CLARIN ERIC

# Open Language Resources for Smarter Artificial Intelligence

Kaja Dobrovoljc

**The ESFRI 20th anniversary conference**

Paris, 25th March 2022

# A Short Introduction

- Linguist / post-doc researcher
  - University of Ljubljana
  - Jozef Stefan Institute
- Language resources and technologies for Slovenian
  - design, development and applications in applied linguistics
- Keen user of **CLARIN** research infrastructure

University *of Ljubljana*

Jožef Stefan Institute

# CLARIN

- **Common Language Resources and Technology Infrastructure**
- On ESFRI roadmap since 2006, ERIC status since 2012, Landmark since 2016

- Provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
  - to **digital language data** (in written, spoken or multimodal form)
  - and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
  - through **a single sign-on** environment
- Serves as an ecosystem for **knowledge sharing and training**
- Is one of the European RIs in the SSH Cluster (aka SCI)

# CLARIN Today

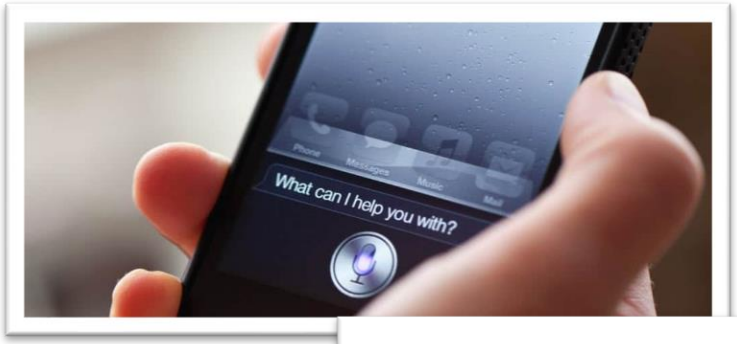
- a distributed network of **70 centres**
- **22 members**: AT, BE, BG, CY, CZ, DE, DK, EE, FI, GR, HR, HU,IS, IT, LT, LV, NL, NO, PL, PT, SE, SI
- **2 observers:** UK, ZA
- **1 third party**

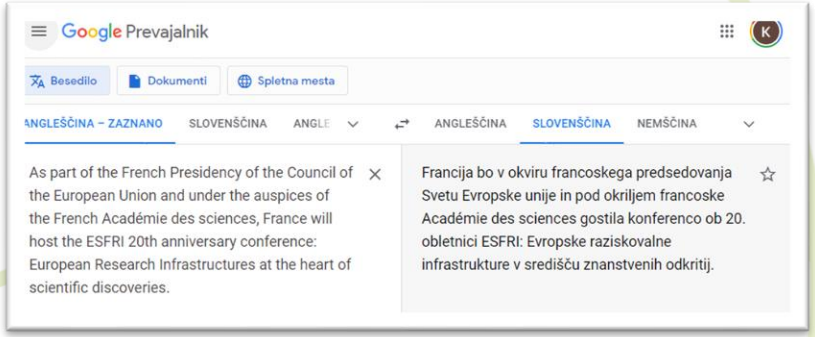

CLARIN.SI

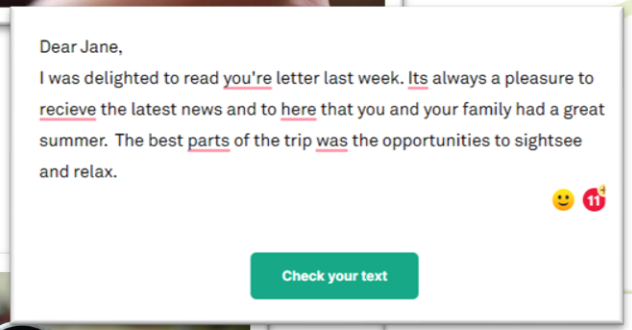
- Slovene CLARIN **national consortium** since 2014
- **Language resources, services and tools for Slovene** and other South-Slavic languages

