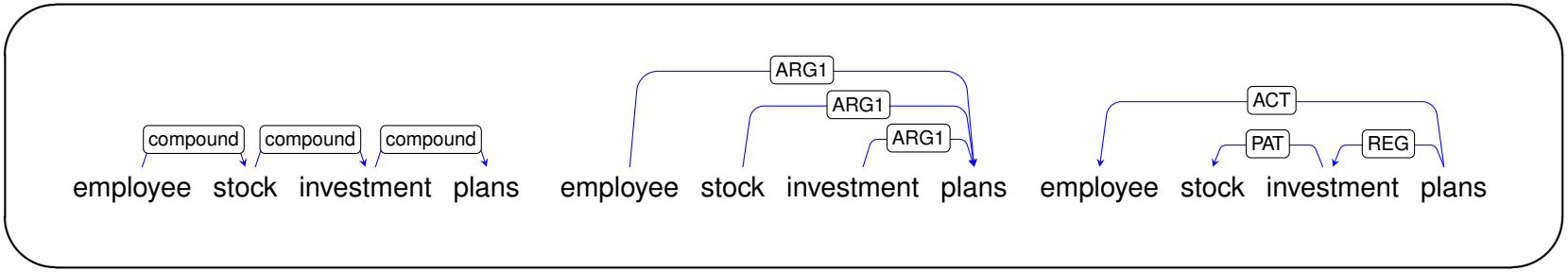


The Language Analysis Portal

Large-Scale, State-of-the-Art Language Analysis for Everyone

Emanuele Lapponi, Milen Kouylekov, Stephan Oepen

Department of Informatics, University of Oslo

lap-developers@ifi.uio.no

A Digital Sea Change: 'Text as Data'

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. [...] This approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. [...] If you tried to read only the entries from the year 2000 alone, [...] it would take eighty years. The sequence of letters is a thousand times longer than the human genome.

[Jean-Baptiste Michel et al., 2010; Science]

A Digital Sea Change: 'Text as Data'

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. [...] This approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. [...] If you tried to read only the entries from the year 2000 alone, [...] it would take eighty years. The sequence of letters is a thousand times longer than the human genome.

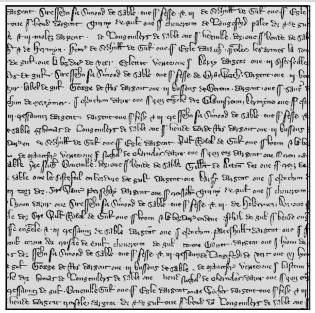
[Jean-Baptiste Michel et al., 2010; Science]

