2nd General Meeting
University of Naples “L’Orientale” Naples, Italy,
8-9 February 2024
https://unidive.lisn.upsaclay.fr/

SAPIENZA NLP

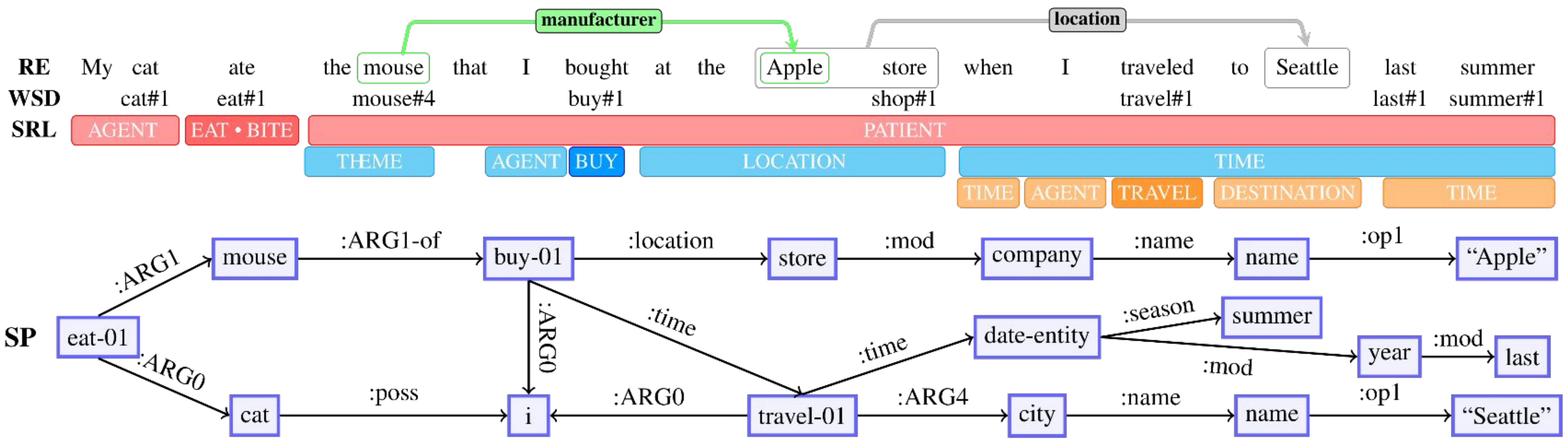
Babelscape

Creating a Multilingual Wide-Coverage Multi-Layered Semantically Annotated Corpus

Simone Conia, Edoardo Barba, Abelardo Carlos Martinez Lorenzo,
Pere-Lluís Huguet Cabot, Riccardo Orlando, Luigi Procopio, Roberto Navigli
Sapienza University of Rome & Babelscape

WG3: Multilingual and cross-lingual language technology

Parallel Annotations for 4 semantic tasks


manufacturer
location
RE My cat ate the mouse that I bought at the Apple store when I traveled to Seattle last summer
WSD cat#1 eat#1 mouse#4 buy#1 shop#1 travel#1 last#1 summer#1
SRL AGENT EAT • BITE PATIENT
THEME AGENT BUY LOCATION TIME
TIME AGENT TRAVEL DESTINATION TIME
SP
eat-01
:ARG1
mouse
:ARG1-of
buy-01
:location
store
:mod
company
:name
name
:op1
“Apple”
:ARG0
cat
:poss
i
:ARG0
:time
travel-01
:time
date-entity
:season
summer
:mod
year
:mod
last
:ARG4
city
:name
name
:op1
“Seattle”

