The Language Analysis Portal

🎨🏗️💡 Design, implementation and use

Emanuele Lapponi (he/him)
Language Technology Group
Institute for Informatics
University of Oslo
`@emanlapponi`

💡 The digital Social Sciences and Humanities (SSH)

- Interest in computational methods has grown substantially in the last decade
- One strain of digital SSH: research built on "more data than you can read"
- NLP tools are becoming an integral part of research outside of CS and Linguistics

CLARIN

- Develop an infrastructure to facilitate NLP ❤️ SSH research
- Part of the CLARINO mandate: build a portal for NLP applications focusing on users with no NLP background

🤔 Undertaking such a project raises questions:

- How should a system the facilitates their use be designed?
- Can off-the-shelf tools be integrated?
- What's a good/scalable/coherent way to do it?
- And what's the point?

Part 1

Political Science & Text-as-data 💡

💡 Classifier evaluation scores as a quantity of interest

- In a parliamentary setting, the contents of a speech should reflect the ideology of the speaker.
- Use party classification scores for parliamentary analysis!

For example: Measure polarization

- Higher accuracy → more polarized*
- Hirst et al** do it for the Canadian Parliament
    - and find that position is a much stronger signal than policy!

...Or ask more involved questions

- Are newcomers to the EU parliament joining groups based on ideology or convenience?***

* Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems

** Party Status as a Confound in the Automatic Classification of Political Speech by Ideology.

*** Lost in Translation? Predicting Party Group Affiliation from European Parliament Debates

Text-as-data pre-processing cookbook

- 🏅 Grimmer and Stewart (2013)* survey the text-as-data field, warning of the dangers of "one-size-fits-all" experiments
- 🤔 They also introduce a recipe for pre-processing, with emphasis on stemming and normalization
- 🙄 Dismissal of expensive techniques that "do little to enhance performance", e.g. context, disambiguation, lemmatization and so on

* Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts

Text-as-data pre-processing cookbook

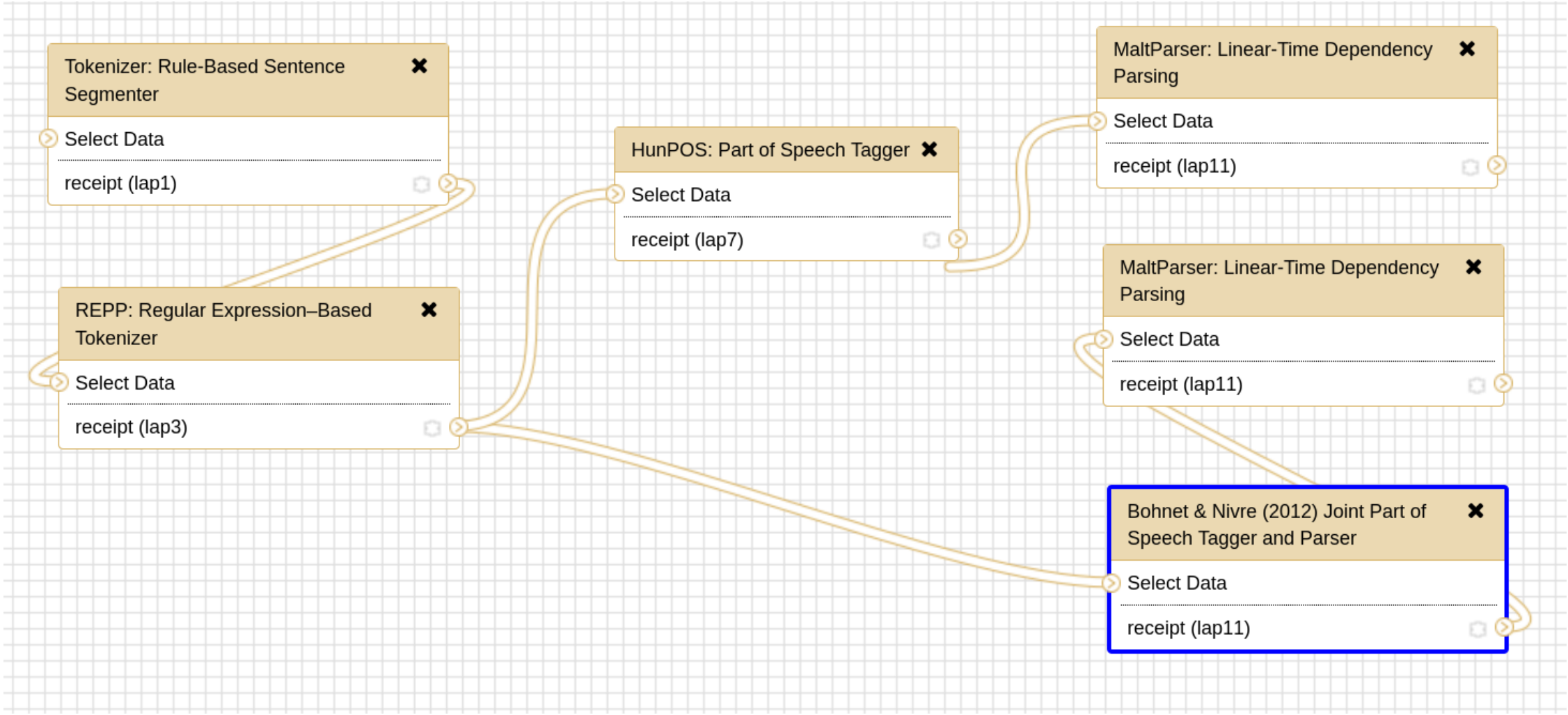
- The recipe becomes gospel
- Even attempts to address the effects of pre-processing on political analysis provide alternate recipes using the same ingredients*
- Experiments miss out on potentially useful (and contentfully significant) tools and techniques!

* Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It

Part 2

LAP & LXF 🎨🏗️

The Language Analysis Portal



Lap eXchange Format

- Scalable to linguistic annotations beneath and above the token level
- Applicable to a wide a range of linguistic phenomena and text annotations
- Scalable to large data volumes

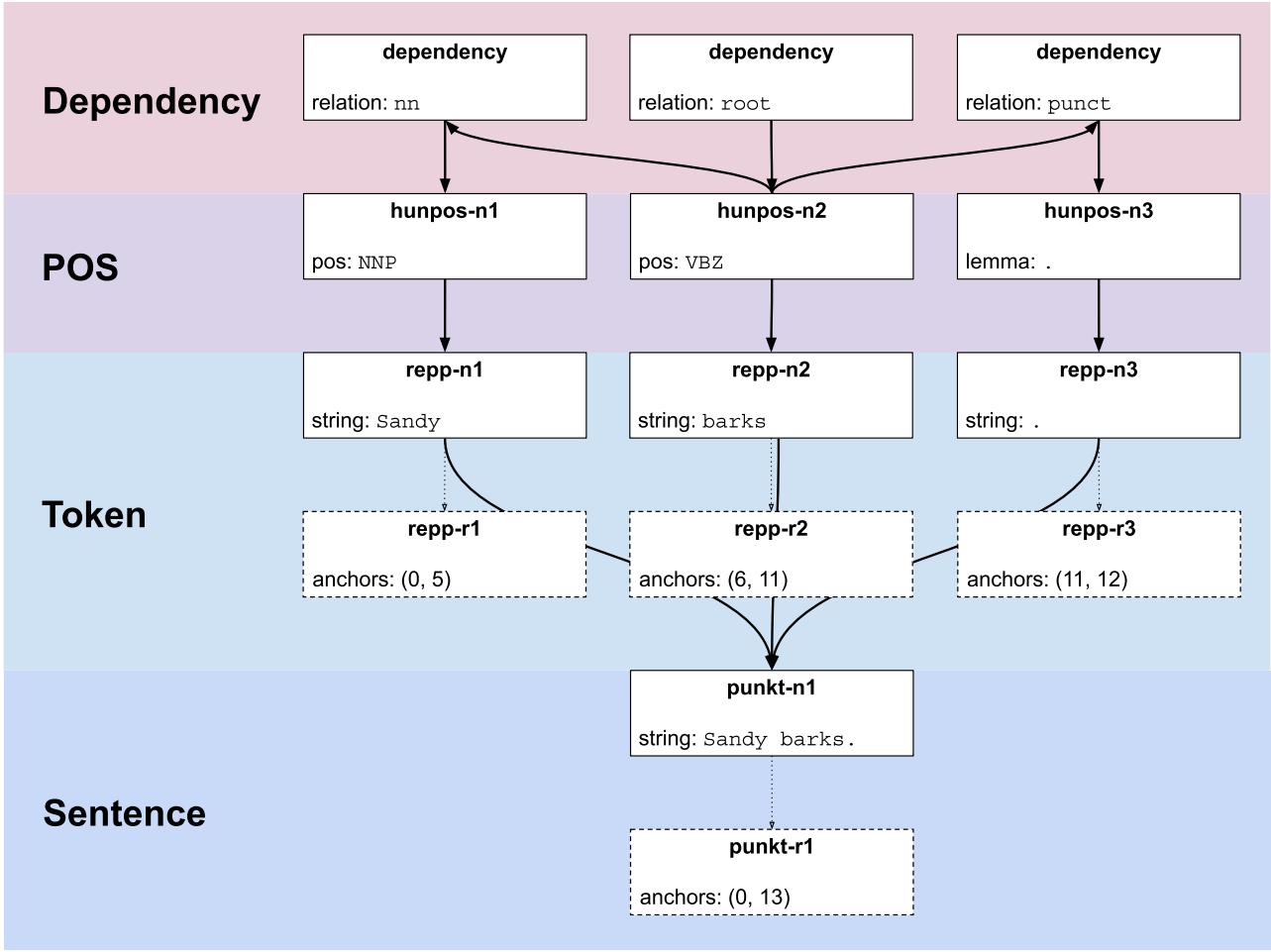
The Language Annotation Framework*

An abstract data model: represent arbitrary text **annotations** in a graph. **Nodes** and **edges** organize the structure and the relations of the annotations, and refer to **regions** in the original data.

- Purpose: full interoperability across annotation formats.
- Machine readable instantiations of the data model exist, but are focused on *representation* rather than *interchange*.

* The Linguistic Annotation Framework. A Standard for Annotation Interchange and Merging

- **LXF** builds on and expands LAF, focusing on intermediate graph representations and building a graph incrementally
- Achieves scalability, both representational and computational, while staying faithful to the original design



Part 4

The Talk of Norway 💡

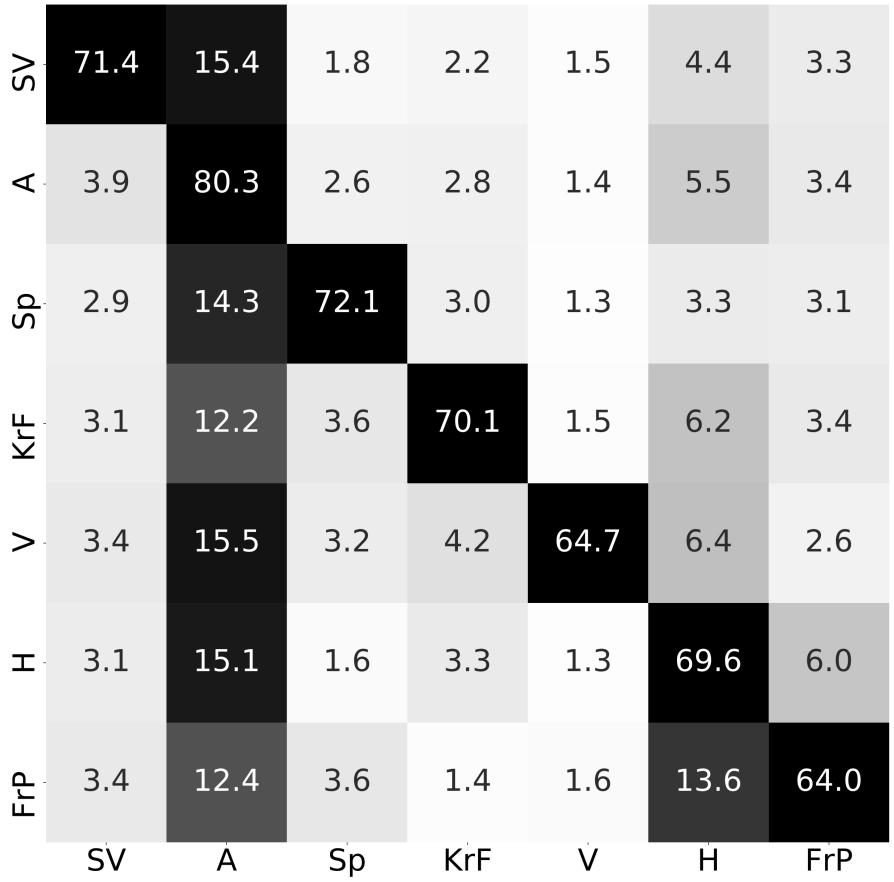
`https://github.com/ltgoslo/talk-of-norway`