Google Books or Europeana, for Example: Digitizing the World Libraries

Text as Data: A Somewhat Crude Analogy

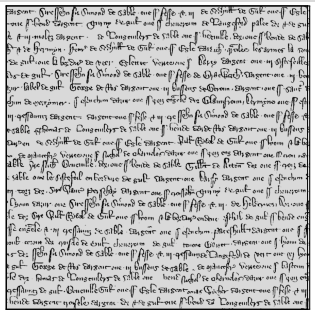
Text as Data: A Somewhat Crude Analogy

Raw Text



Text as Data: A Somewhat Crude Analogy

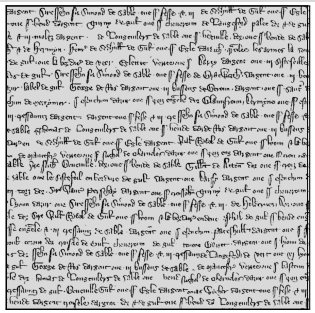
Raw Text



Science

Text as Data: A Somewhat Crude Analogy

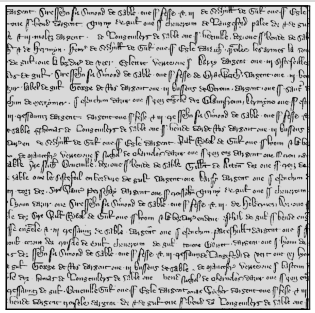
Raw Text



Utility

Text as Data: A Somewhat Crude Analogy

Raw Text

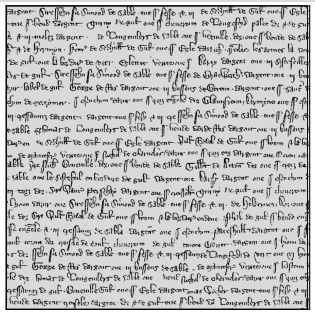


Refinement

Science

Text as Data: A Somewhat Crude Analogy

Raw Text



Refinement

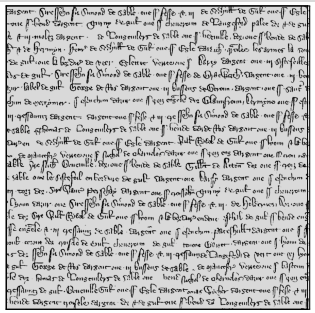
Science

Text as Data: A Somewhat Crude Analogy

Raw Text

Refinement

Science

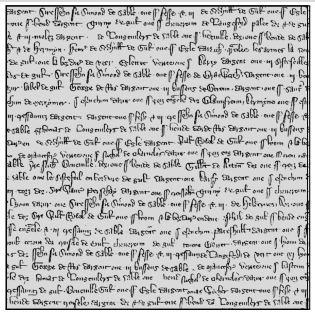


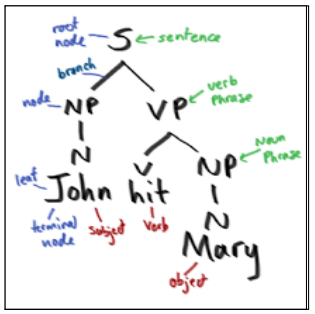
Text as Data: A Somewhat Crude Analogy

Raw Text

Refinement

Science





By Way of Introduction: 'Basic' Linguistic Analysis

Textual Building Blocks (All Too Often Taken for Granted)

- *Sentence Splitting* headlines, bullet lists, interactions with mark-up, etc.
- *Tokenization* punctuation marks, abbreviations, dashes and slashes, etc.

By Way of Introduction: 'Basic' Linguistic Analysis

Textual Building Blocks (All Too Often Taken for Granted)

- *Sentence Splitting* headlines, bullet lists, interactions with mark-up, etc.
- *Tokenization* punctuation marks, abbreviations, dashes and slashes, etc.

The Olympic Committee already regrets choosing Russia for this Winter Olympics .

By Way of Introduction: 'Basic' Linguistic Analysis

Textual Building Blocks (All Too Often Taken for Granted)

- *Sentence Splitting* headlines, bullet lists, interactions with mark-up, etc.
- *Tokenization* punctuation marks, abbreviations, dashes and slashes, etc.

The	Olympic	Committee	already	regrets	choosing	Russia	for	this	Winter	Olympics	.
the	Olympic	Committee	already	regret	choose	Russia	for	this	Winter	Olympics	.

By Way of Introduction: 'Basic' Linguistic Analysis

Textual Building Blocks (All Too Often Taken for Granted)

- *Sentence Splitting* headlines, bullet lists, interactions with mark-up, etc.
- *Tokenization* punctuation marks, abbreviations, dashes and slashes, etc.

The	Olympic	Committee	already	regrets	choosing	Russia	for	this	Winter	Olympics	.
the	Olympic	Committee	already	regret	choose	Russia	for	this	Winter	Olympics	.
DT	NNP	NNP	RB	VBZ	VBG	NNP	IN	DT	NN	NN	.

By Way of Introduction: 'Basic' Linguistic Analysis

Textual Building Blocks (All Too Often Taken for Granted)

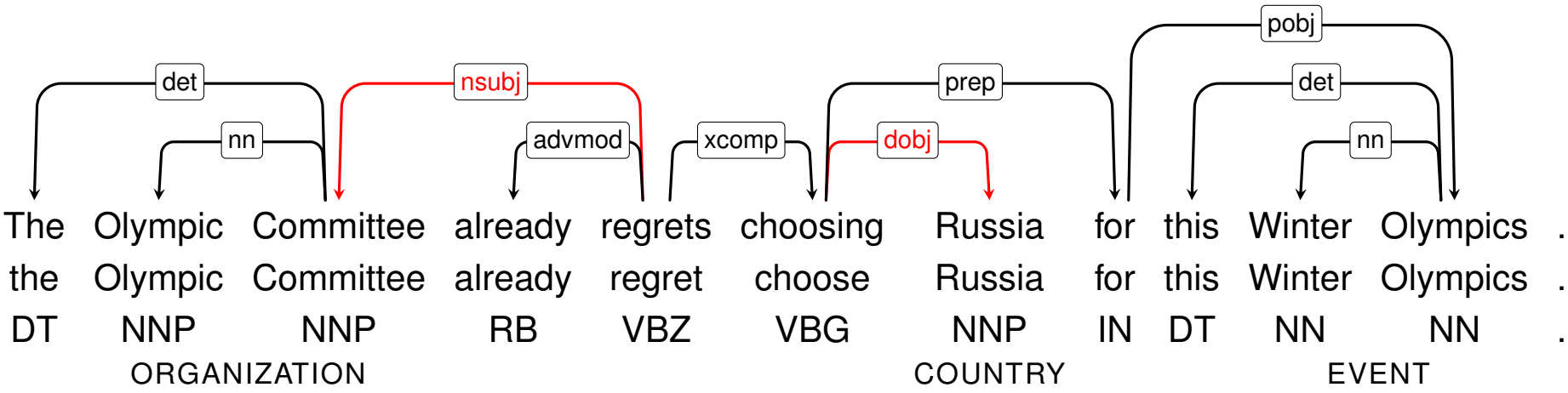
- *Sentence Splitting* headlines, bullet lists, interactions with mark-up, etc.
- *Tokenization* punctuation marks, abbreviations, dashes and slashes, etc.

