Ambiguity, models and resources

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

Cristina Bosco

Ambiguity

Formalising and making all the existing knowledge available to NLP systems cannot be a solution to the problem of ambiguity, for at least two reasons:

- the knowledge is too large to be formalised

- the knowledge varies over time, language, genre, domain …

- linguistic knowledge does not provide guidance how to choose the right interpretation in contexts in which ambiguity occurs

Ambiguity

No precise formalised description of all aspects of language can be complete since the knowledge required to use language includes:

- the meaning of words
- the infinite grammatical structures, in which they can freely occur
- the contexts in which they must be used correctly and properly.

Natural languages are open systems.

This means that we have also to explain all the productions generated in their context, and to preview all future productions that can be generated.

Language as an open system

The goal of NLP is **NOT only** to provide **a description** of natural language that is precise and detailed enough.

To build machines that can cope with language as humans do a MODEL is needed.

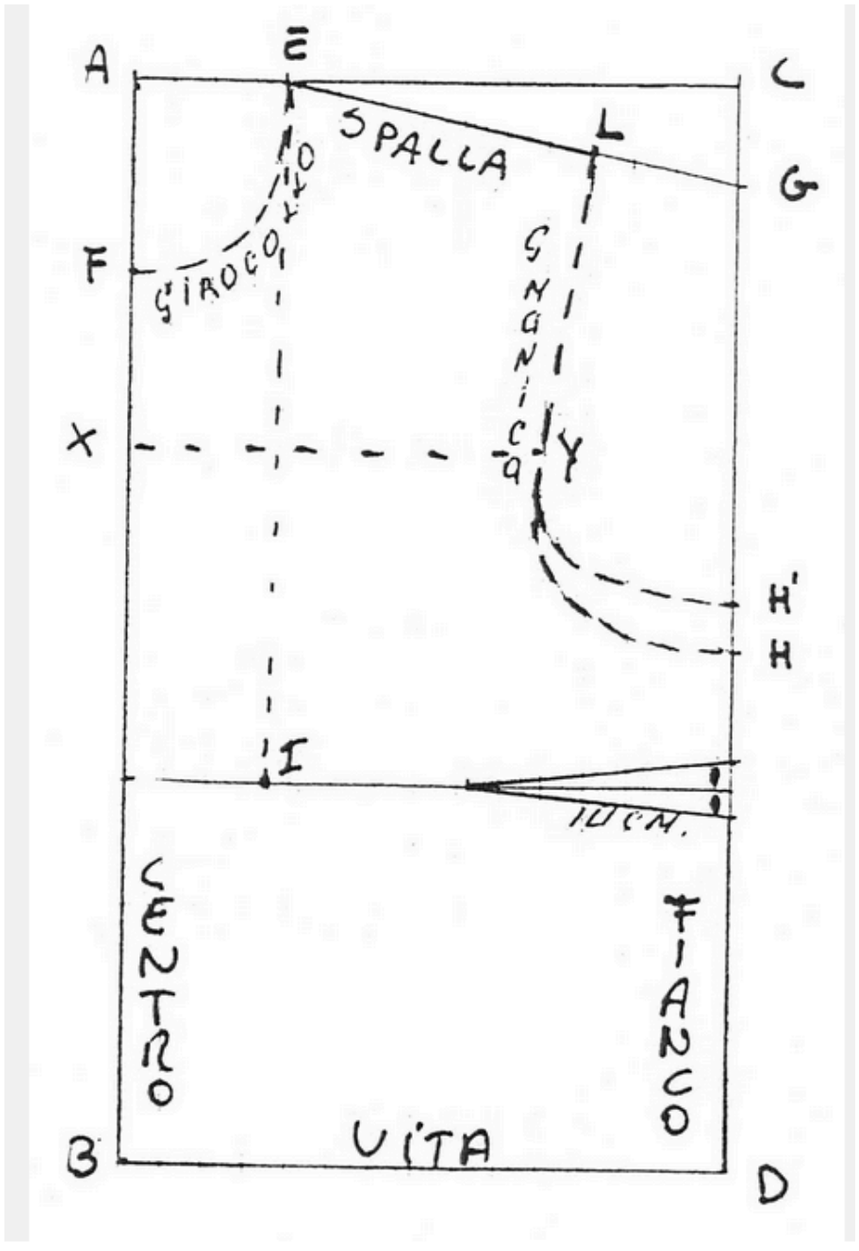
And considering that human language is an open system, we have to built an open model.

Modelling language

What does modeling mean?

A model is a **schematic and abstract** representation of a reality or an object.

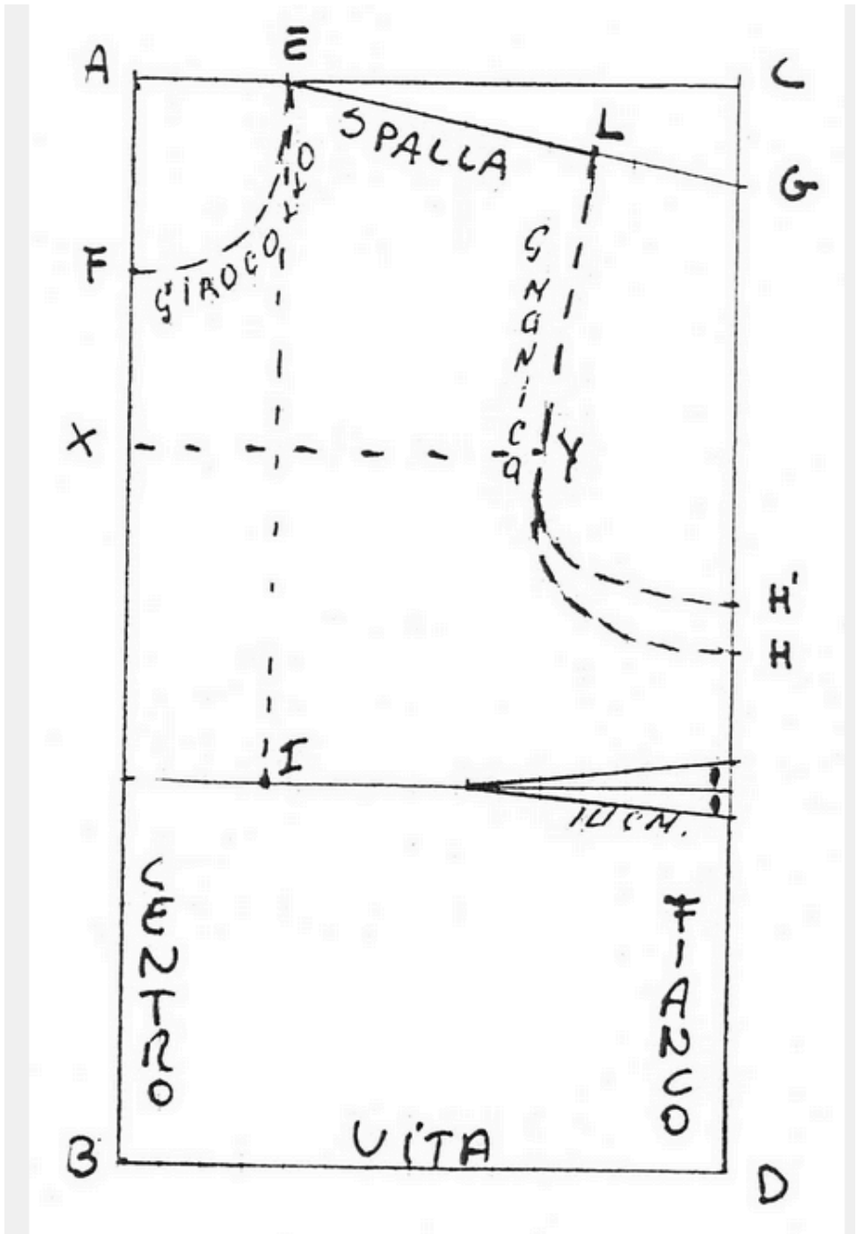
The model of a dress, for example, describes with good approximation the shape, the parts that make it up and the proportions between height and width.
It is **schematic** because it only describes the main components, and it is **abstract** because it does not describe a very precise dress (it does not talk about the type of fabric or the color).