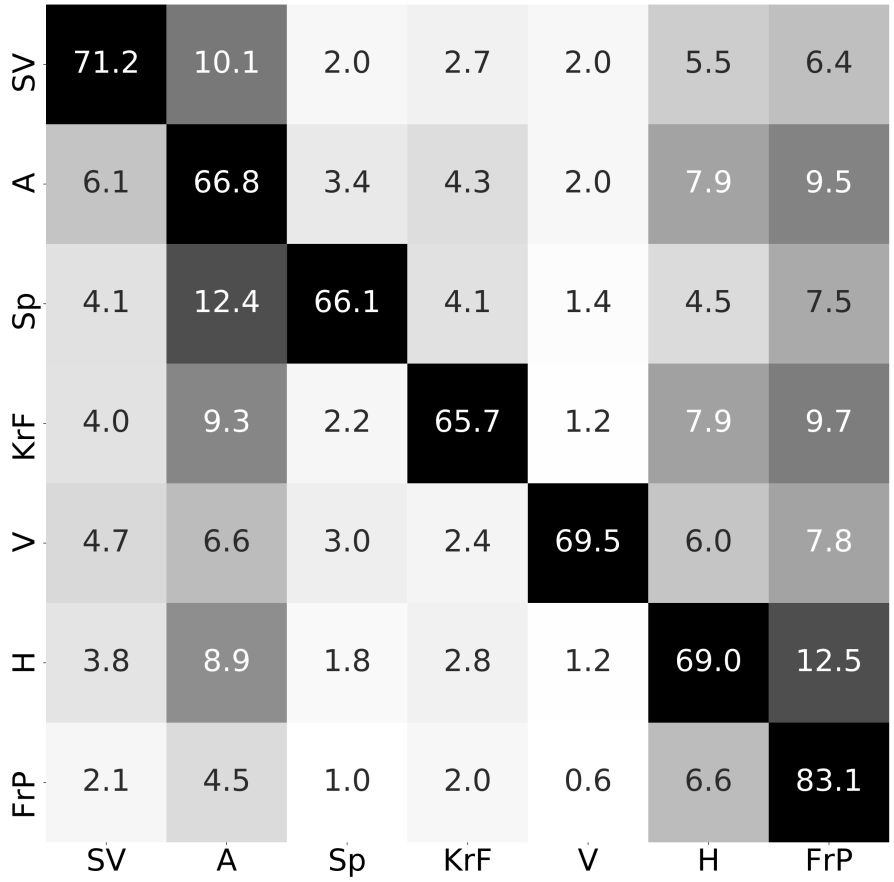
- 18 years of speeches from the Norwegian Parliament, in a novel combination of 80+ meta data variables and language, sentence, token, lemma, PoS and morphology annotations
- An open, plug-and-research resource for text-as-data experimentation

F1 as a quantity of interest, in Norway

- Marked improvements with better normalization, disambiguation and context
- Mixing linguistic and meta-data features helps, too
- Pre-processing does matter for the substance of the analysis

	system	SV	AP	SP	KRF	V	H	FRP	MACRO	ACC
	majority class	—	—	—	—	—	—	—	0.05	0.24
	meta only	0.19	0.38	0.36	0.31	0.24	0.16	0.34	0.28	0.30
unigram	stem	0.57	0.58	0.58	0.51	0.51	0.58	0.62	0.57	—
unigram	token	0.65	0.65	0.66	0.61	0.62	0.65	0.69	0.65	—
unigram	lemma	0.64	0.65	0.65	0.62	0.62	0.64	0.68	0.64	—
unigram	lemma/pos	0.66	0.66	0.67	0.64	0.64	0.66	0.70	0.66	—
unigram	+meta	0.69	0.69	0.72	0.68	0.69	0.68	0.73	0.70	—
n-gram	stem	0.63	0.66	0.65	0.60	0.61	0.65	0.68	0.65	—
n-gram	token	0.67	0.69	0.69	0.66	0.66	0.67	0.71	0.68	—
n-gram	lemma	0.68	0.69	0.69	0.66	0.67	0.68	0.72	0.69	—
n-gram	lemma/pos	0.69	0.70	0.71	0.67	0.69	0.69	0.73	0.70	—
n-gram	+meta	0.71	0.72	0.73	0.71	0.72	0.70	0.75	0.72	—


SV 71.4 15.4 1.8 2.2 1.5 4.4 3.3
A 3.9 80.3 2.6 2.8 1.4 5.5 3.4
Sp 2.9 14.3 72.1 3.0 1.3 3.3 3.1
KrF 3.1 12.2 3.6 70.1 1.5 6.2 3.4
V 3.4 15.5 3.2 4.2 64.7 6.4 2.6
H 3.1 15.1 1.6 3.3 1.3 69.6 6.0
FrP 3.4 12.4 3.6 1.4 1.6 13.6 64.0
SV A Sp KrF V H FrP


