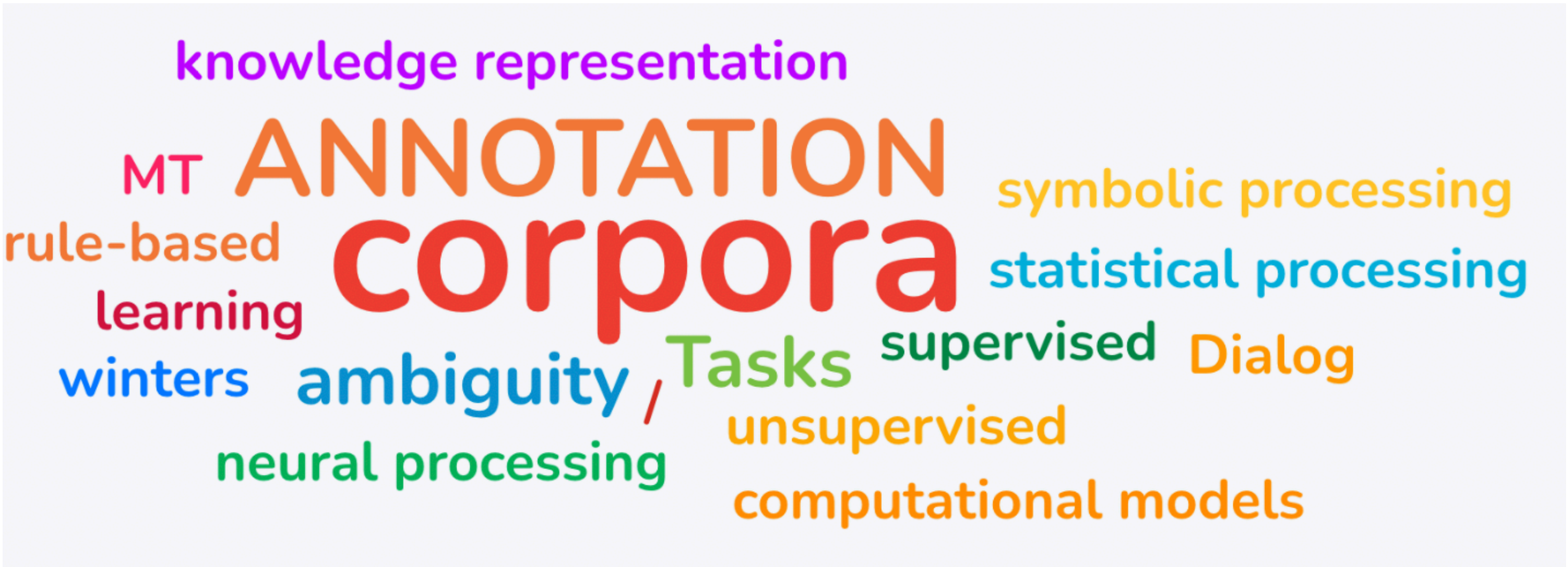
Preliminary Steps for Building Linguistic Resources: Tokenization - PART-I

Resources for Natural Language Processing
LM Language Technologies and Digital Humanities
2024-25

Cristina Bosco

RECAP


knowledge representation
MT
ANNOTATION
symbolic processing
rule-based
corpora
statistical processing
learning
supervised
winters
ambiguity
Tasks
Dialog
unsupervised
neural processing
computational models

Resources and NLP

Linguistic resources are necessary to train NLP statistical models:

Resources and NLP

Linguistic resources are necessary to train NLP statistical models:

- Linguistic knowledge is automatically learned by statistical models from resources (corpora)

MACHINE LEARNING

Resources and NLP

Linguistic resources are necessary to train NLP statistical models:

- Linguistic knowledge is automatically learned by statistical models from resources (corpora)

MACHINE LEARNING

- From annotated resources it is possible to obtain more and more precise knowledge

SUPERVISED MACHINE LEARNING

To take into account: how computers deal with texts

In a large number of cases, dealing with language, **computers** achieve the same results as humans

But they **use completely different procedures** and utilise different forms of knowledge than humans.

An example is how search engines deal with documents collected from the web.