Modelling language

A model is **able to generalize** the properties of a reality or an object

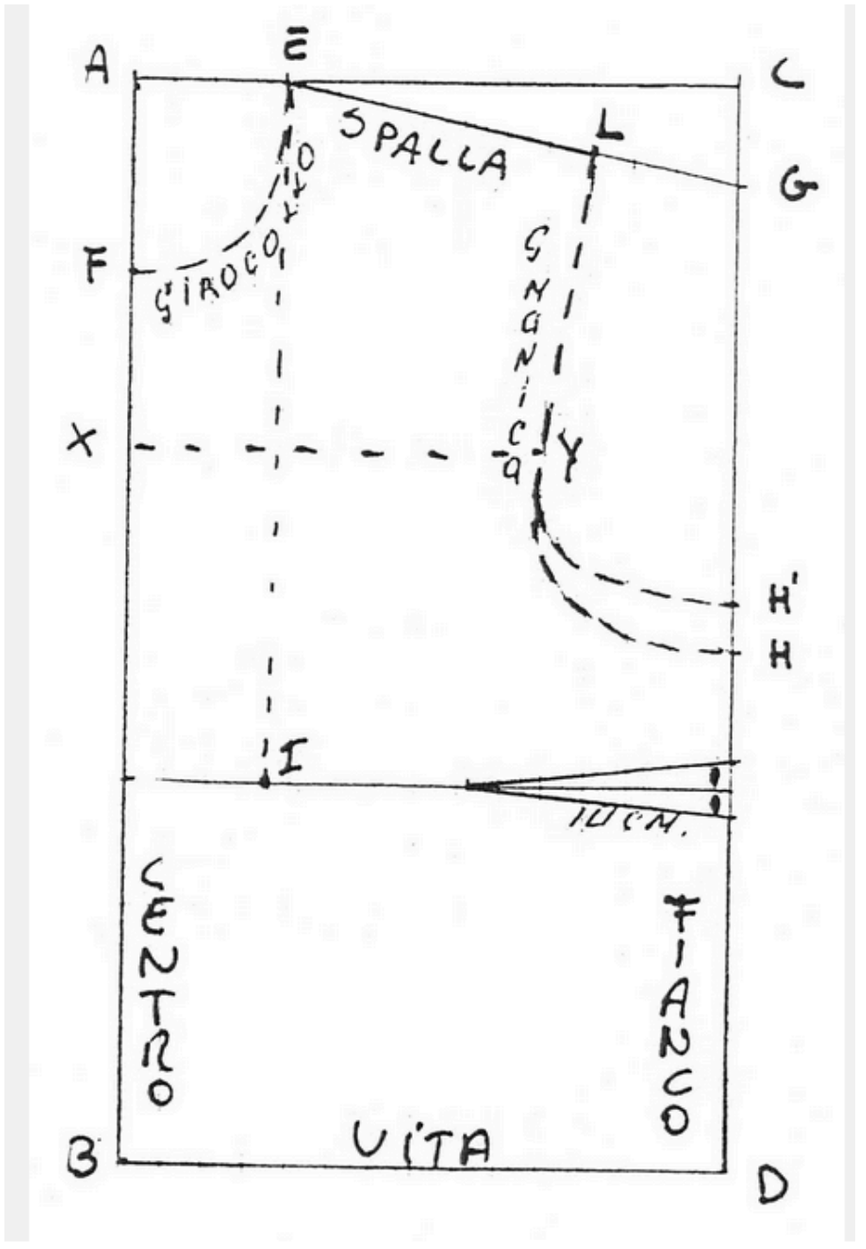
The model of a dress, for example, **describes in a fairly general way** the characteristics that unite so many dresses, but which differ in other aspects, such as color, fabric, size...



Modelling language

A model is **able to generalize** the properties of a reality or an object

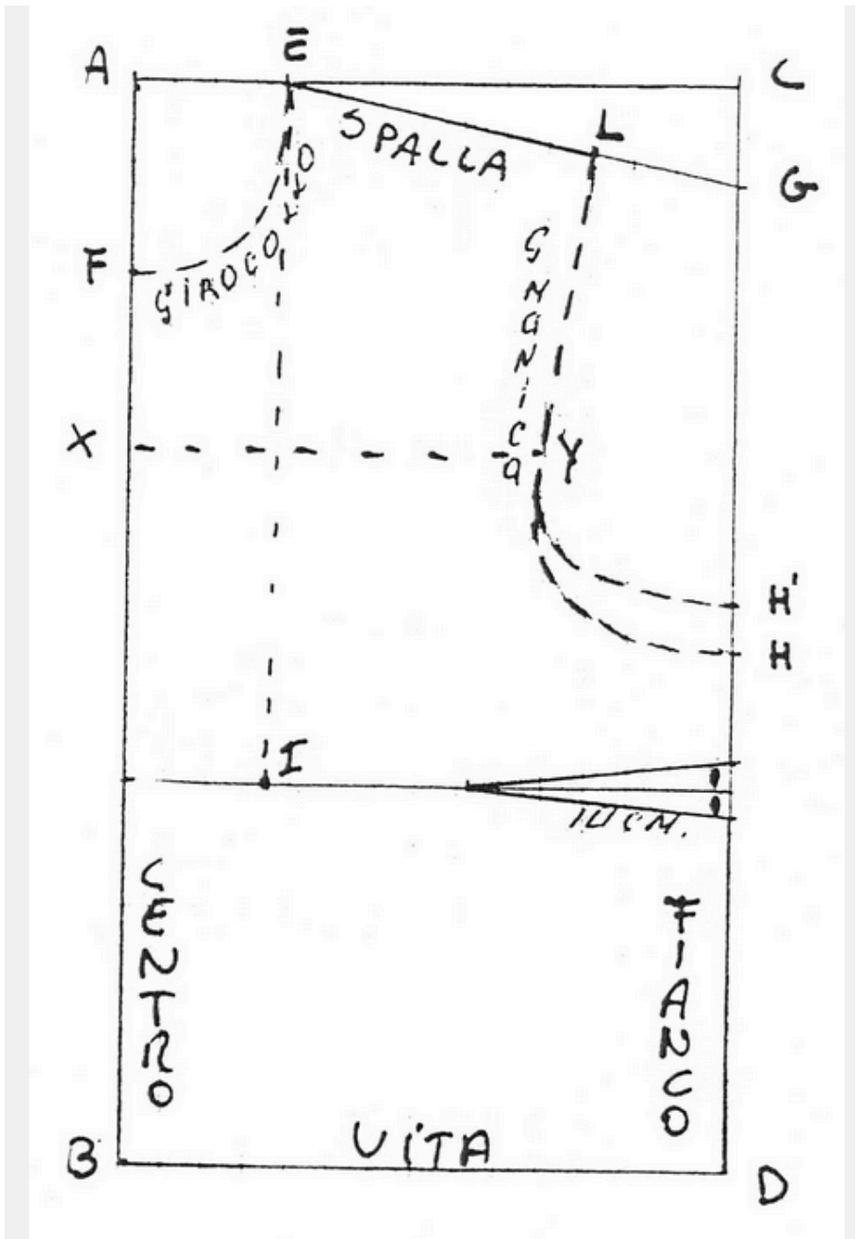
The ability of the model to generalize depends on the fact that the **data used to construct the model are inherently limited manifestations** of the reality or object that the model describes