SV 71.2 10.1 2.0 2.7 2.0 5.5 6.4
A 6.1 66.8 3.4 4.3 2.0 7.9 9.5
Sp 4.1 12.4 66.1 4.1 1.4 4.5 7.5
KrF 4.0 9.3 2.2 65.7 1.2 7.9 9.7
V 4.7 6.6 3.0 2.4 69.5 6.0 7.8
H 3.8 8.9 1.8 2.8 1.2 69.0 12.5
FrP 2.1 4.5 1.0 2.0 0.6 6.6 83.1
SV A Sp KrF V H FrP

Conclusions ✨

- Heterogenous workflows of off-the-shelf NLP tools can be made easily accessible without sacrificing scalability and flexibility
- Easy access to NLP pre-processing tools benefits political research and analysis

🎨 🏗️ LAP & LXF

- A modular system for HPC-ready, user friendly language annotations
- Scalable to large volumes of data and heterogenous linguistic (and non-linguistic!) annotation

💡 Talk of Norway

- An open resource for parliamentary analysis, informed by and developed with Political Scientists
- The first ToN experiments show substantive differences in the analyses built on party classification performance

2013 Towards large-scale language analysis in the cloud
E Lapponi, E Velldal, NA Vazov, S Oepen

2014 Predicting party affiliations from European Parliament debates
B Høyland, JF Godbout, E Lapponi, E Velldal

2014 Off-Road LAF: Encoding and Processing Annotations in NLP Workflows.
E Lapponi, E Velldal, S Oepen, RL Knudsen

2017 Representation and Interchange of Linguistic Annotation. An In-Depth, Side-by-Side Comparison of Three Designs
RE de Castilho, N Ide, E Lapponi, S Oepen, K Suderman, E Velldal, Marc Verhagen

2018 The Talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016
E Lapponi, MG Søyland, E Velldal, S Oepen

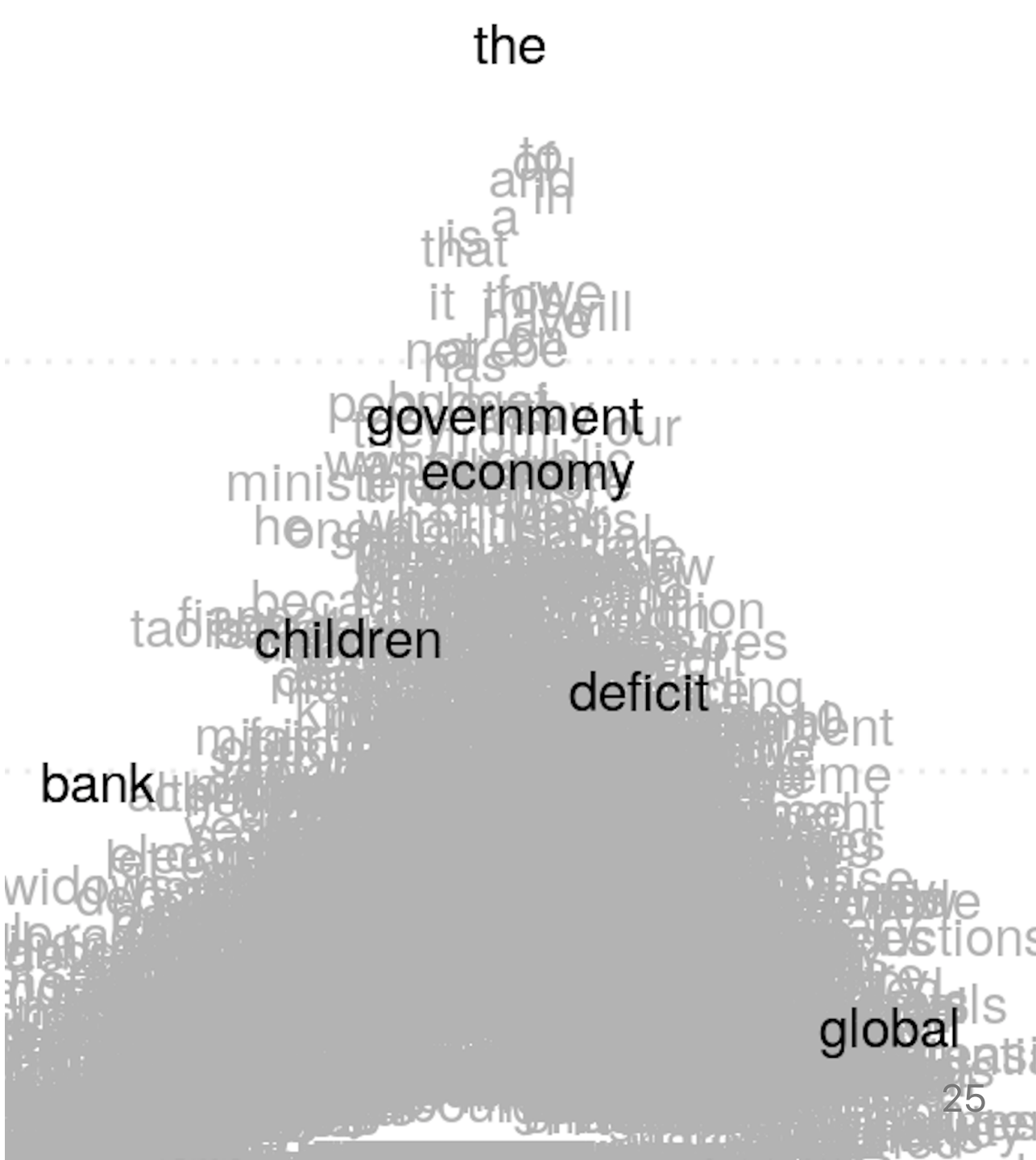
2019 The Language Analysis Portal: Design, implementation and use
E Lapponi

Defense!  