An example: search engines

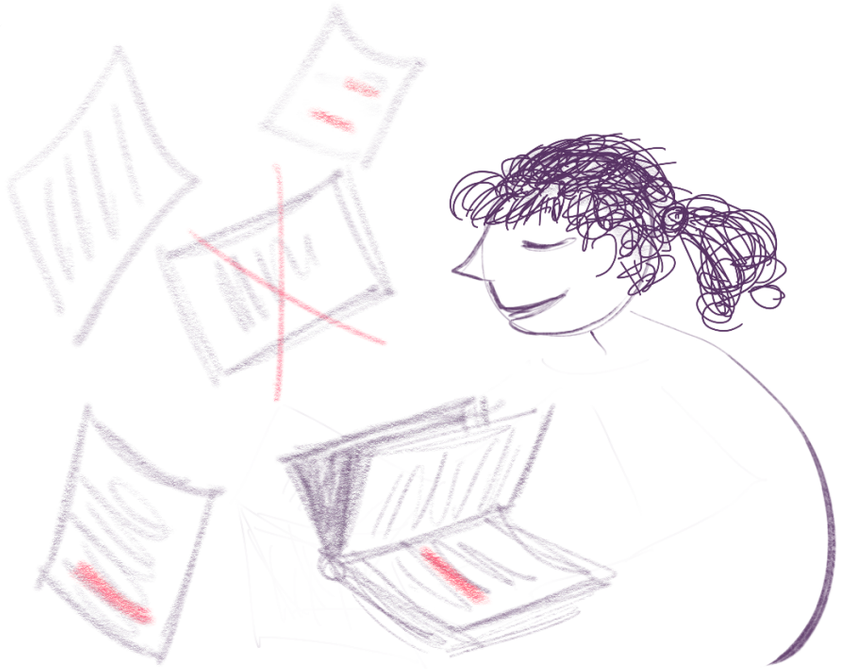
Search engines are tools that provide a basic treatment of the language of documents published on the web.
They find the documents we are interested in.

More precisely, **the task of a search engine is**:

Based on a query (entered by a user), the search engine creates an ordered list of web documents that match the query.

An example: search engines

What kind of procedures do people use to accomplish the task of finding (in a large collection) documents that match a query and ordering them?



To find and order a set of documents, humans **read** them, **understand** their meaning, then **decide whether the content matches the query** and **to what extent it does**.

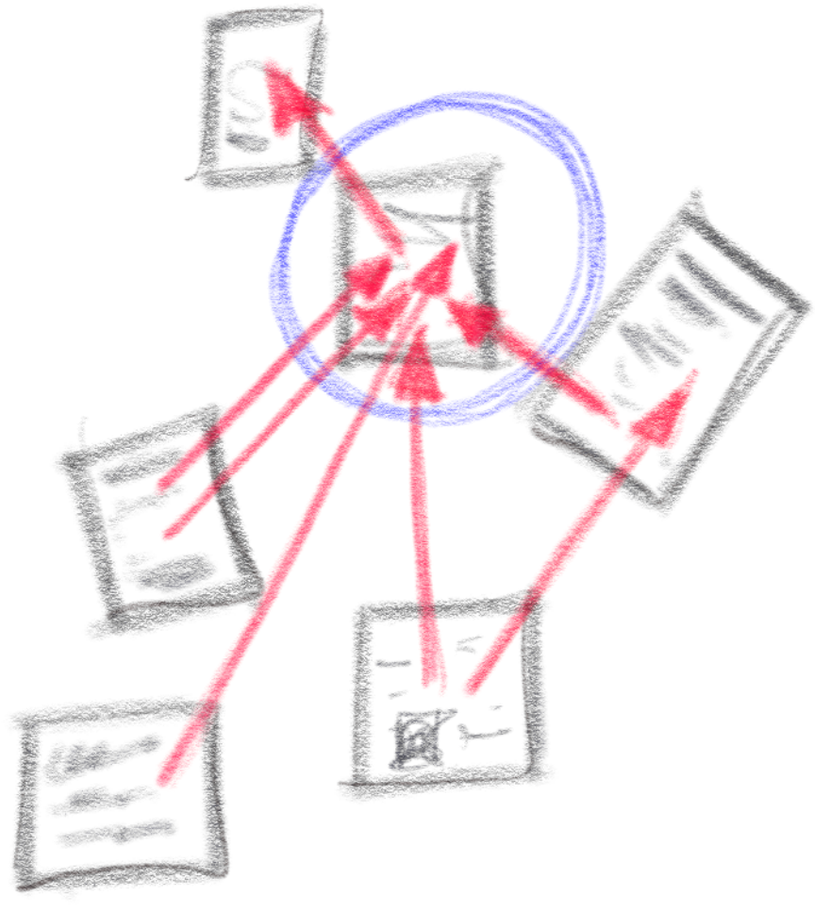
This can be complex and can take a long time if the collection is large as the web!