Modelling language

A valid model must be **able to explain** how it arrives at what it describes

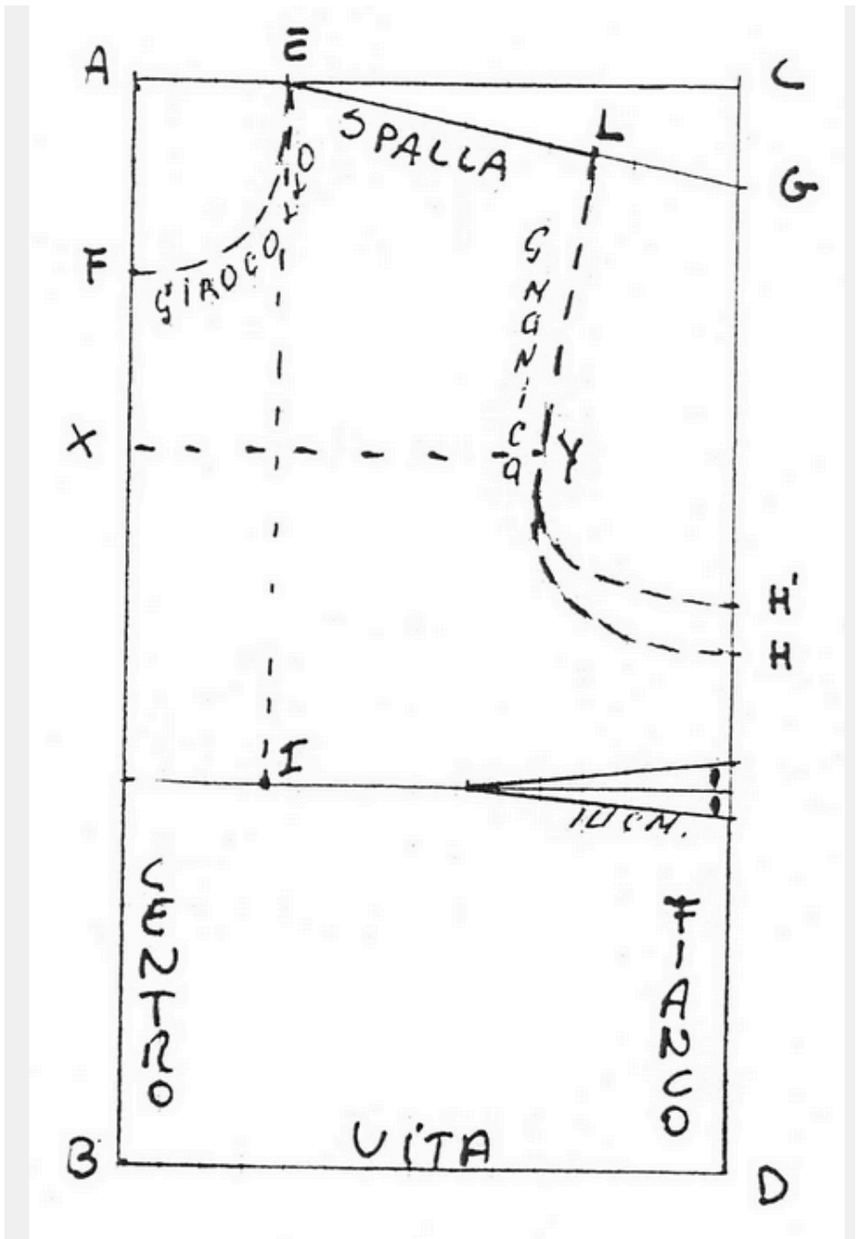
The model of a dress should be able to explain, with some approximation, **how** a dress **is made** and then allow us **to distinguish** the dress from another kind of object, such as a shoe or a belt.



Modelling language

A valid model must be **able to make predictions** about how it arrives at what it describes

The model of a dress should be able to **predict**, with some approximation, **how all the dresses built from the model will be made**, without knowing how many and which ones will actually be made



Modelling language

The main goal of computational linguistics is the development of **computational models of human language**.

A computational model is an **abstract** and **schematic** description of a natural language
that is **able to generalize, explain and predict everything related to language and the use that speakers make of it**.

And it also has a form that makes it usable by a computer!

Open and closed models

A model is closed if the language it describes can be seen as a closed system, i.e. if it is possible to observe all sentences generated using this language.

For example, if an L language is no longer spoken, we have all the sentences ever produced in L, and we can model it as s a closed system.

This can be done whether our objective is to explain all the productions of this language L.

Open and closed models

But a language is in itself an open system: it includes a finite set of words and rules that allows the generation of infinite sentences!

Typically, **a language is an open system**, since new sentences are constantly being generated using its words and rules.

The purpose of a model of an open system is not only to **explain** the set of productions of that language generated until the moment of our observation, but also to **predict** how sentences generated in the future might be formed.

Open and closed models

The distinction between open and closed models is therefore in the eyes of the observer:

Explaining and predicting the behavior of a language system are not very different things, what changes is the set of data considered.

Explaining means looking for a coherent description of the language, while predicting means having criteria to recognize new sentences as consistent with the language described in the model.

NLP models

In NLP, language is now treated as an open system and through statistical and neural approaches.

The **first attempts** to model language (from 1950 to 1990) assumed that language is a closed system, to simplify the task and because they are based on symbolic approaches with good explanatory skills but reduced predictive abilities.

These NLP systems can be suitable for *toy domain*, (often very) limited (manageable) portions of language with small sets of information (rules and vocabulary). Not for language at large.