Dataset Creation

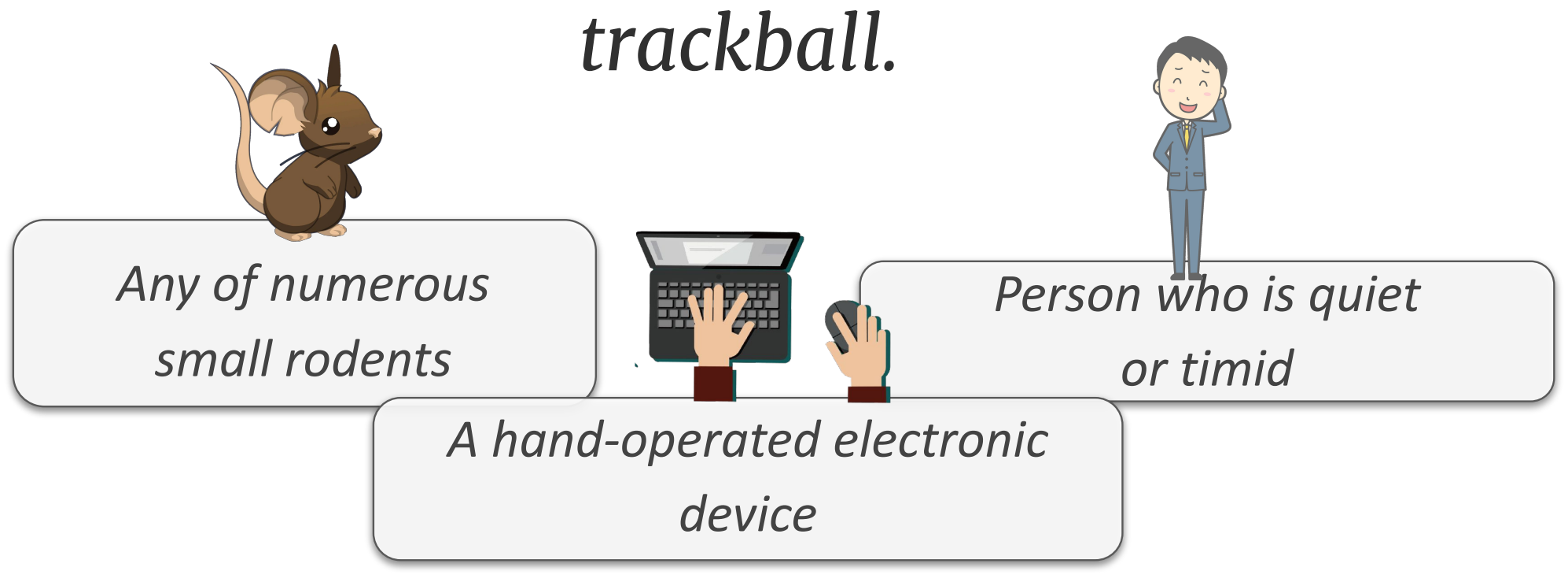
Wikipedia

English Italian German Spanish French

- Preserve pages that exist in all 5 languages
- 440,000 articles per language
- More than 7M sentences per language

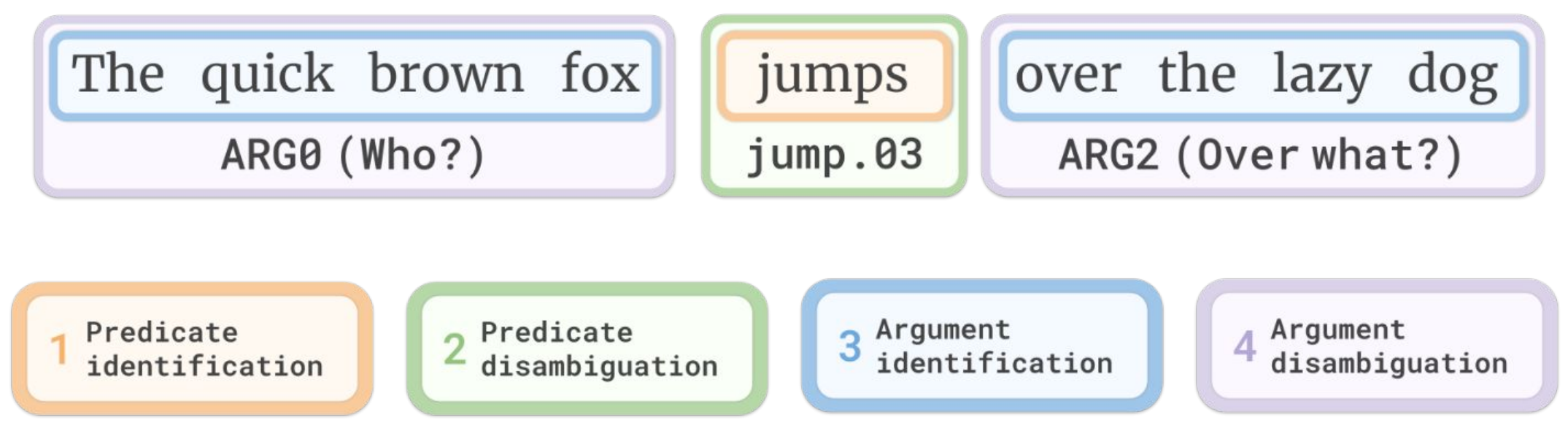
Word Sense Disambiguation (WSD)

A <u>mouse</u> takes much more space than a trackball.