An example: search engines

A search engine applies a completely different procedure.

It collects all the documents that more or less match with the user query.

To decide the order in which documents matching the user query must be given to the user, it simply **counts the links that enter each document** (starting from other documents).
The more a document is linked the more it is considered as valuable … **regardless of the meaning** of the text in that document.

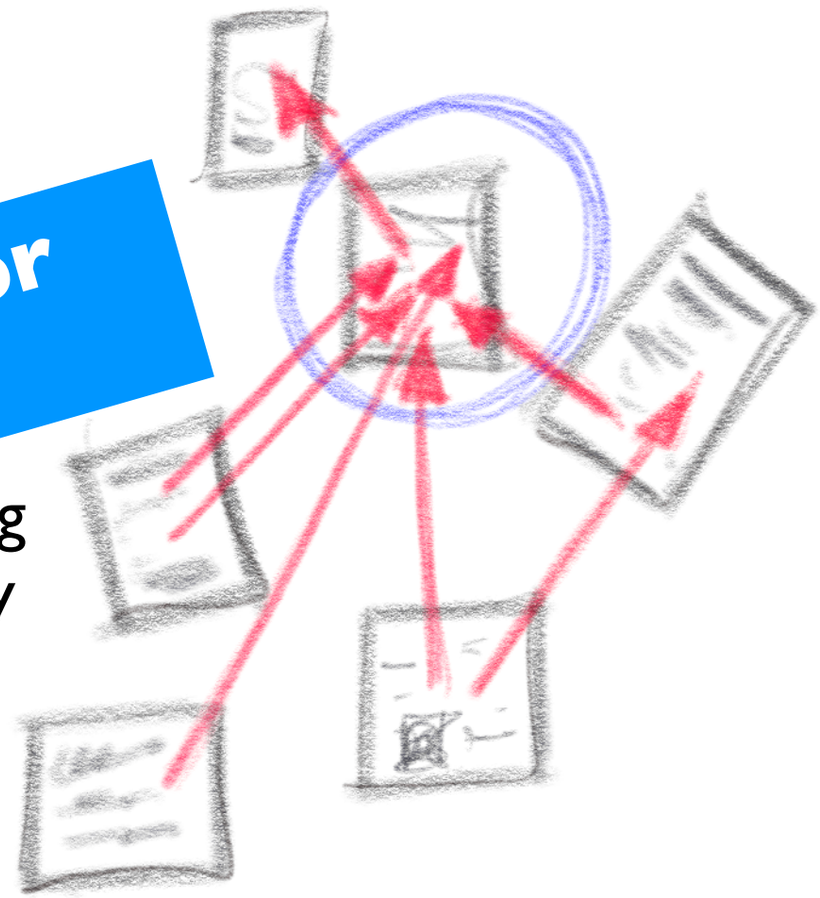


An example: search engines

A search engine apply a completely different procedure.

It collects all the documents that more or l
match with the user query.

To decide th ents matching
th to the user, it simply
s that enter each
d ent (starting from other documents).
The more a document is linked the more it is considered as valuable … **regardless of the meaning** of the text in that document.