# Language Technology

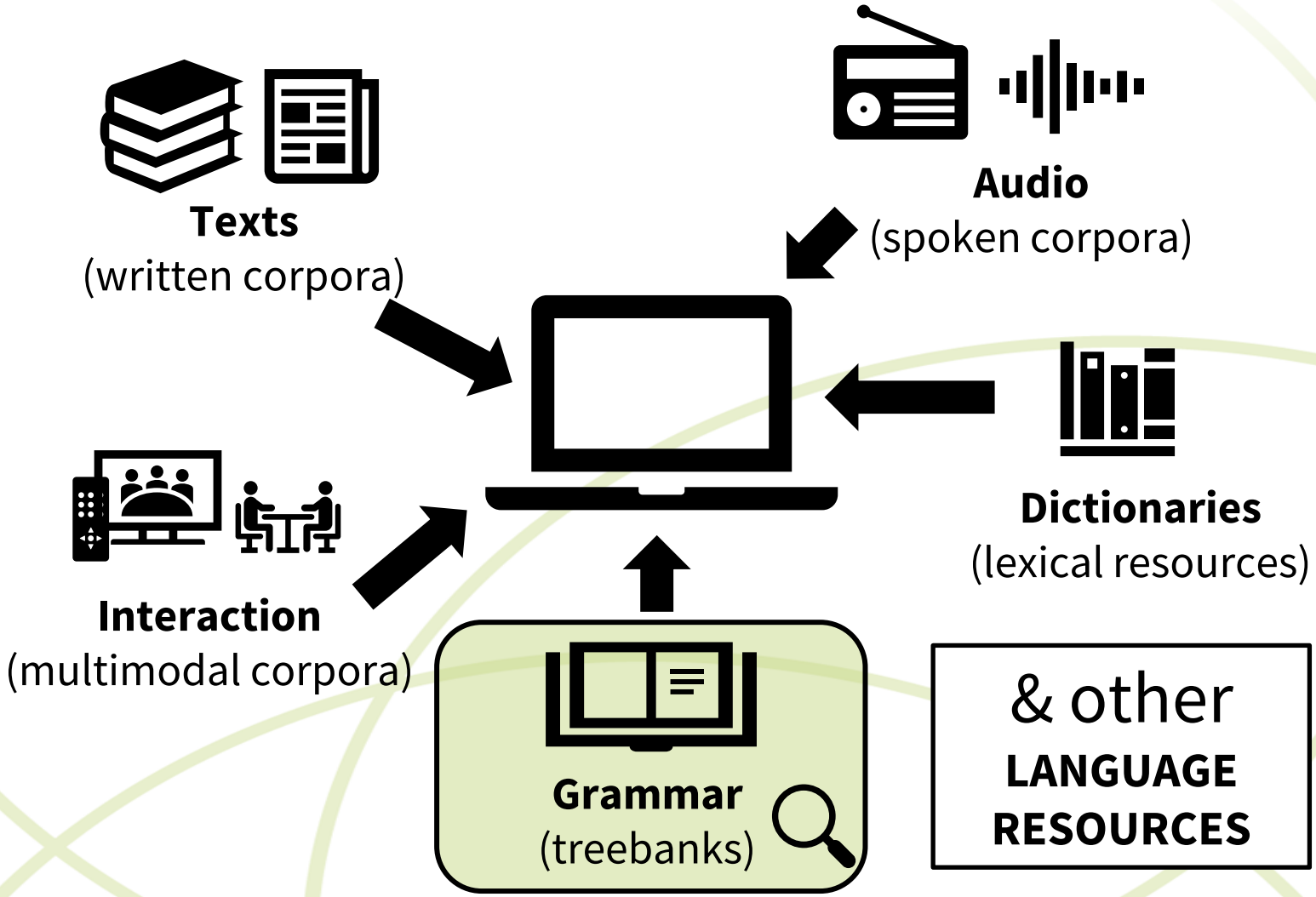
- A branch of Artificial Intelligence that enables machines to understand human language.