Director's cut 🎬

Using NLP techniques for political analysis

- For example: place words on a left-right political axis*



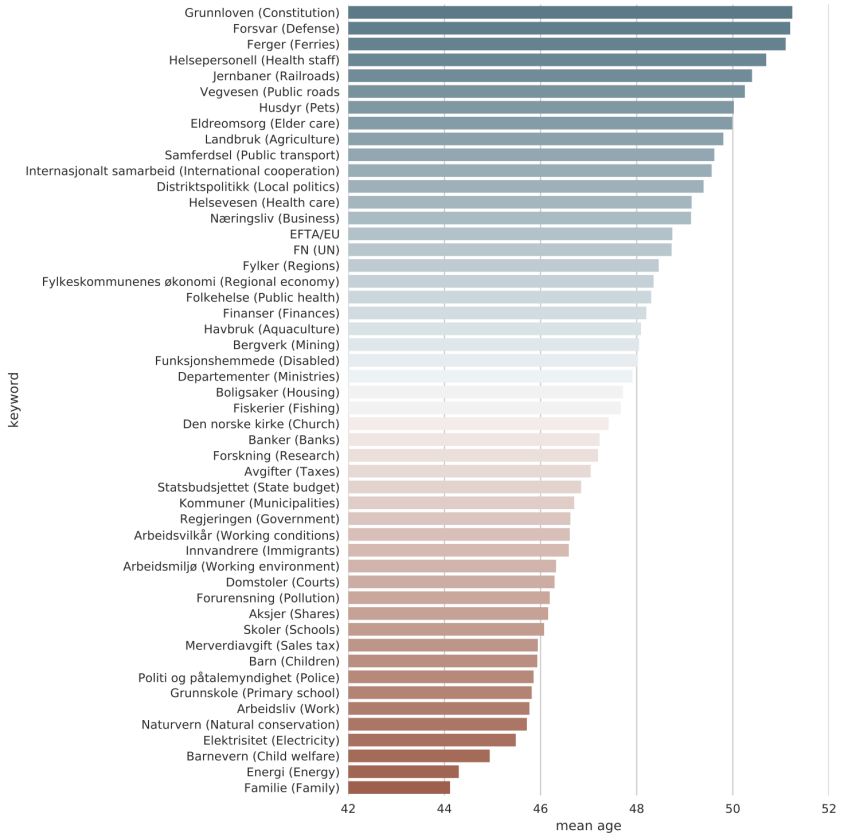
* A Scaling Model for Estimating Time-Series Party Positions from Texts

Getting our feet wet

- Preliminary experiments, going off the recipe
- Signs that linguistic features might contribute positively to these methods

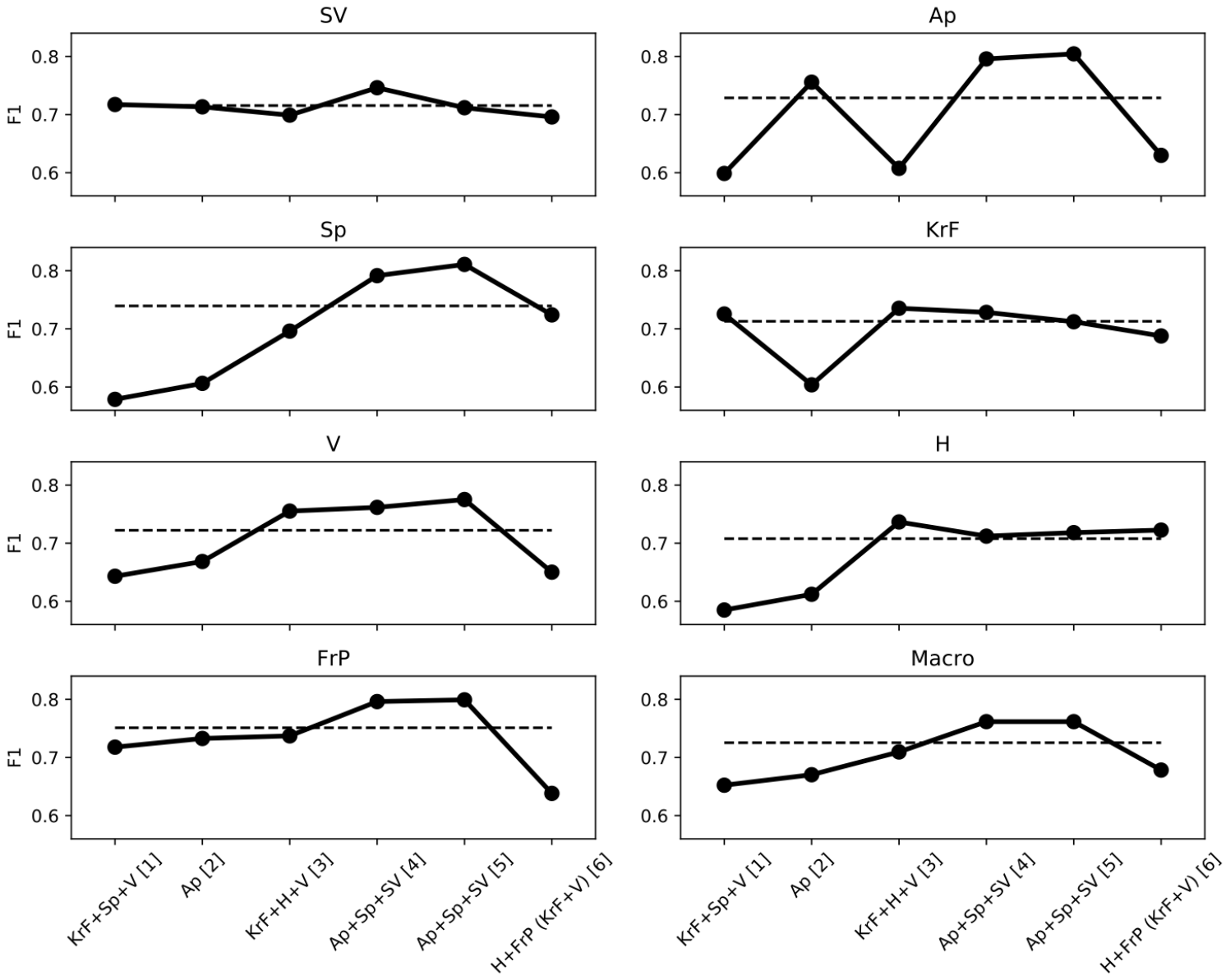
	Baseline	stem	dep/stem
Acc	0.394	0.476	0.492
P	0.065	0.439	0.458
R	0.166	0.399	0.393
F_1	0.094	0.418	0.423