Text meaning does matter for search engines

Overview

- Resources, tasks and structure
- Annotation
- Segmentation:
  - Tokenization, tokens, lexemes and morphemes
  - Challenges in tokenization

Resources and tasks

Several different types of linguistic resources exist which are built **for different tasks** with different characteristics:

Resources and tasks

Several different types of linguistic resources exist which are built **for different tasks** with different characteristics:

- The **structure** varies from a simple collection of words (dictionaries, lexical, thesauri …) to collection of sentences (treebanks, parallel corpora, …) or whole documents (corpora for sentiment analysis …)

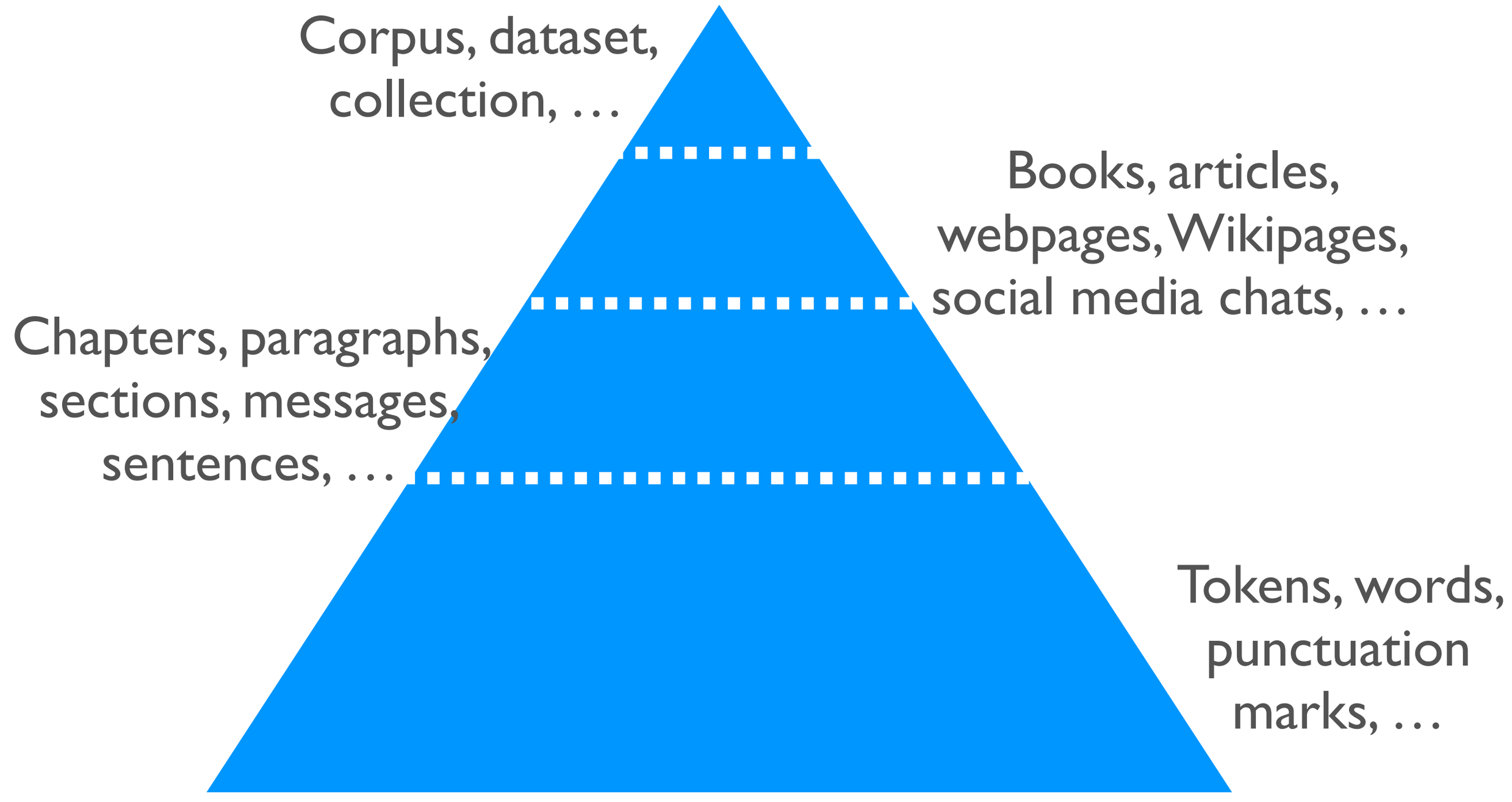
Resources and tasks

Several different types of linguistic resources exist which are built **for different tasks** with different characteristics:

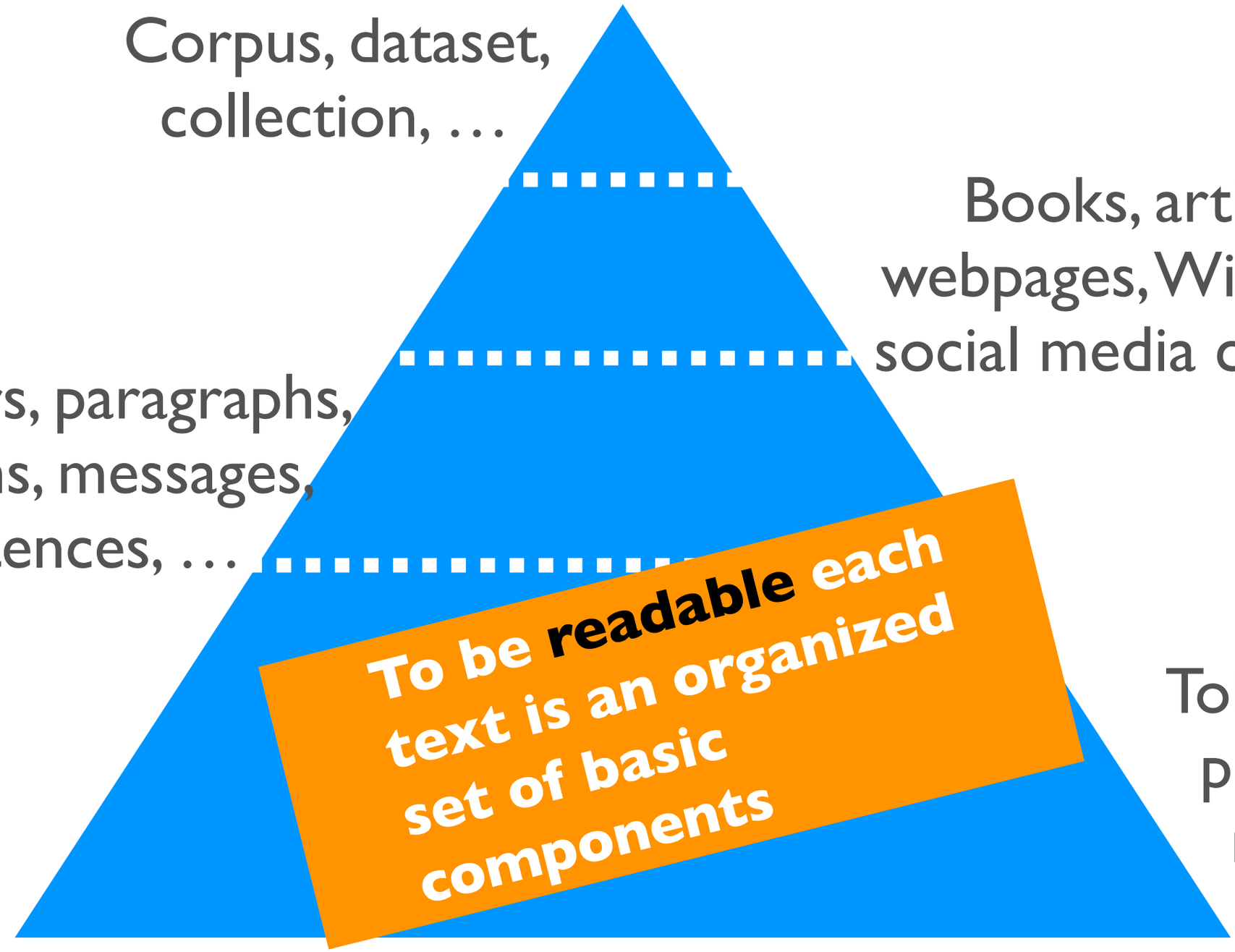
- The **structure** varies from a simple collection of words (dictionaries, lexical, thesauri …) to collection of sentences (treebanks, parallel corpora, …) or whole documents (corpora for sentiment analysis …)

- There is a close **relationship between the structure and the annotation**: certain forms of annotation are for single words / sentences (morphology / syntax) while other forms are for whole documents (sentiment)

Structure and organisation of resources



Structure and organisation of resources



Annotation: a definition

Annotation consists of adding information to the pure text.
The main aim it at making **explicit the linguistic knowledge which is implicit in data**.

