

# Building a Language Analysis Portal with Galaxy

Emanuele Lapponi,<sup>1</sup> Erik Velldal,<sup>1</sup> Nikolay Vazov,<sup>2</sup> Stephan Oepen<sup>1</sup>

(1) Language Technology Group, Department of Informatics, University of Oslo

(2) Research Support Services Group, University Center for Information Technology, University of Oslo



## WHAT?

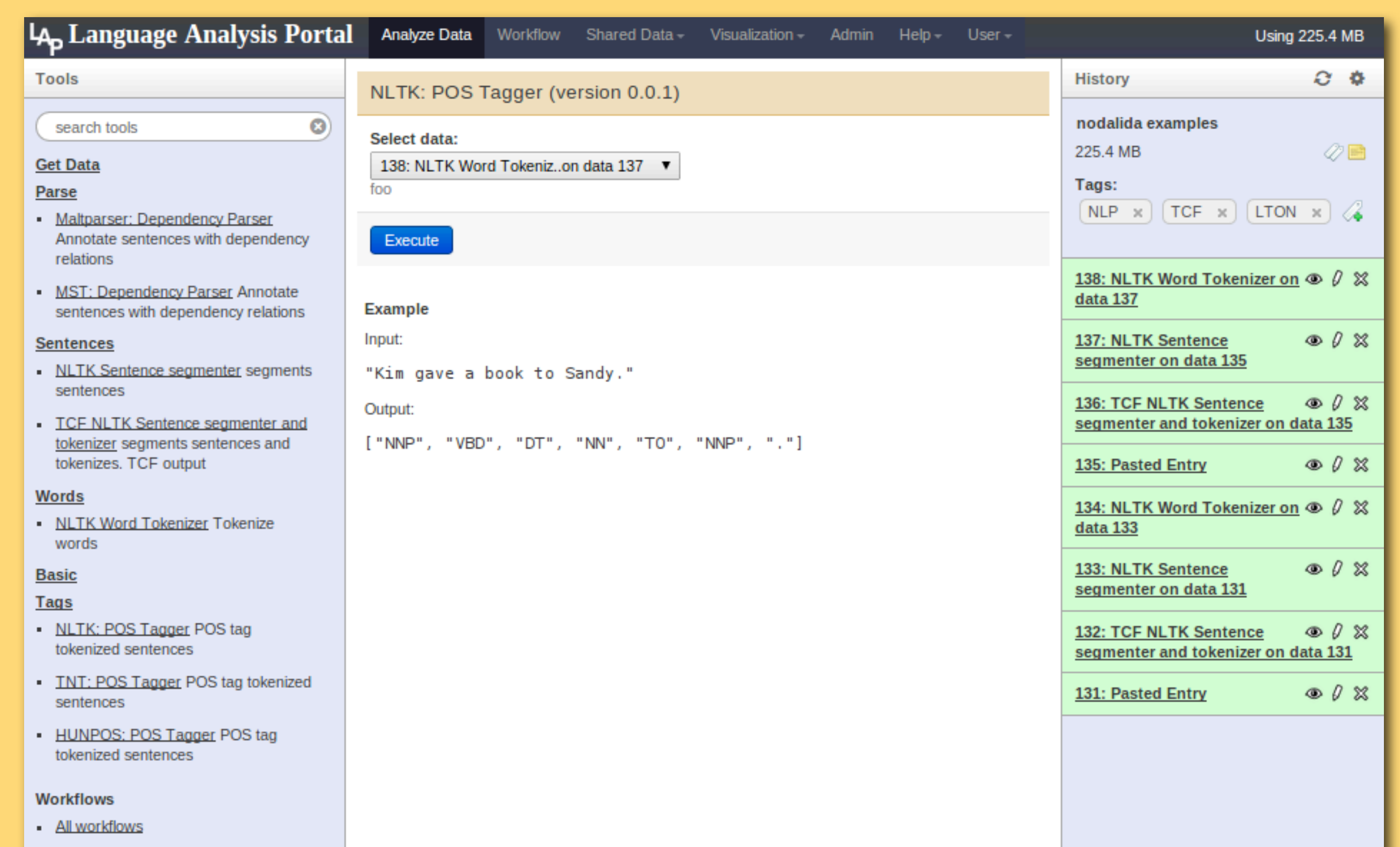
- LAP: Language Analysis Portal
- A platform-independent, web-based portal for natural language processing (NLP)
- A workflow-centric Galaxy: NLP tools typically depend (often multiple) upstream annotators: LAP will host a varied set of tools that can be chained to form complex workflows
- Galaxy will pass jobs to a High Performance Computing (HPC) Linux cluster
- Part of the Norwegian branch of the pan-European CLARIN infrastructure initiative