The	Olympic	Committee	already	regrets	choosing	Russia	for	this	Winter	Olympics	.
the	Olympic	Committee	already	regret	choose	Russia	for	this	Winter	Olympics	.
DT	NNP	NNP	RB	VBZ	VBG	NNP	IN	DT	NN	NN	.
	ORGANIZATION					COUNTRY			EVENT		

By Way of Introduction: ‘Basic’ Linguistic Analysis

Textual Building Blocks (All Too Often Taken for Granted)

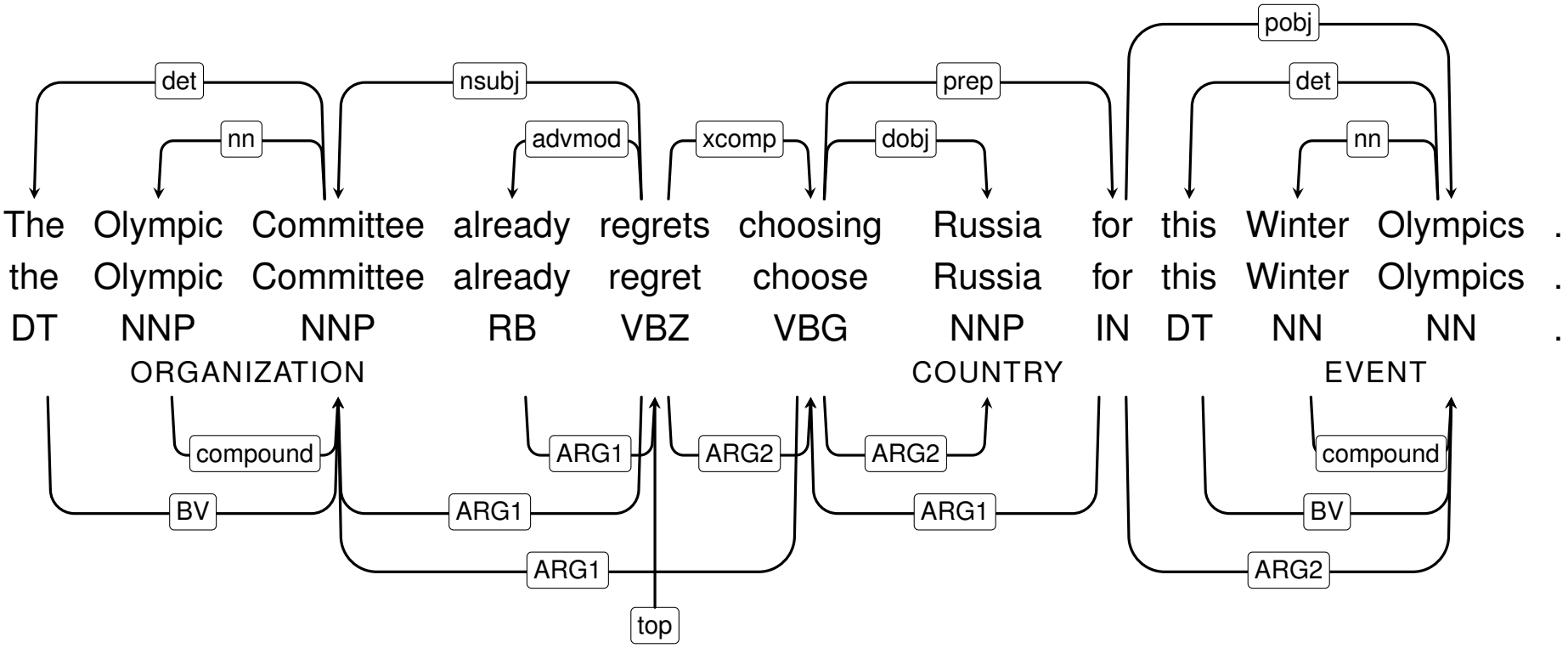
- *Sentence Splitting* headlines, bullet lists, interactions with mark-up, etc.
- *Tokenization* punctuation marks, abbreviations, dashes and slashes, etc.