Any of numerous small rodents
Person who is quiet or timid
A hand-operated electronic device

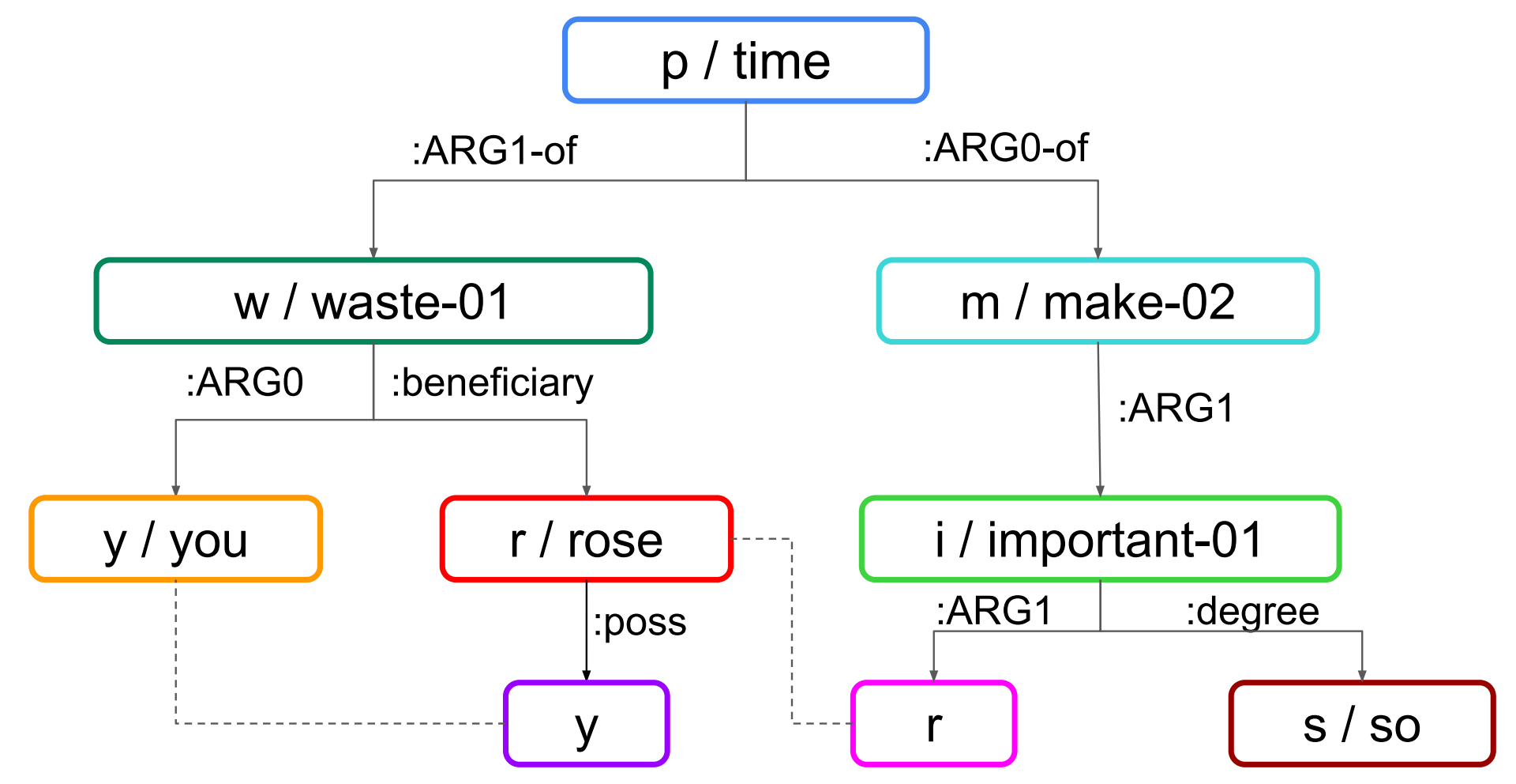
- Assigning a word in context its most appropriate meaning from a predefined sense inventory
- We use ESCHER (Barba et al., 2021) to tag the corpus.
- BabelNet 5 as the default sense inventory