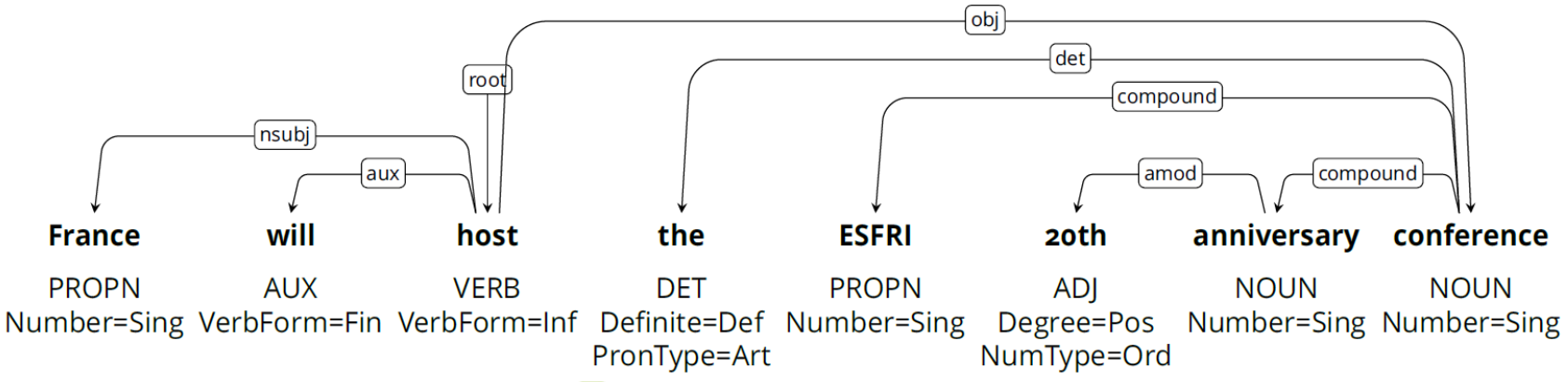




# How do Machines Learn a Language?



# Treebanks

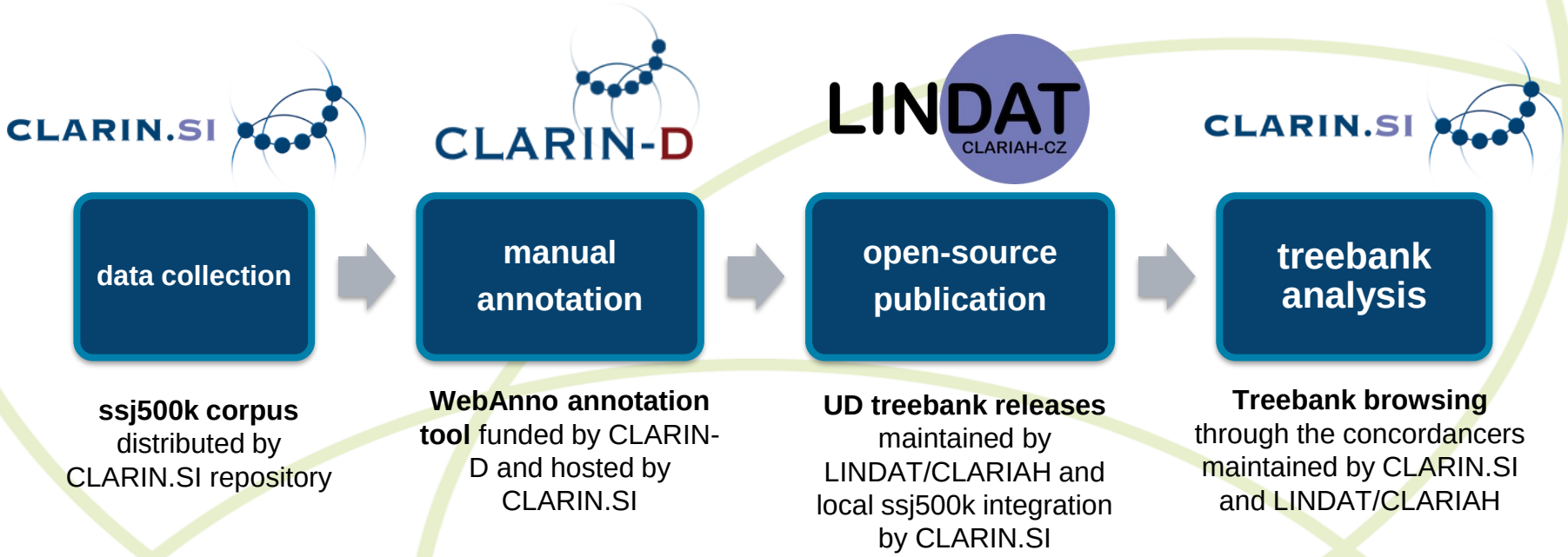
- Collections of **grammatically annotated sentences**
- **Syntactic structure** of a sentence represented as a tree