By Way of Introduction: ‘Basic’ Linguistic Analysis

Textual Building Blocks (All Too Often Taken for Granted)

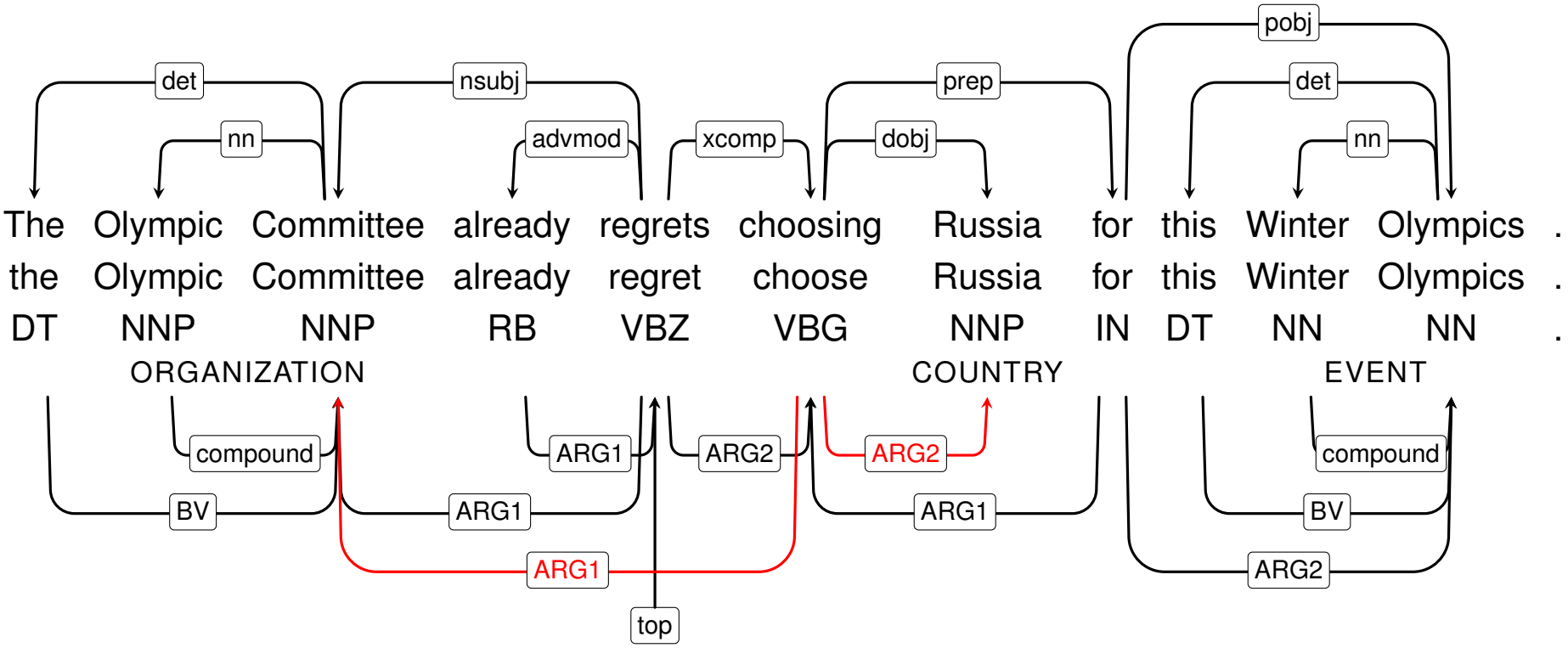
- *Sentence Splitting* headlines, bullet lists, interactions with mark-up, etc.
- *Tokenization* punctuation marks, abbreviations, dashes and slashes, etc.



By Way of Introduction: ‘Basic’ Linguistic Analysis

Textual Building Blocks (All Too Often Taken for Granted)

- *Sentence Splitting* headlines, bullet lists, interactions with mark-up, etc.
- *Tokenization* punctuation marks, abbreviations, dashes and slashes, etc.