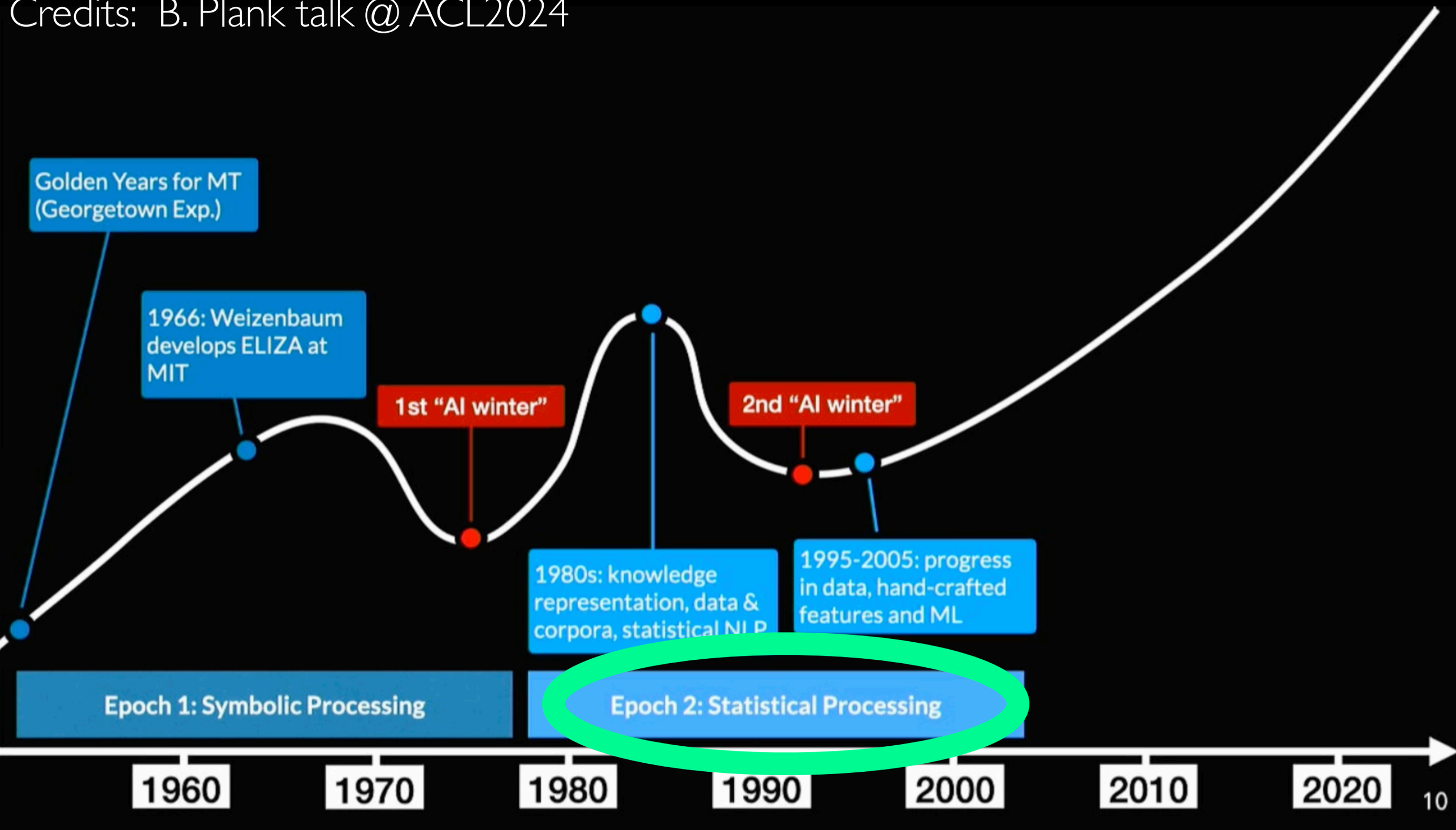
NLP models

Statistical models try to capture the regularities of natural language.

They use the probability that a grammatical category or structure occurs to predict it.

NLP: statistical processing

Credits: B. Plank talk @ ACL2024



NLP models

In NLP, has been used symbolic, statistical and neural linguistic models.

The distinction between these types of models is primarily a historical question:

- the first models of computational linguistics were **symbolic**

- since the 1990s, **statistical** models (machine learning)

- in the last 10 years, simple statistical models have been gradually replaced by **neural** models (deep learning).

NLP models

Symbolic models are based on rules defined by experts that are applied to the data analysis.

Some examples:

- for a tokenizer, a rule may be that the space character is considered a separator between two tokens

- for a pos tagger, a rule may be to recognize a token that begins with a capital letter and is not at the beginning of a sentence as a proper noun

NLP models

Statistical models basically consider similar rules as symbolic models, but 1) learn them directly from large data samples to which they are applied by machine learning techniques and 2) associate probabilities to these rules.

An example:

- a statistical pos tagger assumes that a token is a common name because it follows an article. This is based on statistical knowledge and more precisely on how often the token that follows an article is a common noun.

NLP models

Neural models differ with and between statistical models depending on the algorithms and principles they use.

Neural models are statistical models that use specific algorithms, called neural networks, that have a structure that simulates the behavior of human neurons.
Learning requires a large amount of data and is referred to as deep learning.

NLP models

Model	Data	Are results explainable?
Simbolic	No	Yes
Statistical	Thousands	In part
Neural	Billions	No

Ambiguity and models

Can be models helpful also for solving ambiguity?

Yes, statistical models can be helpful for addressing ambiguity !

Statistical models are able to generalise knowledge available in linguistic data, and mostly in corpora. They can also associate each of the characteristics of a language with the probability of its occurrence.

Ambiguity and models

Ambiguity consists in the fact that there are two alternative solutions to a case: e.g.
two syntactic structure for the English sentence

‘Mary see a man with a telescope’

It is Mary using the telescope? or the man using the telescope?

Both solutions are valid, i.e. they do not violate the rules of grammar.

Ambiguity and models

But from a statistical point of view the two solutions are the same?

One of the two solutions might be **more probable**.

This means that statistical probability can be used as a criterion to put in order the two alternative and to decide between them.

The solution chosen on the basis of probability theory is definitely not correct, but probably better than the others.

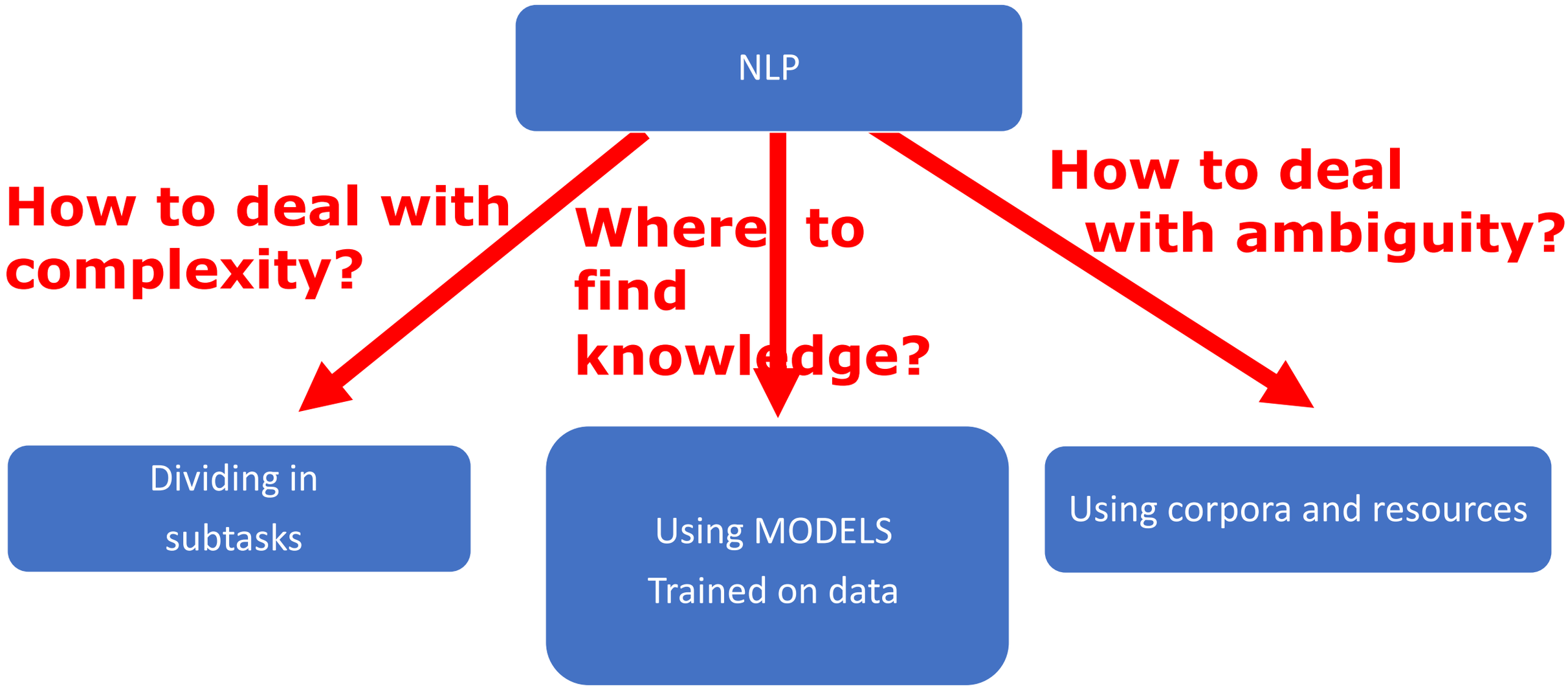
Ambiguity and models

From where can be the probability learned?