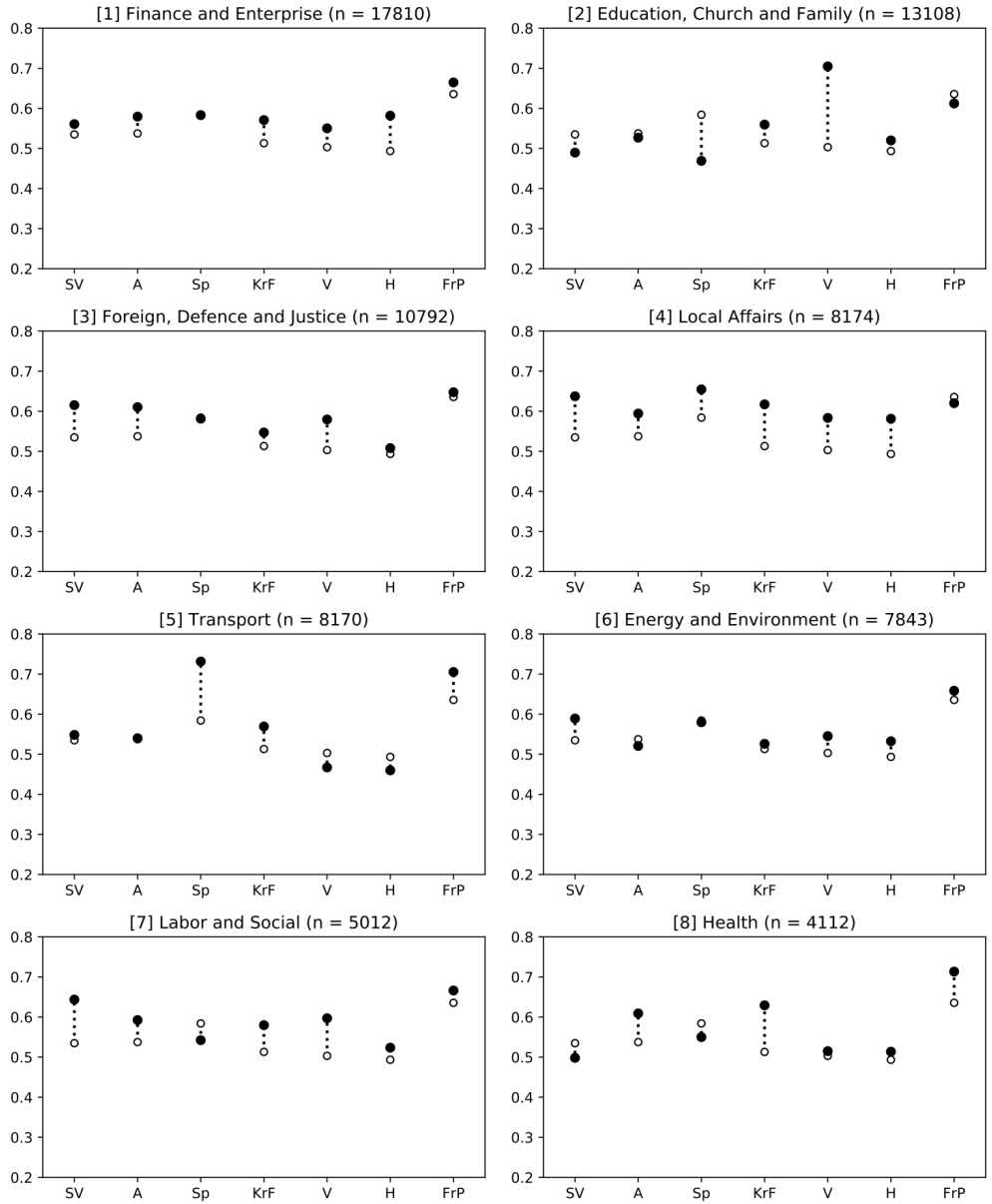
* Predicting Party Affiliations from European Parliament Debates



	SV	FrP
1	ramme videregående skole *affect high schools*	bevilge mer midler *allocate more funds*
2	tøyle grådige markedskrefter *control greedy market forces*	bygge ny vei *build new roads*
3	være liten plass *be little room*	putte mer penger *put more money*
4	vurdere ulike virkemidler *consider different means*	bygge firefelts vei *build four-lane highways*
5	medføre doblet slaktevekt *cause doubling of the carcass weight*	effektivisere norsk økonomi *streamline the norwegian economy*
6	nå ambisiøse mål *reach ambitious goals*	innføre lokal kontantstøtte *introduce local child care allowance*
7	være betydelige kostnader *be substantial costs*	bruke nytt verktøy *use new tools*
8	redusere forurensende utslipp *reduce pollutant emissions*	gi betydelige samfunnsgevinster *give significant social gains*
9	gjøre større utslippskutt *do bigger emission cuts*	utføre kriminelle handlinger *commit criminal acts*
10	få mindre høyrepolitikk *get less right-wing politics*	sikre økt pasientbehandling *ensure increased patient treatment*

On their own, the meta data variables are useful for statistics (what do younger vs. older MEP talk about?), and can be used in combination with the linguistic annotations to quickly go from raw data to policies


SV
Ap
Sp
KrF
V
H
FrP
Macro
F1
0.8
0.7
0.6
KrF+Sp+V [1]
Ap [2]
KrF+H+V [3]
Ap+Sp+SV [4]
Ap+Sp+SV [5]
H+FrP (KrF+V) [6]


[1] Finance and Enterprise (n = 17810)
[2] Education, Church and Family (n = 13108)
[3] Foreign, Defence and Justice (n = 10792)
[4] Local Affairs (n = 8174)
[5] Transport (n = 8170)
[6] Energy and Environment (n = 7843)
[7] Labor and Social (n = 5012)
[8] Health (n = 4112)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
SV
A
Sp
KrF
V
H
FrP

What is an NLP tool?