Semantic Role Labeling (SRL)


The quick brown fox
ARG0 (Who?)
jumps
jump.03
over the lazy dog
ARG2 (Over what?)
1 Predicate identification
2 Predicate disambiguation
3 Argument identification
4 Argument disambiguation

- Automatically answer: Who did What to Whom, Where, When and How?
- We adopt Multi-SRL (Conia and Navigli, 2020) for tagging.
- We use two predicate inventories VerbAtlas and Propbank.

Semantic Parsing (SP)



EN It is the time you have wasted for your rose that makes your rose so important

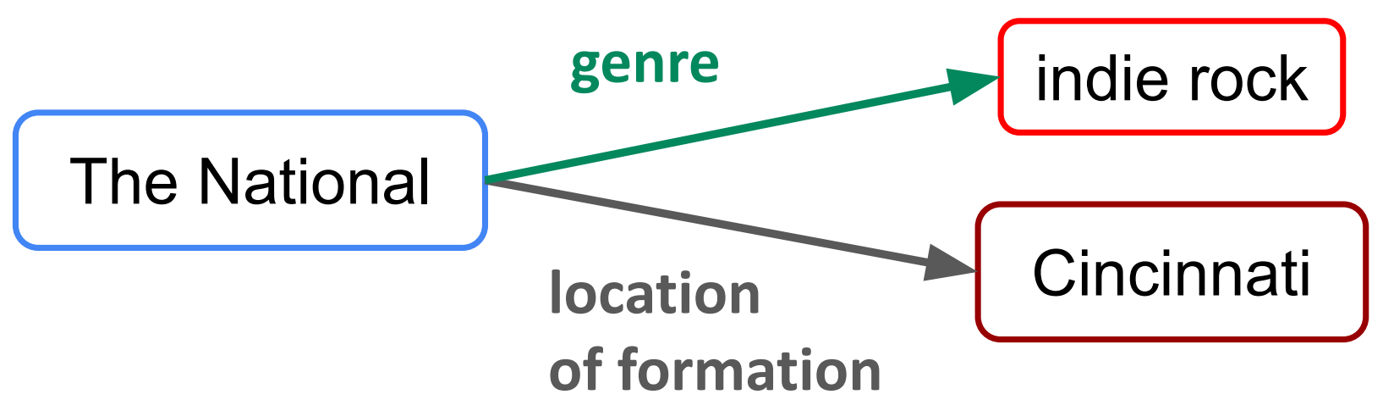
ES El tiempo que perdiste por tu rosa hace que tu rosa sea tan importante

IT È il tempo che tu hai perduto per la tua rosa che ha fatto la tua rosa così importante

- Encoding the meaning of a sentence into a machine-interpretable structure.
- We focus on Abstract Meaning Representation (AMR).
- We train our own multilingual parsers based on SPRING (Bevilacqua et al., 2021)

Relation Extraction (RE)

I <u>The National</u> sono un gruppo musicale <u>indie rock</u> formatosi a <u>Cincinnati</u>.



- Extract semantic relationships between entities from unstructured text.
- We use a two-step approach mixing both predictions by mREBEL Huguet Cabot et al., 2023) and a pipeline leveraging hyperlinks from Wikipedia and WSD predictions.
- The relation types set is taken from Wikidata.

Insights and Opportunities

Our annotations will allow for open-access high quality data for training WSD, SRL, SP and RE systems

Our work can serve as a benchmark for multilingual WSD by leveraging Wikipedia hyperlinks

Our work will allow to explore the interconnection between these tasks

Do SRL predicate senses coincide with WSD senses?

What are the interactions between SRL and SP structures

Can we leverage WSD senses to have richer concept-based RE

FUTURE AI RESEARCH
PNRR MUR project
PE0000013-FAIR

COST EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

UniDive

Funded by the European Union