The probability can be learned directly from language, that is from a corpus (collection of texts)

How good are the probabilities depends on the corpus we used to calculate them.

The larger is the corpus, the higher is the probability that it includes all the possible words and structures of the language.


NLP
How to deal with complexity?
Where to find knowledge?
How to deal with ambiguity?
Dividing in subtasks
Using MODELS Trained on data
Using corpora and resources

1800

- Corpora used in linguistics

1950

- ARTIFICIAL INTELLIGENCE
- Turing: 1950 *Computing machinery and intelligence*
- **NLP: RULE- BASED APPROACHES**
- Chomsky: 1957 *Syntactic Structures*
- Tesnière: 1959 *Eléments de Syntaxe Structurelle*

1990 …

- **NLP: CORPUS-BASED APPROACHES**

Ambiguity and resources

Linguistic resources are particularly useful for learning all the information about how a language works and for dealing with the ambiguity that occurs in the language.

The most used resources are corpora.

The main goal of modern corpora is to help NLP systems build language **models**, not just describe languages.

Ambiguity and corpora

With the development of corpora and the NLP approaches that use them, a significant paradigm shift took place: **from formalising language to modelling it**.

To **formalise** a language L means to describe it accurately and precisely according to certain predetermined criteria.

Modelling a language L, on the other hand, means capturing the mechanisms that determine how it works. A good model allows us to interpret all the structures and phenomena that occur in L, in cases that have already been observed, but also in cases that have not been previously observed so far.

Corpora

A corpus (from Latin, plural corpora) is a **large collection of linguistic data** such as written texts or transcriptions of speech.

Much of the **NLP** currently relies on corpora, but corpora are first and foremost a creation of **applied linguistics**.

Corpora in linguistics

In linguistics, the practice of collecting and directly observing corpora has long been used to study languages, and is known as **corpus linguistics**.

A corpus is the place where evidence can be found to confirm (or deny) a variety of hypotheses about linguistic phenomena.

Corpora in linguistics

Corpora have been used by linguists to investigate languages but also to compile:

- **dictionaries**, such as the *American Heritage Dictionary of the English Language* (starting in 1969) or the *Grande dizionario italiano dell'uso* (GRADIT, starting in 1999)
- **grammar guides**, such as *A Comprehensive Grammar of the English Language (*1985)

Corpora in linguistics

Corpus linguistics consists in studying a language L as it is expressed in a collection of real texts in L.

According to corpus linguistics, reliable analysis of a language is more likely to be possible with **corpora collected in the "natural" context** in which L is used, and which avoid experimental interference. Data collected in this way are called *unrestricted*.

Corpora in NLP

Starting from the late 80s, in NLP **corpora** are used to learn the knowledge needed to deal with languages.

The rules governing a language can be indeed inferred from a corpus of that language by applying **machine learning**.

The linguistic **knowledge inferred from a corpus** associated with each linguistic object (rules, words, etc.) is associated with a **frequency** that can be used as a criterion for selecting the meaning of a word or the structure of a sentence in ambiguity cases also.

NLP and corpora

Prior to the Chomskyan Revolution and the birth of NLP, much language study was based on direct observation of data produced by speakers.

Corpus-based approaches, which aim to derive rules from large collections of data (*corpora*) that represent actual language use, have been widely used for the last two centuries.

This tradition was overshadowed for a time by Chomskyan theories in **linguistics** and especially in **NLP**, in which formalization was seen as a necessary condition for the development of systems.

NLP: a little bit of history

Nevertheless, corpus-based research continued in some circles until it revived in the last 1980s.

In particular, the pioneering work begun in 1949 by the Jesuit priest **Robert Busa**, with the support of Thomas Watson of IBM, is in this corpus-based tradition.

He developed the first machine-readable corpus: the *Index Thomisticus*, a 10 million-word corpus of medieval philosophy in Latin, encoded first on punch-cards and later with various electronic supports.

Corpora

Corpora are usually constructed so as to be **balanced** and **representative** of a particular language.

This is particularly important if a corpus is to be used for quantitative analysis: if the corpus data is skewed or unrepresentative then results of the analysis may not be reliable.

These considerations do not apply on corpora collected for the literary or historical interest of the documents they include.

Corpora

To be **representative**, a corpus is designed to include a wide coverage with respect to some specific dimension it must represent: dialect, time frame, demographic category (age, gender …) or phenomenon (morphological, syntactic, semantic …).

"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)

Corpora

To be **balanced**, a corpus must include a large variety of samples of all the different phenomena that are part of the language.

In practice, the accepted balance is determined by the intended uses of the corpus.

Corpora

Modern corpora are stored in computer memories and can be larger than ancient corpora.

Corpora tend to be of a finite **size** and to remain fixed once they have been created, with the exception of monitor corpora which are continually updated with new material.

Monitor corpora can be particularly useful for compilers of dictionaries who need to be able to track new words entering the language and the changing or declining use of old ones.

Corpora

How large the corpus should be?
The size of the corpus depends upon the purpose for which it is intended as well as on some practical considerations:

- The kind of query users want to answer using the corpus

- The methodology used by the users to study the data

- The availability of the source of data.

Corpora

Year	Name of the Corpus	Size (in words)
1960s - 70s	Brown and LOB	1 Million words
1980s	The Birmingham corpora	20 Million words
1990s	The British National corpus	100 Million words
Early 21st century	The Bank of English corpus	650 Million words

Corpora

- The **COBUILD** Bank of English (Collins Birmingham University International Language Database): the full Collins corpus currently includes 4.5 billion words, while the Bank of English is a subset of 650 million words from a carefully chosen selection of sources used for building Collins dictionaries

- The CORpus di Italiano Scritto (**CORIS**): the corpus includes around 100 million words from daily newspapers, weeklies, monthly magazines and books (fiction and non fiction).

Corpora for NLP

- In NLP corpora are the main source from where to learn linguistic knowledge
- They are usually large corpora, millions of words
- They are used to find all the statistics about words, grammatical categories, syntactic structures

Corpora and statistics

- The statistical knowledge extracted from a corpus must be organised according to a statistical model
- One of the simplest statistical model is that based on N-grams
- An N-gram is a sequence of N words

N-grams

3-gram

Mary is running in the park with her dog

Mary is running

is running in

running in the

in the park

the park with

park with her

with her dog

N-grams

3-gram

Mary is running in the park with her dog

Mary is running

is running in

running in the

in the park

the park with

park with her

with her dog

N-grams

3-gram

Mary is running in the park with her dog

Mary is running

is running in

running in the

in the park

the park with

park with her

with her dog

N-grams

3-gram

Mary is running in the park with her dog

Mary is running

is running in

running in the

in the park

the park with

park with her

with her dog

N-grams

3-gram

Mary is running in the park with her dog

Mary is running

is running in

running in the

in the park

the park with

park with her

with her dog

N-grams

3-gram

Mary is running in the park with her dog

Mary is running

is running in

running in the

in the park

the park with

park with her

with her dog

N-grams

3-gram

Mary is running in the park with her dog

Mary is running

is running in

running in the

in the park

the park with

park with her

with her dog

Bi-grams are all pairs of adjacent words that occur in the sentences of the corpus.