## WHY?

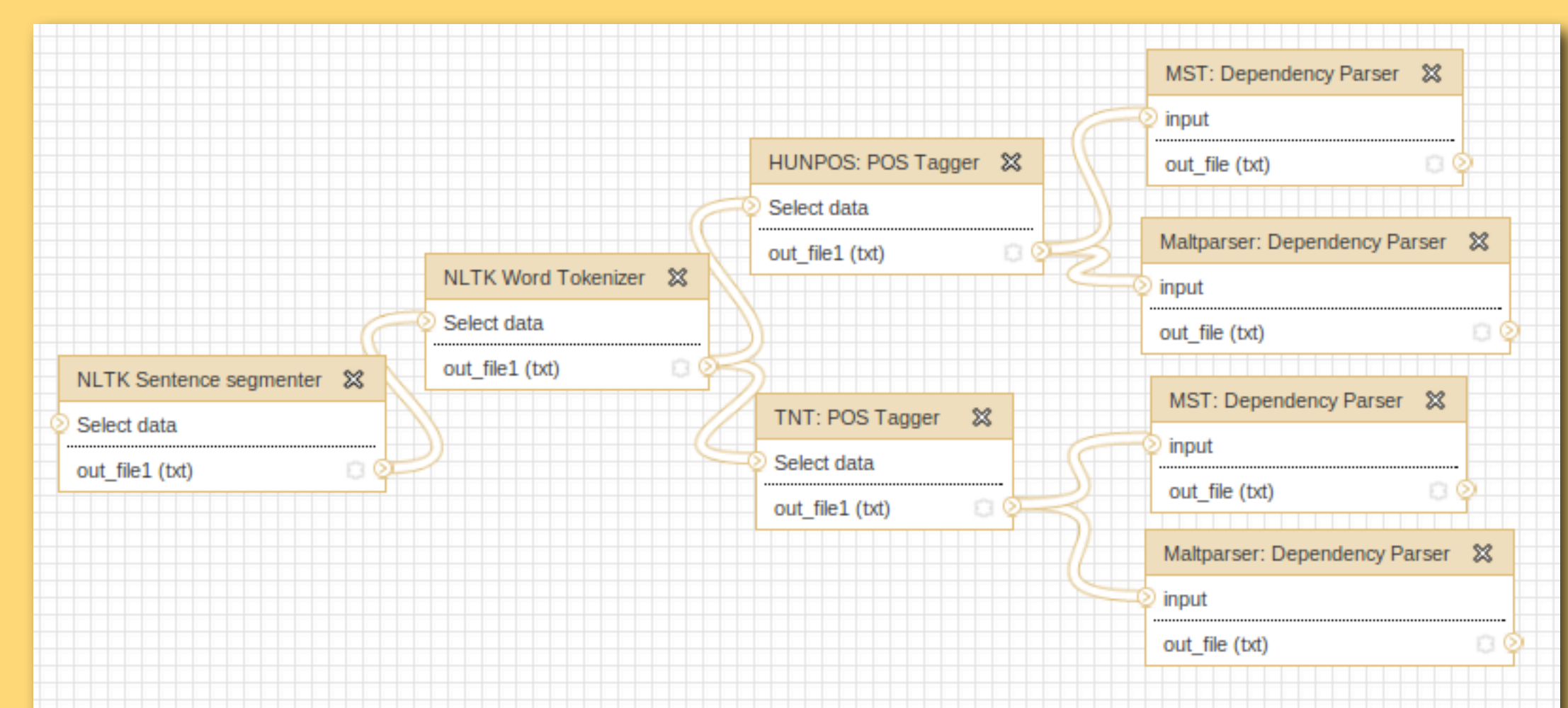
- Boost the availability and usability of large-scale language analysis within and outside the field
- Reach out to researchers from less technically oriented disciplines
- Enable HPC-powered experimentation for users who might not have access to the computing power necessary to process larger data sets
- Enable language technology-based research in the humanities and social sciences
- Facilitate reproducibility of experiments
- Transfer of knowledge from other Galaxy portals maintained at the University of Oslo, such as Bioportal and The Genomic HyperBrowser

## HOW?

- Several expected extensions to Galaxy:
- Completely new suit of tools, targeting NLP
- Predefined workflows for layered LT tasks such as syntactic parsing and document classification
- All tools chained together through a standardized interchange format (LAF / GrAF)
- Federated identity management (through Feide)
- Connection to an HPC backbone
- Built in awareness of tool dependencies (e.g. required upstream processing) in the workflow manager



Screenshot of LAP Galaxy.



A workflow with four endpoints.

<http://www.mn.uio.no/ifi/english/research/projects/clarino/>