Many linguistic knowledge can be annotated (morphological, syntactic, semantic…).

The data are segmented according to the type of knowledge that must be annotated: lexical referring to individual words, syntactic referring to entire sentences, semantic referring to phrases or messages (which may include several sentences) ...

From data to information

- **Data** = raw texts

- **Information** = knowledge that can be extracted from raw texts

- **Annotation** = raw texts with associated information

From data to information

- **Data** = raw texts

- **Information** = knowledge that can be extracted from raw texts

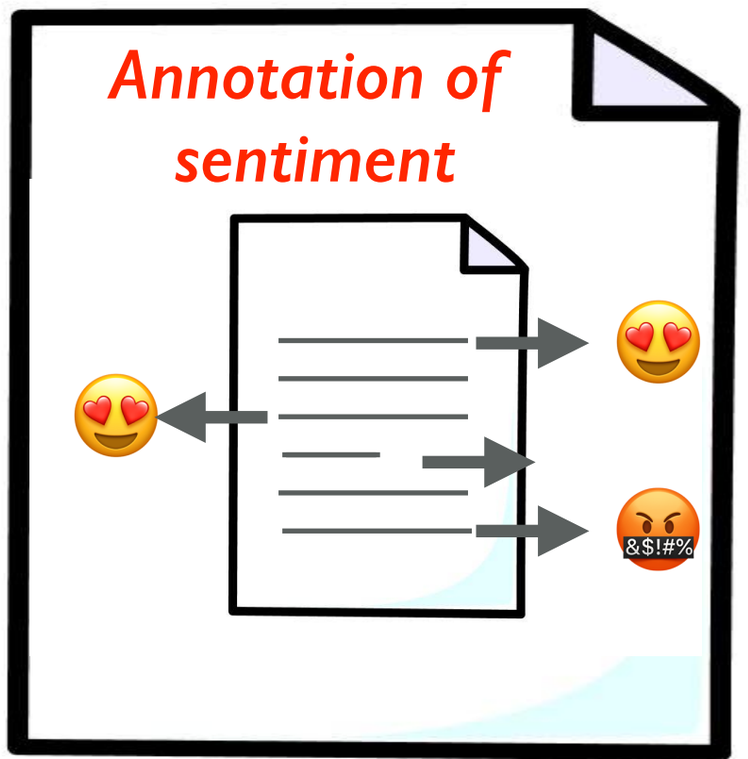
- **Annotation** = raw texts with associated information

From data to information

- **Data** = raw texts

- **Information** = knowledge that can be extracted from raw texts

- **Annotation** = raw texts with associated information



Resources we focus on: annotated corpora

Corpora are large collections of texts (singular ***corpus***, plural ***corpora***) and the mostly used resources for NLP.