For each bi-gram a **frequency** cab be extracted that provides information about the probability that the words in the bi-gram are syntactically related.

During the training of the NLP system the **model** of the language in the corpus is built: all bi-grams are collected and associated with their frequency.

Annotated corpora

Corpora can be:

- **Not annotated**: collections of linguistic data in electronic format
- **Annotated**: collections of linguistic data in electronic format where some form of knowledge is made explicit and properly associated with each datum.

Linguists have differing views about the annotation of a corpus, ranging from minimal to very detailed annotation.

In **NLP** corpora are usually annotated according to the task for which they must be used.

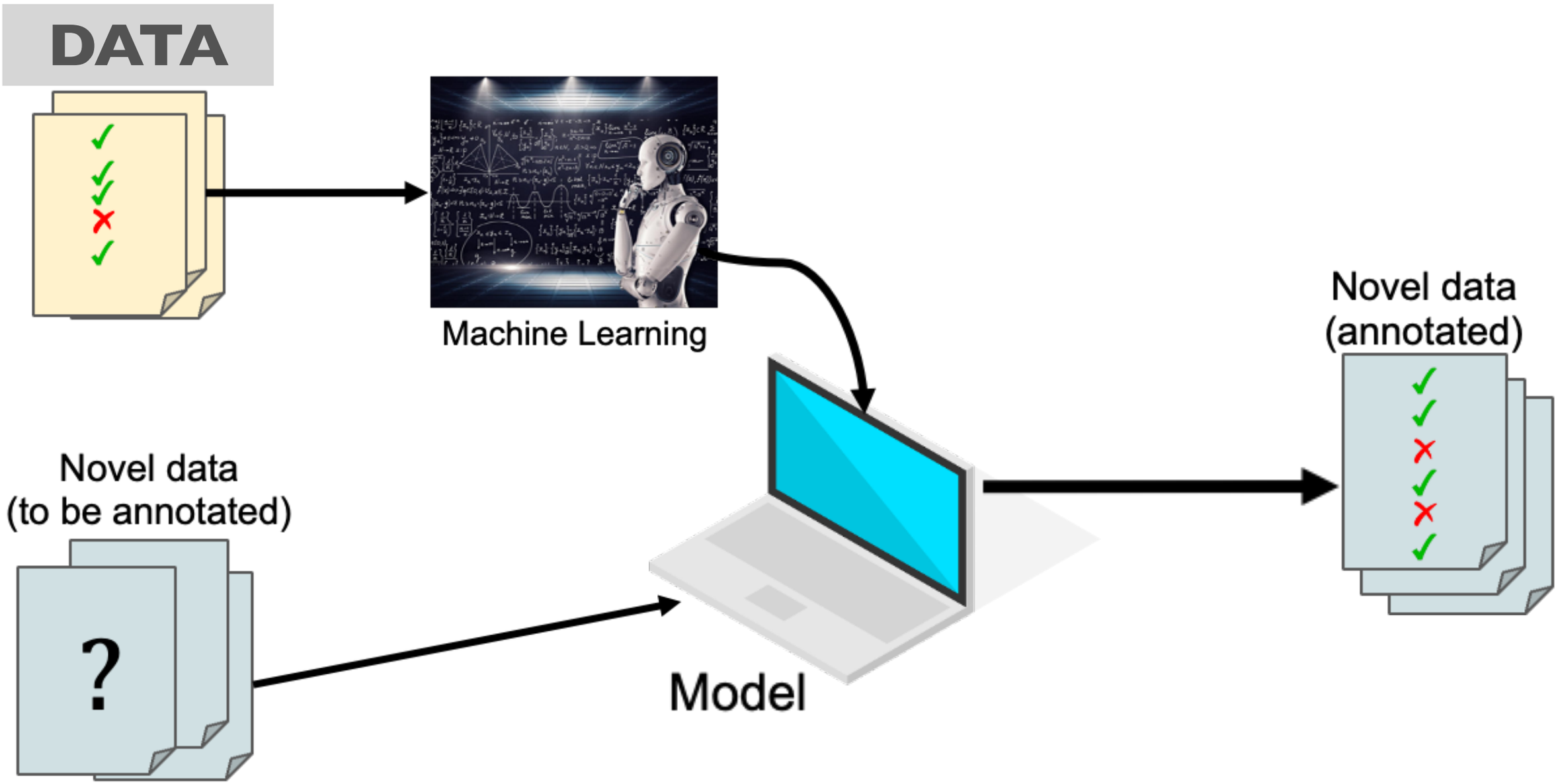
NLP and annotated corpora

Annotated corpora are extensively used in NLP. Their usefulness is in direct proportion to the wealth of information they contain in their annotation.

However the annotation is a time-consuming task that has not usually been applied to very large corpora.

The annotation makes explicit some form of knowledge that is implicit in linguistic data. Machine learning can be applied to extract and generalise that knowledge.

How machine learning works?



Annotate or not annotate?

An example of (very small, 4 sentences) English corpus about dogs:
the doggy corpus

- the dog is in the garden
- the dog is barking
- dogs are sleeping in the garden
- the dog chased the cat

Annotate or not annotate?

The *doggy corpus* can be used as it is, **UNANNOTATED.**

Here the corpus is in the plain form, without annotation:

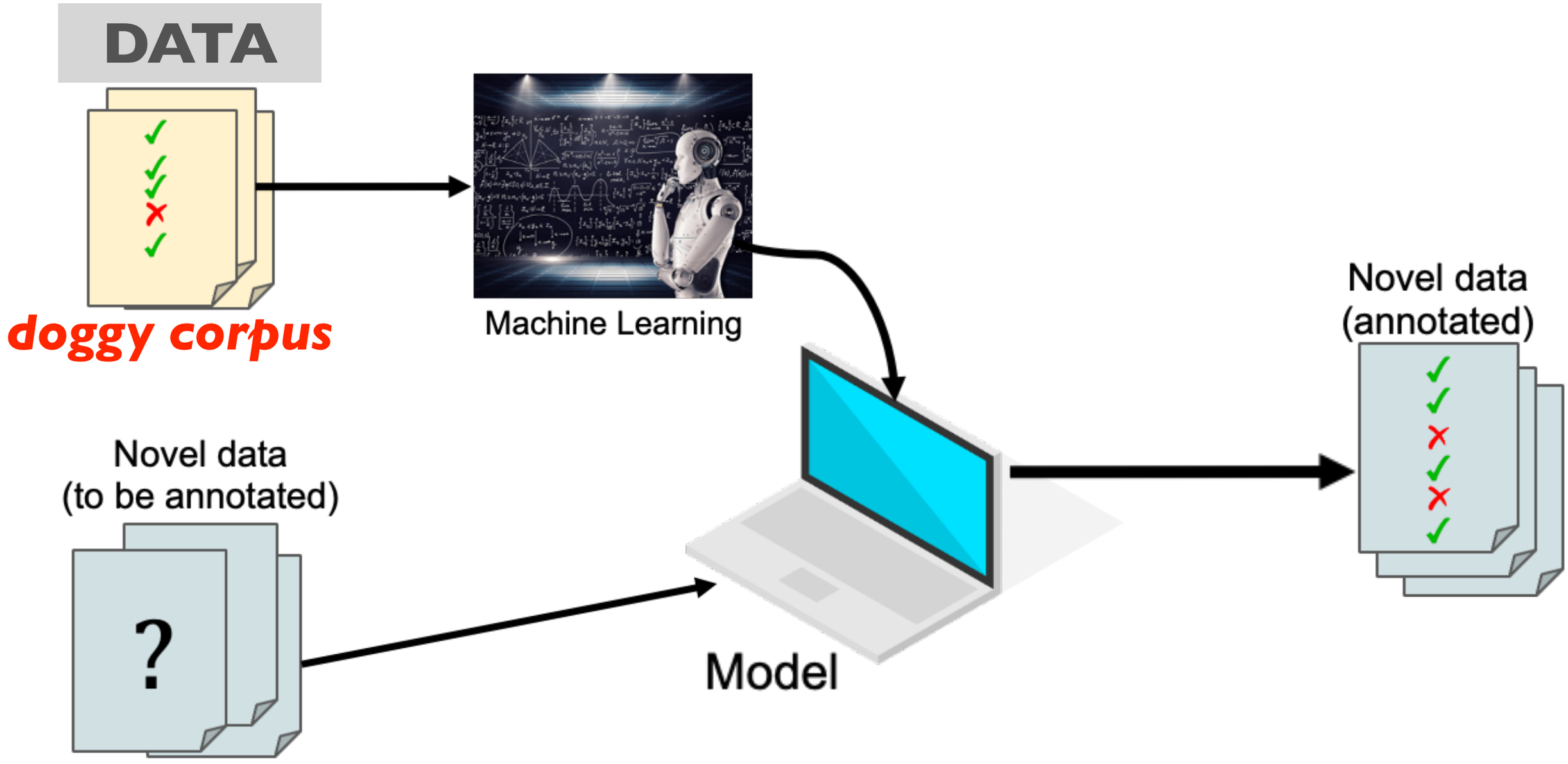
- the dog is in the garden
- the dog is barking
- dogs are sleeping in the garden
- the dog chased the cat