- Tokenizers, PoS taggers, lemmatizers, syntactic parsers +++
- Come in many flavors, adhere to different traditions, have different pros and cons (depending on the task)
- It's a good idea to test different tools before settling for one for a given task!

Tool interoperability and interchange formats

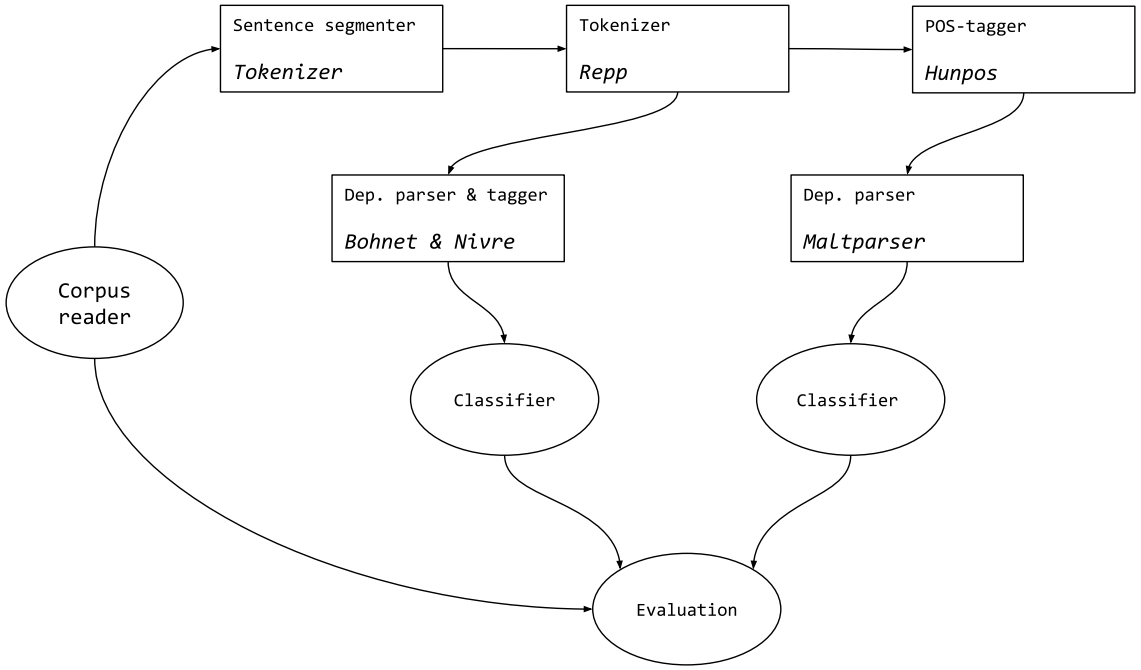
```
Sandy NNP
barks NNS
.     .
```

```
1 Sandy _ _ NNP _ _ _ _ _
2 barks _ _ NNS _ _ _ _ _
3 .     _ _ . _ _ _ _ _
```

```
1 Sandy _ _ NNP _ 2 nn    _ _
2 barks _ _ NNS _ 0 null  _ _
3 .     _ _ .   _ 2 punct _ _
```

What is a workflow?

- Combine tools to get annotations for some downstream task (linguistic or otherwise)
- Tools better be compatible



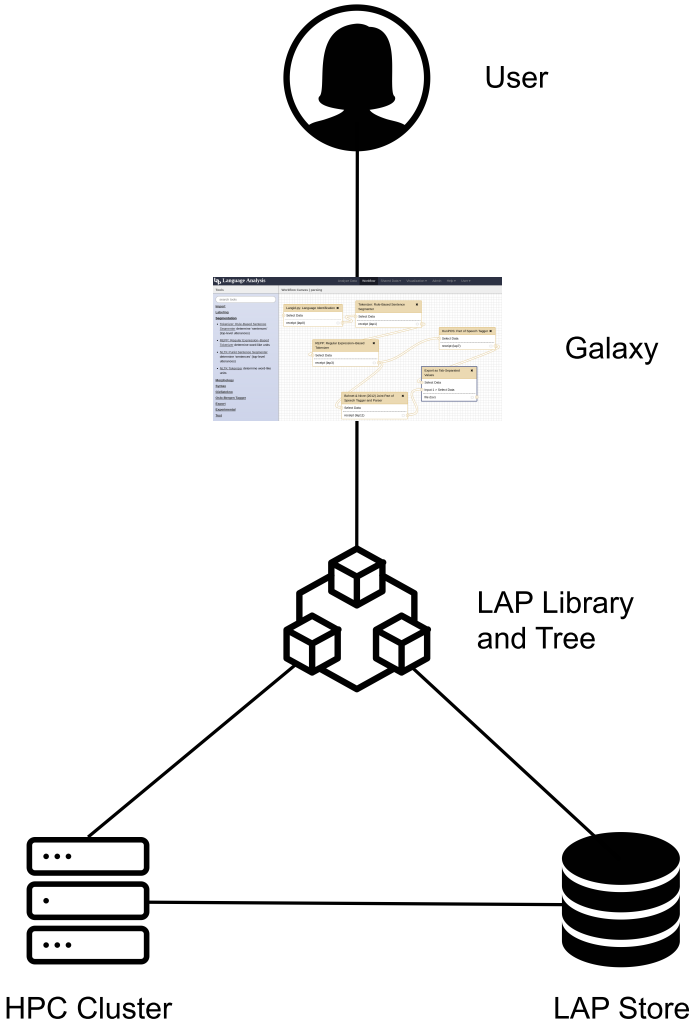
What is a portal?