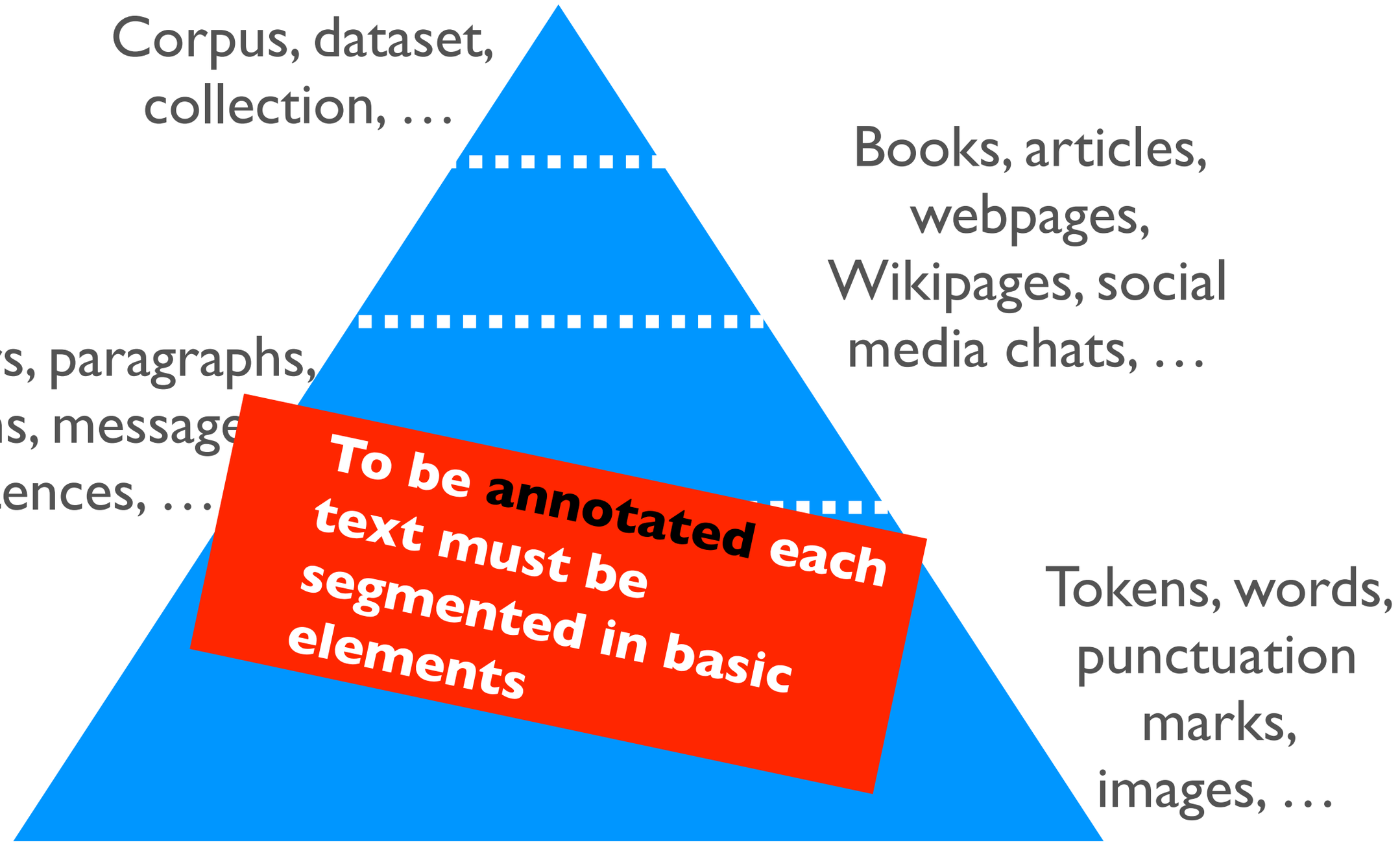
Corpora can be **annotated** for different tasks, but the more used annotations applied on corpora are:

Part of Speech tagging > **morphological** analysis

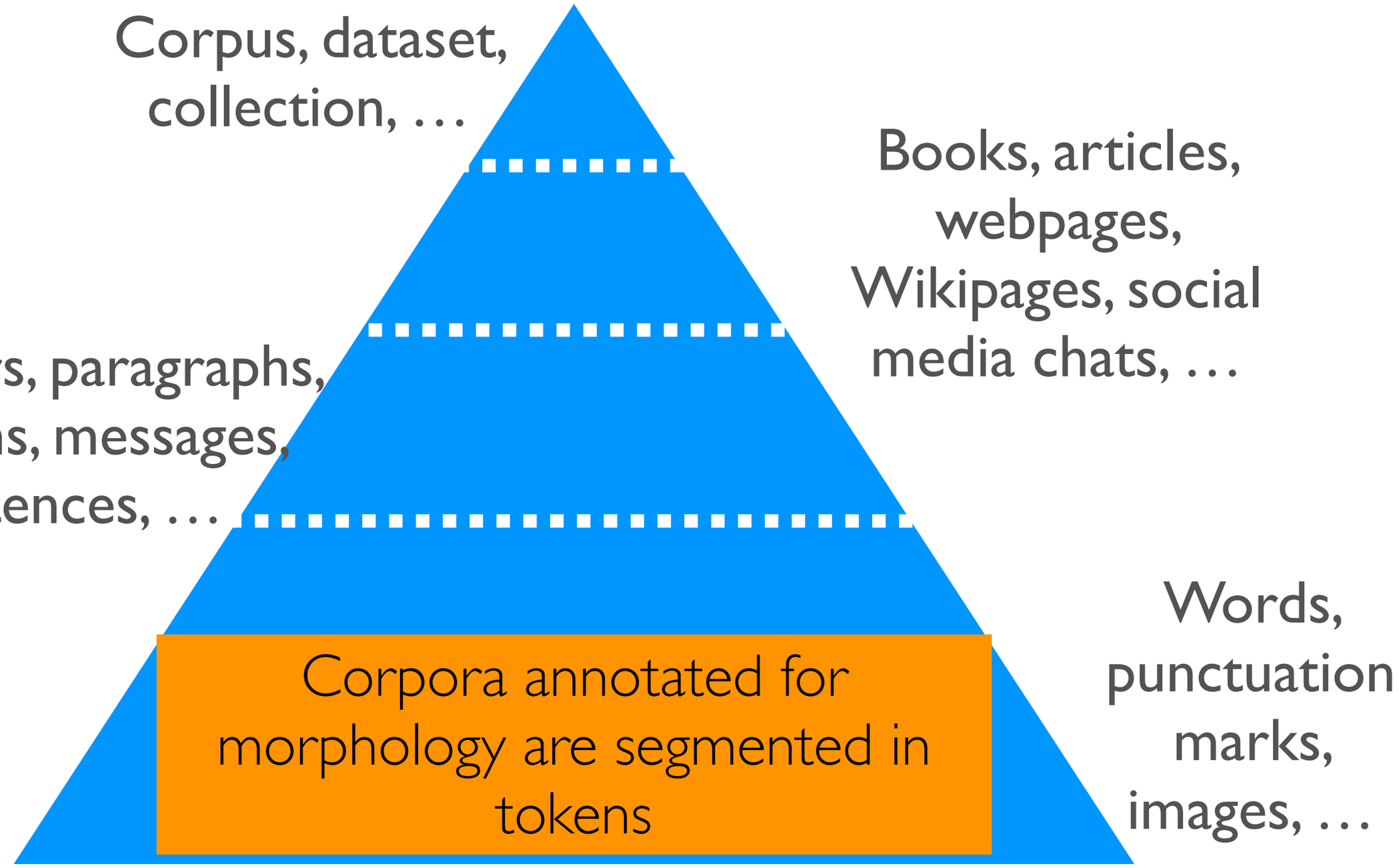
Parsing > **syntactic** analysis

The availability of corpora annotated with morphology and syntax is crucial for performing preliminary steps of complex tasks too, such as QA, Sentiment analysis, dialog, MT …

Structure and organisation of resources



Structure and organisation of resources



Segmentation

Preliminarily to any form of analysis linguistic data must be segmented in the basic elements to which the annotation must be associated.

For the morphological annotation the basic elements are TOKENS.

What is a token?

How does the segmentation in tokens work?

Tokenization is the basic task of preparing the text for later analysis, *normalizing* it and converting it into a format that is suitable for the computational analysis that will applied later.

In particular, it serves to remove the "noise" caused by the presence of characters that could cause problems during subsequent analysis and to give these characters a meaning.

Token and tokenisation

The tokens in the sentence

There are several difficult topics in this course.

can be obtained by applying tokenization.
They are 9 and precisely

<There>$_1$ <are>$_2$ <several>$_3$ <difficult>$_4$
<topics>$_5$ <in>$_6$ <this>$_7$ <course>$_8$ <.>$_9$

Token and tokenisation

In computer science a *token* is the minimal unit of information.

In NLP, **token** means *lexical token*, simply elements on which later analysis can be easily applied.

Tokenization consists of dividing the text into minimal units containing morphological information called tokens.

It is the first step in language analysis and it is necessary e.g. to create corpora annotated for morphology and syntax or to apply other forms of analysis.

Exercise about tokenisation

Exercise A and B