In Support of Digital Humanities and Social Sciences

Predicting Party Affiliations from European Parliament Debates

Bjørn Høyland
Department of Political Science
University of Oslo
bjorn.hoyland@stv.uio.no

Jean-François Godbout
Department of Political Science
University of Montreal
godboutj@umontreal.ca

Emanuele Lapponi
Department of Informatics
University of Oslo
emanuel@ifi.uio.no

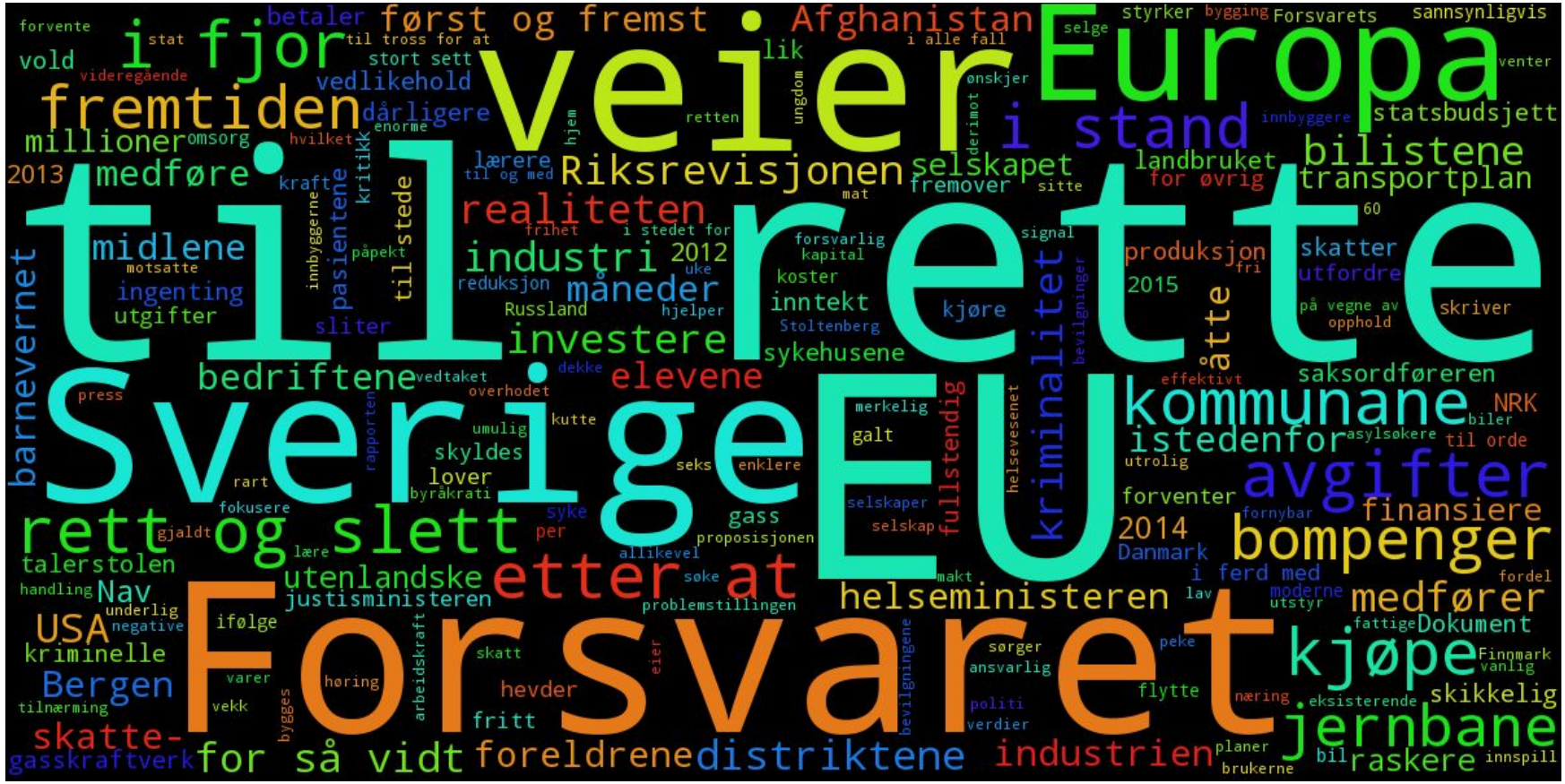
Erik Velldal
Department of Informatics
University of Oslo
erikve@ifi.uio.no