- No local installation is required
- Data upload and results download
- Graphical user interface to configuring and running tools
- Not developer tools like **NLTK** and **SpaCy**!

What is big data?

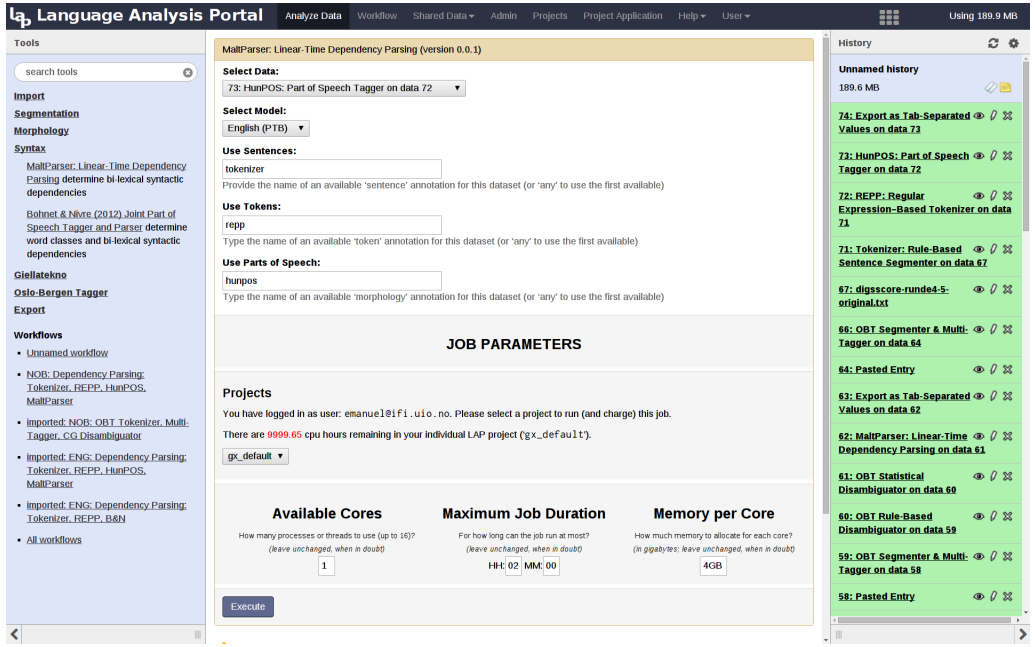
- Big enough to make processing it on a laptop impractical

LAP Architecture

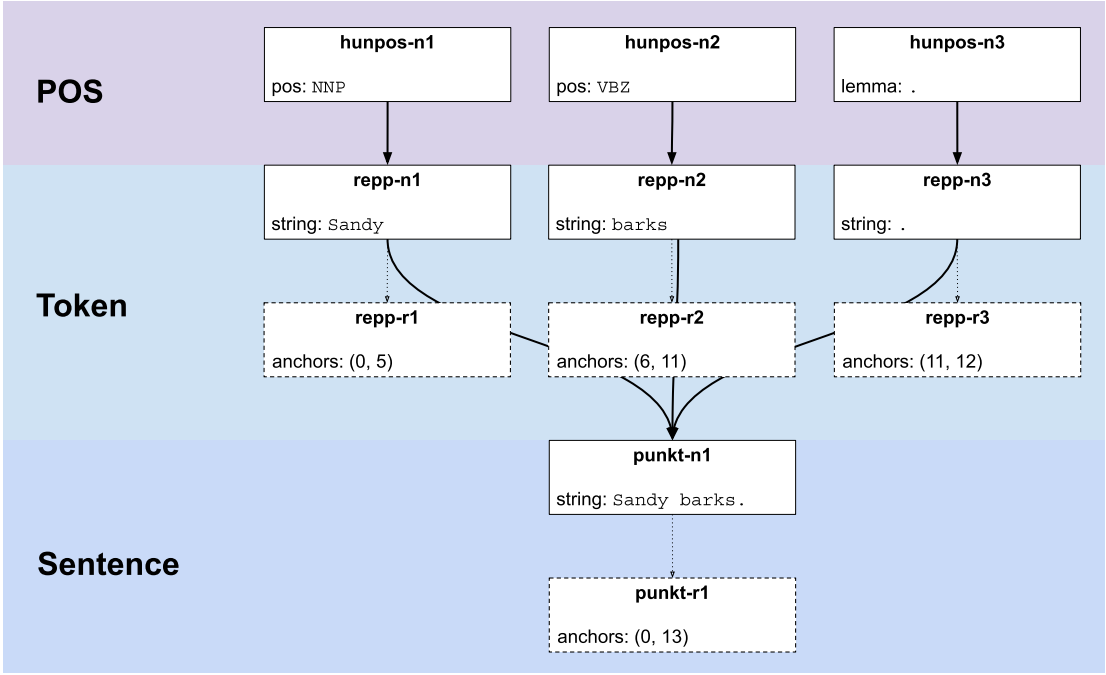


Graphical user interface

- LAP is GUI agnostic
- Blazing the trail for using Galaxy for NLP



```
{
    "annotations": {
        "morphology": [
            "hunpos"
        ],
        "sentence": [
            "nltk_punkt"
        ],
        "token": [
            "repp"
        ]
    },
    "annotators": {
        "hunpos": "26a43d0c-f3db-11e8-b17a-00259075dac6",
        "nltk_punkt": "1d112232-f3db-11e8-a3ee-00259075db92",
        "repp": "20aaddc0-f3db-11e8-b01c-b083fed3d77f"
    },
    "media": {
        "text": "0d3b5012-f3db-11e8-91a5-b083fed3d77f"
    },
    "receipt_origin": "hunpos"
}
```

Our perspective

- Thoroughly navigating the NLP sea before doing political analysis means no political analysis
- Building NLP-for-SSH infrastructures might just be a good idea
- Doing it without navigating the SSH sea likely is not: if we build it, will they come?