- Useful for:
  - **linguistics:** development of grammatical theories
  - **parsing technologies:** automated large-scale annotation and downstream applications
  - **language-based research in general:** complex data investigations

# Developing a Treebank for Slovenian

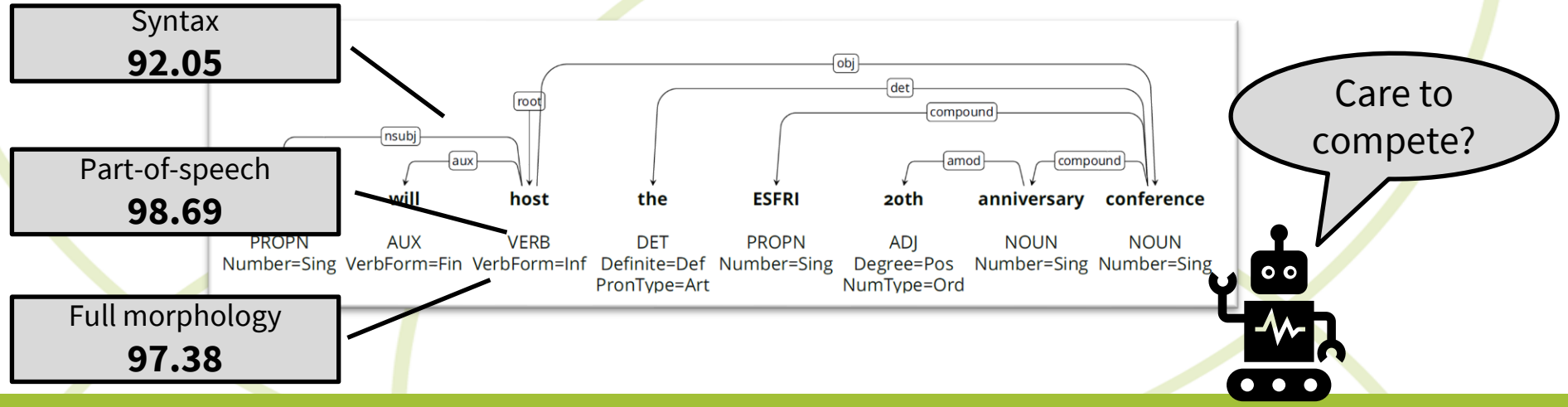
- **SSJ-UD:** reference treebank for Slovenian, cross-linguistically interoperable annotation scheme (UD)
- A lengthy process dependent on CLARIN-maintained **data, services and tools**



# Treebank in Action

- Used in modelling several **state-of-the-art parsing tools** for multilingual natural language processing

- Including the CLASSLA-Stanza tool developed by the **CLARIN Knowledge Centre for South Slavic languages**



# Conclusion

- **Open language resources are essential to scientific advances** in multilingual artificial intelligence and other language-related disciplines.
- The development, distribution and exploitation of these resources has been made significantly easier by the ESFRI ERIC **Common Language Resources and Technology Infrastructure (CLARIN)**.
- A case in point: **A decade of rapid advances in Slovenian language technologies** based on CLARIN(.SI)-maintained data, services and tools.

- It is not just about the infrastructure, but also **the people behind it**.


# Thank you!