Abstract

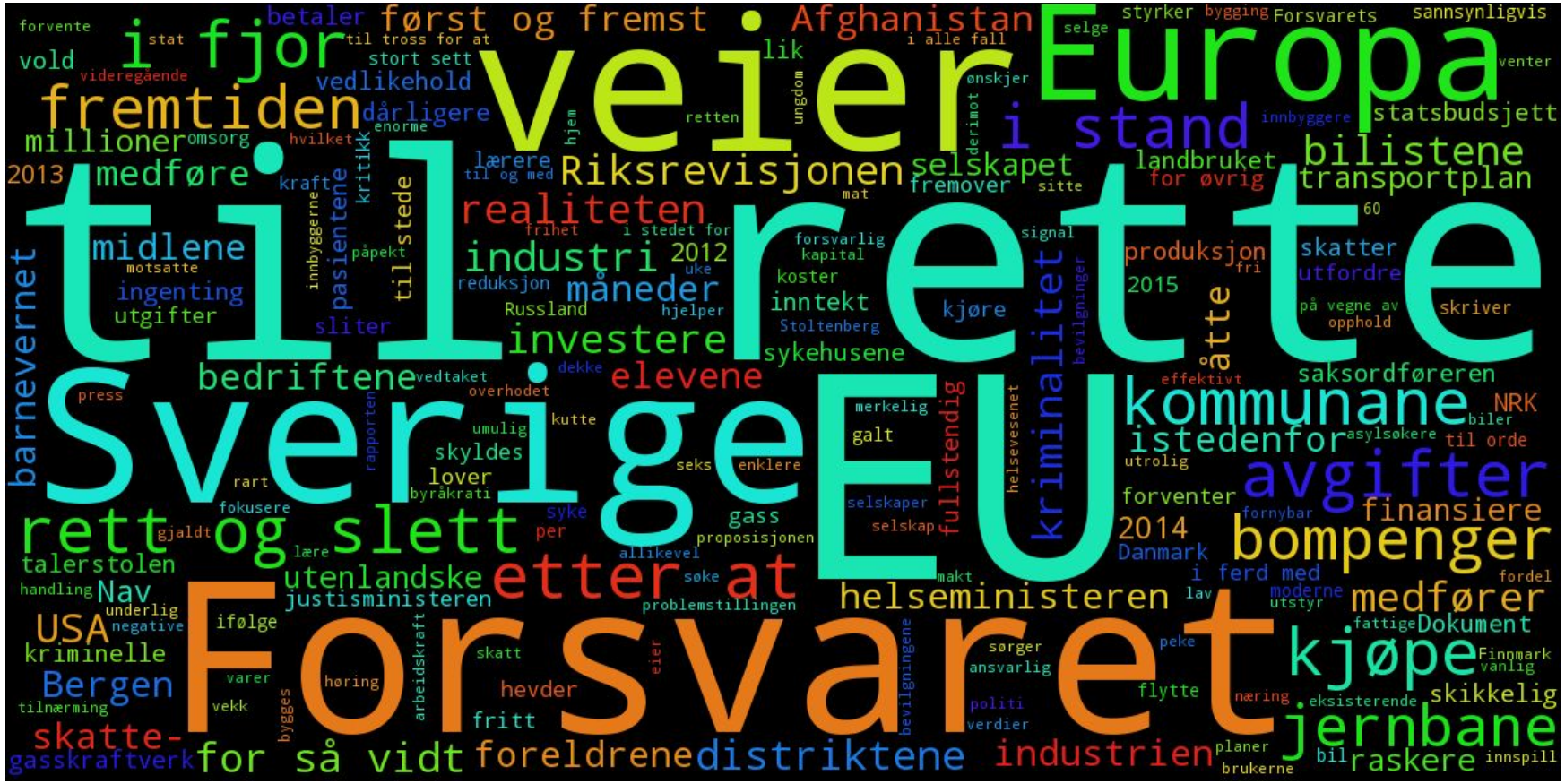
This paper documents an ongoing effort to assess whether party group affiliation of participants in European Parliament debates can be automatically predicted on the basis of the content of their speeches, premise for the success of such an approach is that differences in ideology and belief systems are reflected in differences in choice of words in plenary debates. Another premise is that a shared belief system translates to the same choice of party group. As discussed below, systematic differences in prediction performance in the data can be used

(2014 Workshop on Language Technologies and Computational Social Science)

A Trivial Example: Benefits of Linguistic Analysis



A Trivial Example: Benefits of Linguistic Analysis



- 250,000 speech transcripts from the Norwegian Parliament: 1998–2016;
- rich and structured metadata available; e.g. all speeches from one party.

Seeking to ‘Amplify’ the Linguistic Signal

Seeking to ‘Amplify’ the Linguistic Signal

- Morphological analysis through the Oslo–Bergen Tagger;
- ignore everything but nouns; use lemmata (‘base forms’).

LAP: High-Level Goals and Design

Motivation

- Lower barrier to entry for non-technical users: state-of-the-art tools;
- unified, Web-based point of entry; balance ease of use and flexibility.

Core Elements

- **Tools** Pre-installed and -configured, multi-lingual, common tasks;
- **Interfaces** Portal access; data injection and export; parameterization;
- **Plug & Play** 'mix and match' tools: unified LAP interchage format.

Scalability

- Submit to Norwegian eInfrastructure: 11,000$^+$ cores at your fingertips.

LAP: High-Level Goals and Design

Motivation

Incremental, demand-driven approch to tool integration;

$\rightarrow$ so far, languages and tools represented in Norway.

