. The metadata can be encoded in the corpus text, or held in a separate document or database.
- **Textual markup** encodes information within the text other than the actual words, for example, the sentence breaks or paragraph breaks in a written text.

CORPUS INFORMATION

In spoken corpora, the information conveyed by the **metadata and textual markup** may be very important to the analysis.

- The metadata would typically identify the speakers in the text and give some useful background information on each of them, such as their age and sex.
- Textual markup would then be used to indicate utterance boundaries.

CORPUS INFORMATION

- We can also encode linguistic information within a corpus text, so that we can describe it as analytically or **linguistically annotated**.
- For instance, the angle-bracket tags of XML can easily be used to indicate where a noun phrase begins and ends:

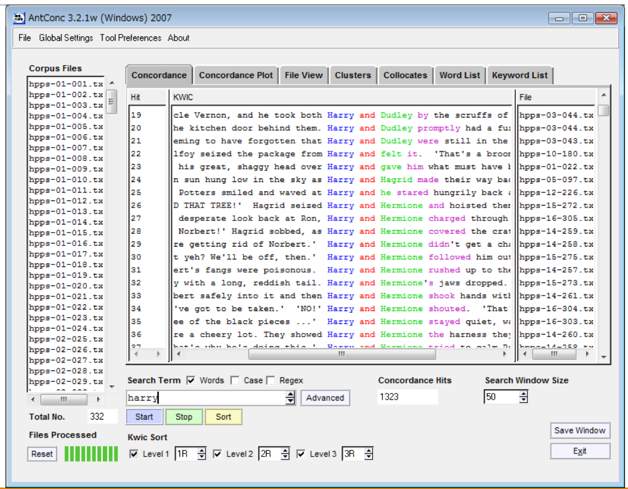
<np>The cat</np> sat on <np>the mat</np> .

CORPUS INFORMATION

- A wide range of annotations have been applied automatically to English text, by analysis software (also called taggers) such as:
  - constituency parsers such as Fidditch
  - dependency parsers such as the Constraint Grammar system
  - part-of-speech taggers such as CLAWS
  - semantic taggers such as USAS
  - lemmatizers or morphological stemmers

CORPUS ANALYSIS SOFTWARES

- A concordancer allows to search a corpus and retrieve from it a specific sequence of characters of any length-perhaps a word, part of a word, or a phrase.



CORPUS EXAMPLES

Name of corpus	Year published	Size	Collection contents
British National Corpus (BNC)	1991–1994	100 million words	Cross section of British English, spoken and written
American National Corpus (ANC)	2003	22 million words	Spoken and written texts
Corpus of Contemporary American English (COCA)	2008	425 million words	Spoken, fiction, popular magazine, and academic texts

Sec. 13.3.1

EXAMPLES OF TURKISH CORPORA (Ç.Ü. Türkoloji)

Project title: Türkiye Türkçesi Çevrim İçi Haber Metinlerinde Yeni Sözlerin (Neolojizm) Otomatik Çıkarımı (111K223 TÜBİTAK SOBAG)

Project Statistics	
News sources	643
Distinct link count	12,763,184
Data size	1,055 GB
Total token	41,224,731
Total bigram	416,306,339

23

Sec. 13.3.1

EXAMPLES OF TURKISH CORPORA (METU Corpus)

METU Turkish Corpus (MTC)

- METU Turkish Corpus is a collection of 2 million words of post-1990 written Turkish samples. A subset of the corpus is used in METU-Sabanci Turkish Treebank.
- The words of METU Turkish Corpus were taken from 10 different genres

METU-Sabanci Turkish Treebank

- Morphologically and syntactically annotated treebank corpus of 7262 grammatical sentences.
- The structure of METU-Sabanci Turkish Treebank is based on XML.

24

EXAMPLES OF TURKISH CORPORA - TSCORPUS

- Published corpora serves a dataset o derived from various sources; online newspapers, forums, social media, academic papers etc.
- 1,329,708,730 tokens for now, still processing
- Available online with part-of-speech and morphological tagging.
- The project is free for academic studies and researches.

CORPUS EVALUATION ALGORITHMS

When working with corpus datasets in NLP, three major types of machine-learning algorithms are typically used:

- Supervised learning
- Unsupervised learning
- Semi-supervised learning

CORPUS EVALUATION ALGORITHMS

Supervised learning

Any technique that generates a function mapping from inputs to a fixed set of labels (the desired output).

The labels are typically metadata tags provided by humans who annotate the corpus for training purposes.

CORPUS EVALUATION ALGORITHMS

Unsupervised learning

Any technique that tries to find structure from an input set of unlabeled data.

CORPUS EVALUATION ALGORITHMS (Unsupervised)

Algorithms	Tasks
Clustering	Genre classification, spam labeling
Decision trees	Semantic type or ontological class assignment, coreference resolution
Naïve Bayes	Sentiment classification, semantic type or ontological class assignment
Maximum Entropy (MaxEnt)	Sentiment classification, semantic type, or ontological class assignment
Structured pattern induction (HMMs, CRFs, etc.)	POS tagging, sentiment classification, word sense disambiguation

CORPUS EVALUATION ALGORITHMS

Semi-supervised learning

Any technique that generates a function mapping from inputs of both labeled data and unlabeled data; a combination of both supervised and unsupervised learning.

EXAMPLES OF CORPUS TOOLS

- **Stanford CoreNLP** provides a set of human language technology tools. It can give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, phrases and syntactic dependencies.
- **Brat Rapid Annotation Tool** is a browser-based rapid annotation tool for text annotation; that is, for adding notes to existing text documents.
- **GATE** is a well established open-source suite of tools for NLP tasks in general and named entity recognition/semantic tagging in particular.