Annotate or not annotate?

Here in ***the doggy*** corpus all the grammatical categories have been **ANNOTATED,** made explicit with the red labels associated to the words:

- the**[art]** dog**[noun]** is**[verb]** in**[prep]** the**[art]** garden**[noun]**
- the**[adj]** dog**[noun]** is**[verb]** barking**[verb]**
- dogs**[noun]** are**[verb]** sleeping**[verb]** in**[prep]** the**[art]** garden**[noun]**
- the**[art]** dog**[noun]** chased**[verb]** the**[art]** cat**[noun]**

We train machine learning on the doggy corpus



We can apply to the **UNANNOTATED** doggy corpus machine learning using a very simple statistical approach based on bi-grams. For instance the bi-grams for

- the dog is in the garden
are

the dog

dog is

is in

in the

the garden

We can apply to the **ANNOTATED** doggy corpus the same machine learning method using a very simple statistical approach based on bi-grams. For instance the bi-grams for

- the**[art]** dog**[noun]** is**[verb]** in**[prep]** the**[art]** garden**[noun]**

are

the**[art]** dog**[noun]**

dog**[noun]** is**[verb]**

is**[verb]** in**[prep]**

in**[prep]** the**[art]**

the**[art]** garden**[noun]**

The bi-grams of the whole **UNANNOTATED** doggy corpus are:

the dog, dog is, is in, in the, the garden

the dog, dog is, is barking

dogs are, are sleeping, sleeping in, in the, the garden

the dog, dog chased, chased the, the cat

The bi-grams of the whole **ANNOTATED** doggy corpus are:

the**[art]** dog**[noun]** **,** dog**[noun]** is**[verb]** **,**
is**[verb]** in**[prep]** **,** in**[prep]** the**[art]** **,**
the**[art]** garden**[noun]**

the**[art]** dog**[noun]** **,** dog**[noun]** is**[verb]** **,**
is**[verb]** barking**[verb]**

dogs**[noun]** are**[verb]** **,** are**[verb]** sleeping**[verb]** **,**
sleeping**[verb]** in**[prep]** **,** in**[prep]** the**[art]** **,**
the**[art]** garden**[noun]**

the**[art]** dog**[noun]** **,** dog**[noun]** chased**[verb]** **,**
chased**[verb]** the**[art]** **,** the**[art]** cat**[noun]**

What knowledge can be learned from the **UNANNOTATED** doggy corpus?

the dog,	dog is,	is in,	in the,	the garden
the dog,	dog is,	is barking		
dogs are,	are sleeping,	sleeping in,	in the,	the garden
the dog,	dog chased,	chased the,	the cat	

What knowledge can be extracted from the **ANNOTATED** corpus?

the[art] dog[noun] , dog[noun] is[verb] ,
is[verb] in[prep] , in[prep] the[art] ,
the[art] garden[noun]

the[art] dog[noun] , dog[noun] is[verb] ,
is[verb] barking[verb]

dogs[noun] are[verb] , are[verb] sleeping[verb] ,
sleeping[verb] in[prep] , in[prep] the[art] ,
the[art] garden[noun]

the[art] dog[noun] , dog[noun] chased[verb] ,
chased[verb] the[art] , the[art] cat[noun]

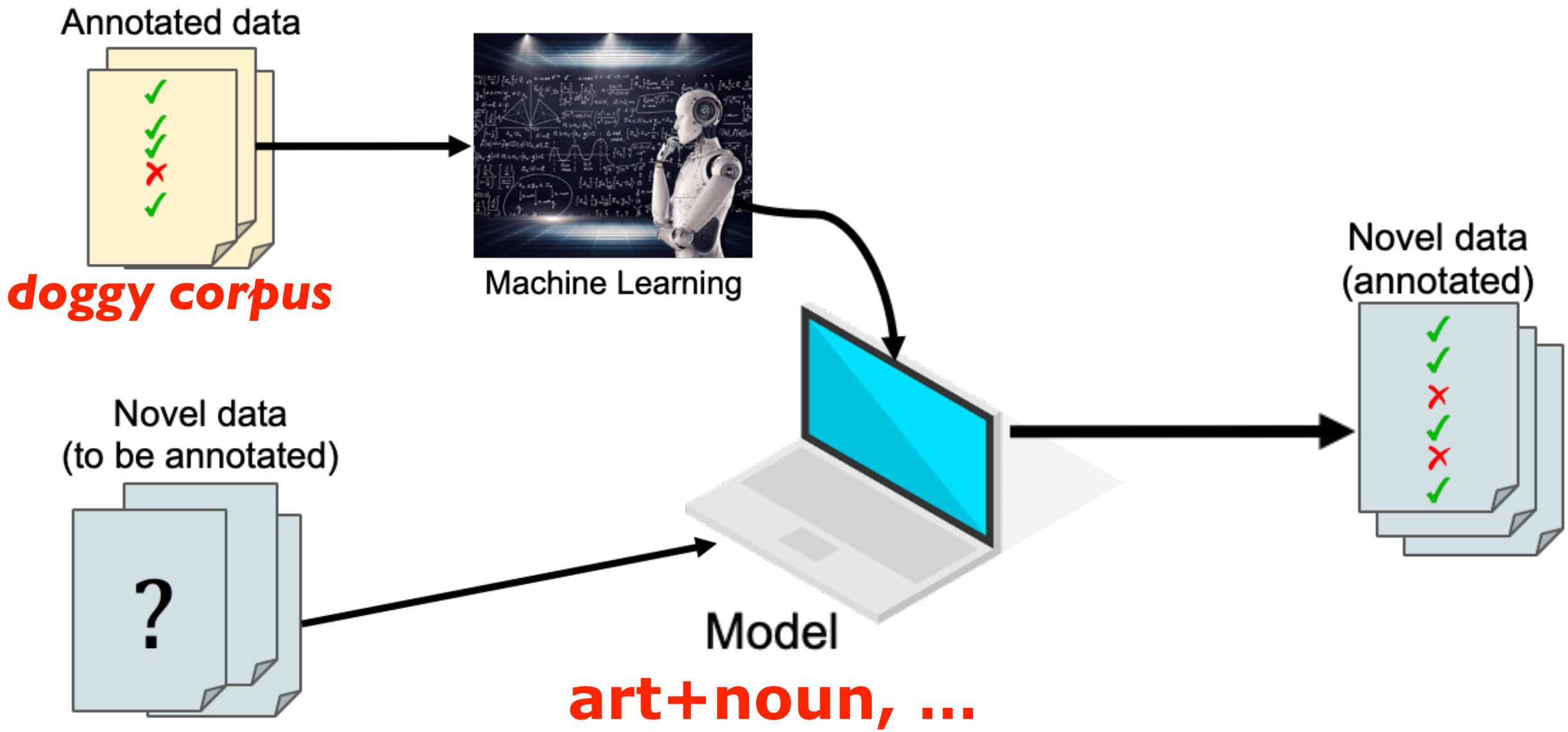
Some example of knowledge that can be extracted from the **UNANNOTED** doggy corpus using bi-grams:

- The word “the" often co-occurs with the word "dog"
- How often?
  - In 3 over 6 cases (bi-grams) the word "the" is followed by the word "dog" > 50%

Some example of knowledge that can be extracted from the **ANNOTATED** doggy corpus using bi-grams:

- Articles and nouns seem to be linked by some relation since they often co-occur
- How often?
  - In 6 over 6 cases (bi-grams) the article is followed by the noun > 100%
  - In 6 over 7 cases the noun is preceded by the article > 90%

We train machine learning on the doggy corpus



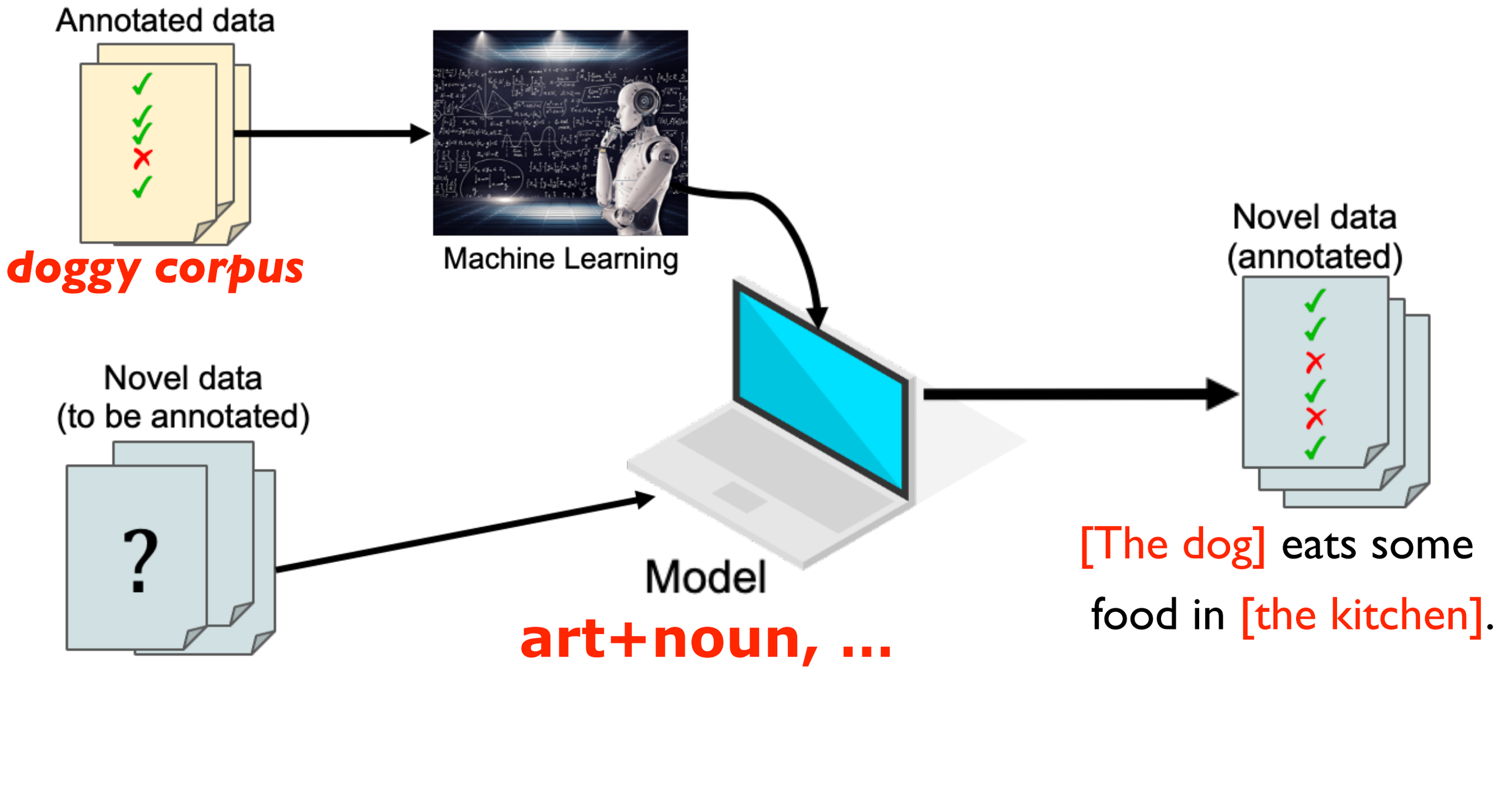
After the training, the **model** is used for the analysis of new sentences.

The dog eats some food in the kitchen.

The model allows the system to decide that article and noun form together a unit (called noun phrase) within the sentence

[The dog] eats some food in [the kitchen].

We train machine learning on the doggy corpus



Lessons learned from the doggy corpus:

- **a model of language depends on the corpus used and that a different corpus allows machine learning to build different models**

- **the annotation provides more precise information about the behaviour of words**

- **the annotation reduce the sparsity of data and allows us to cluster the knowledge**

What is a linguistic resource for NLP?

An organized set of linguistic objects associated with an explicit encoding of some linguistic knowledge.

It can be a collection of sentences, of words, of meanings, of expressions,
because to deal language
we always need knowledge about language and about the relationships linking words each others and with world.

Resources and events

An increasing portion of researches in NLP area has been devoted to linguistic resources.
This is attested by several dedicated events:

- the **Language Resources and Evaluation Conference** (LREC)
- the main NLP conferences: the Meeting of the Association for Computational Linguistics (ACL), the Meeting of the European Chapter of the Association for Computational Linguistics (EACL), the Empirical Methods in NLP (EMNLP), Computational Linguistics conference (COLING).

Every year in these events, novel resources were introduced and the exiting ones exploited for putting forward the state-of-the-art of NLP.

Resources for NLP

The complexity of knowledge involved in linguistic resources makes their development a very **difficult and time-consuming task**, in particular when the annotation includes different layers of knowledge and/or fine-grained annotations.

This also depends on the fact that a resource can be only partially built by an **automatic tool** (the tool that can later take advantage of it!) and usually requires careful **hand-made validation** to be released

Resources for NLP

A resource is the **result of a team's work**, nobody can develop a resource alone, also because annotation usually represents the knowledge of a whole speakers' community (> performance!).

The quality and usefulness of a developed resource can be only improved by its **usage by a large community** of users during years, but the usability and portability of a resource can be limited by several factors and the adequacy to a **standard** (when a standards exist!) is a hot issue.