Core Elements

- **Tools** Pre-installed and -configured, multi-lingual, common tasks;
- **Interfaces** Portal access; data injection and export; parameterization;
- **Plug & Play** 'mix and match' tools: unified LAP interchage format.

Scalability

- Submit to Norwegian eInfrastructure: 11,000$^+$ cores at your fingertips.

LAP: High-Level Goals and Design

Motivation

Incremental, demand-driven approch to tool integration;
$\rightarrow$ so far, languages and tools represented in Norway.

Core Elements

In parallel, reach out to candidate consumers (SSH researchers);
jointly implement 'use case' studies, e.g. Høyland, et al. (2014).

- **Plug & Play** 'mix and match' tools: unified LAP interchage format.

Scalability

- Submit to Norwegian eInfrastructure: 11,000$^+$ cores at your fingertips.

LAP: High-Level Goals and Design

Motivation

Incremental, demand-driven approch to tool integration;
$\rightarrow$ so far, languages and tools represented in Norway.

Core Elements

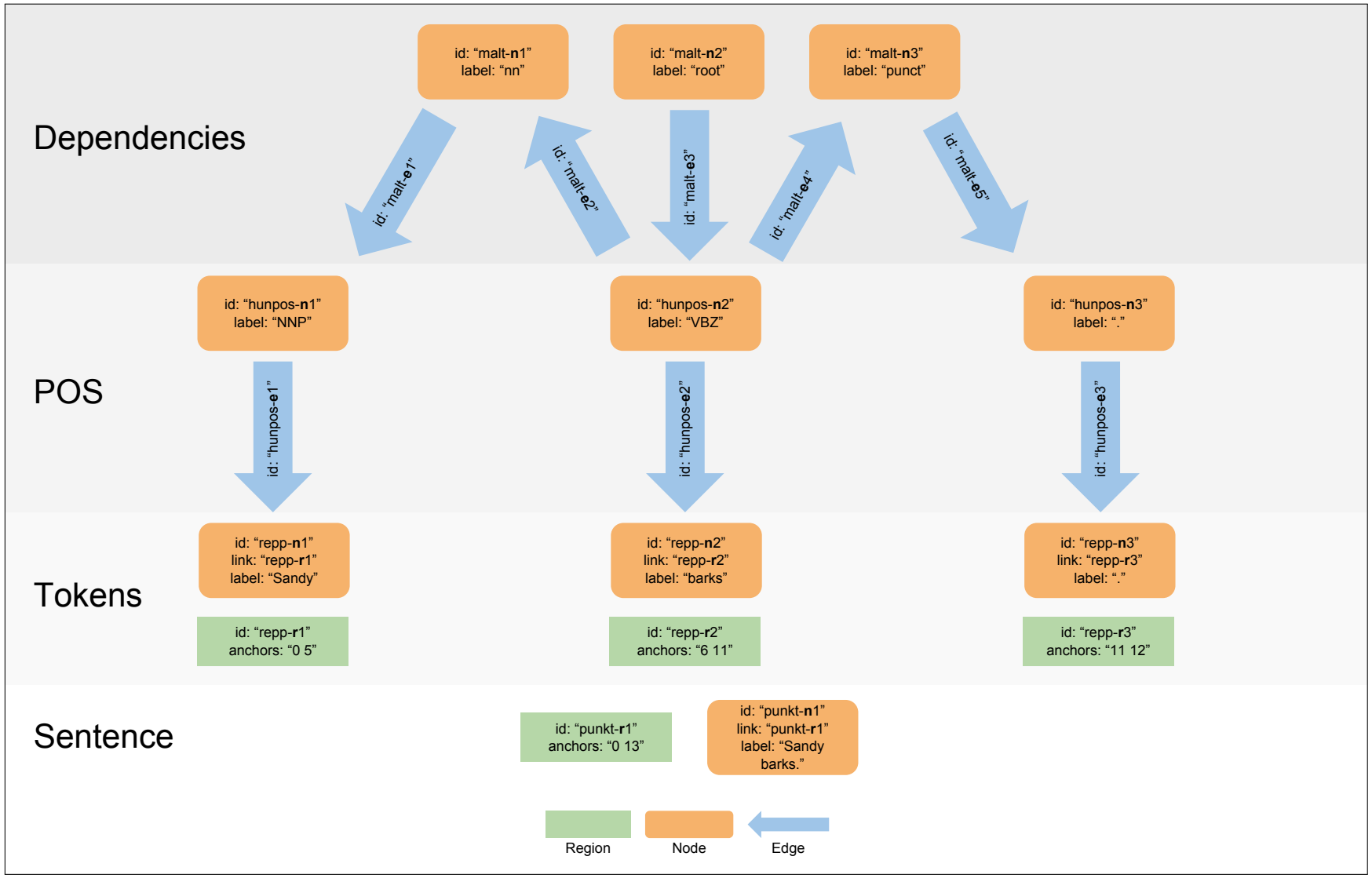
In parallel, reach out to candidate consumers (SSH researchers);
jointly implement 'use case' studies, e.g. Høyland, et al. (2014).

- **Plug & Play** 'mix and match' tools: unified LAP interchage format.

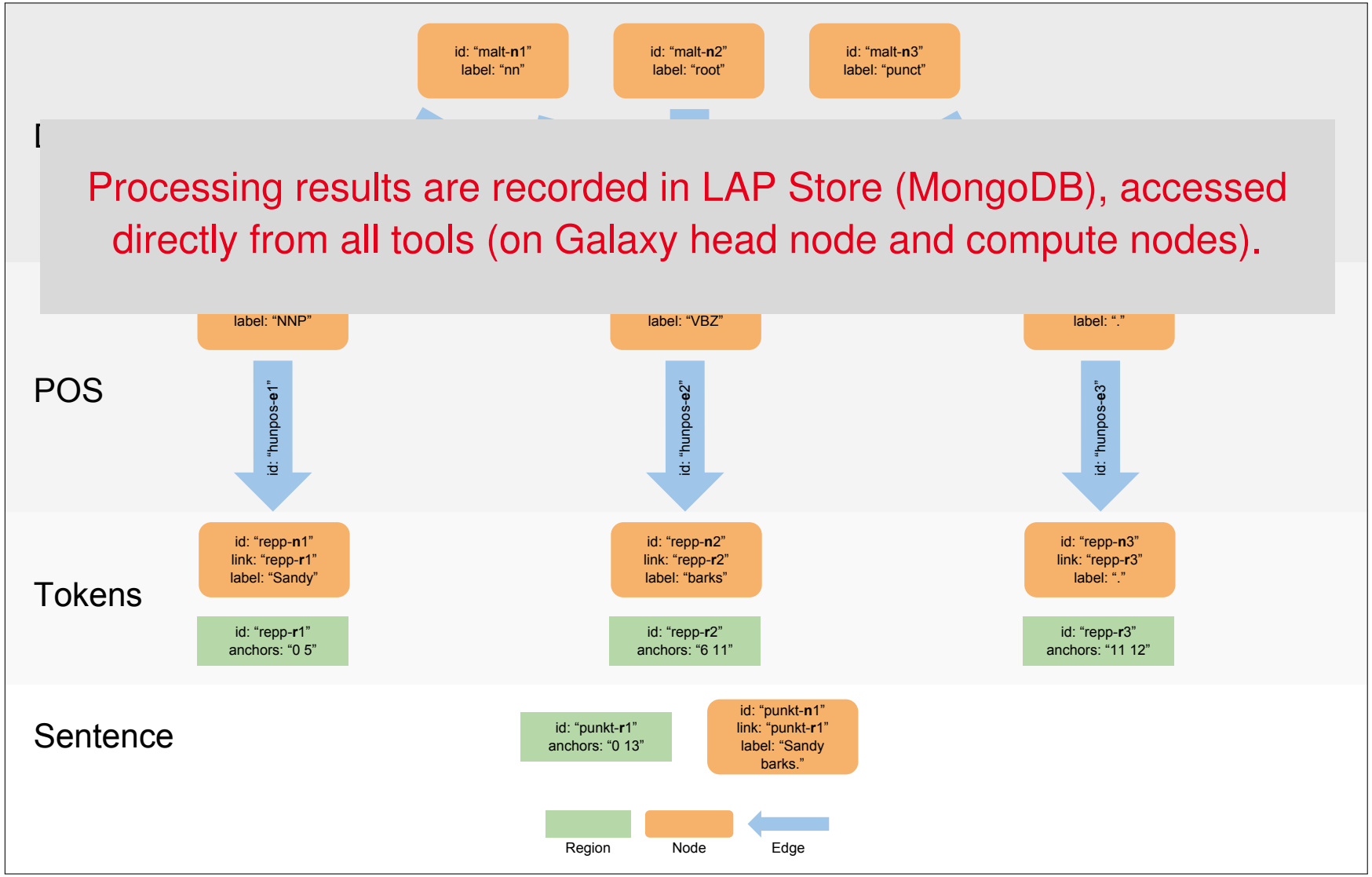
CLARINO: Funded by Norwegian Research Council (2013–2018).

- Submit to Norwegian eInfrastructure: 11,000$^+$ cores at your fingertips.

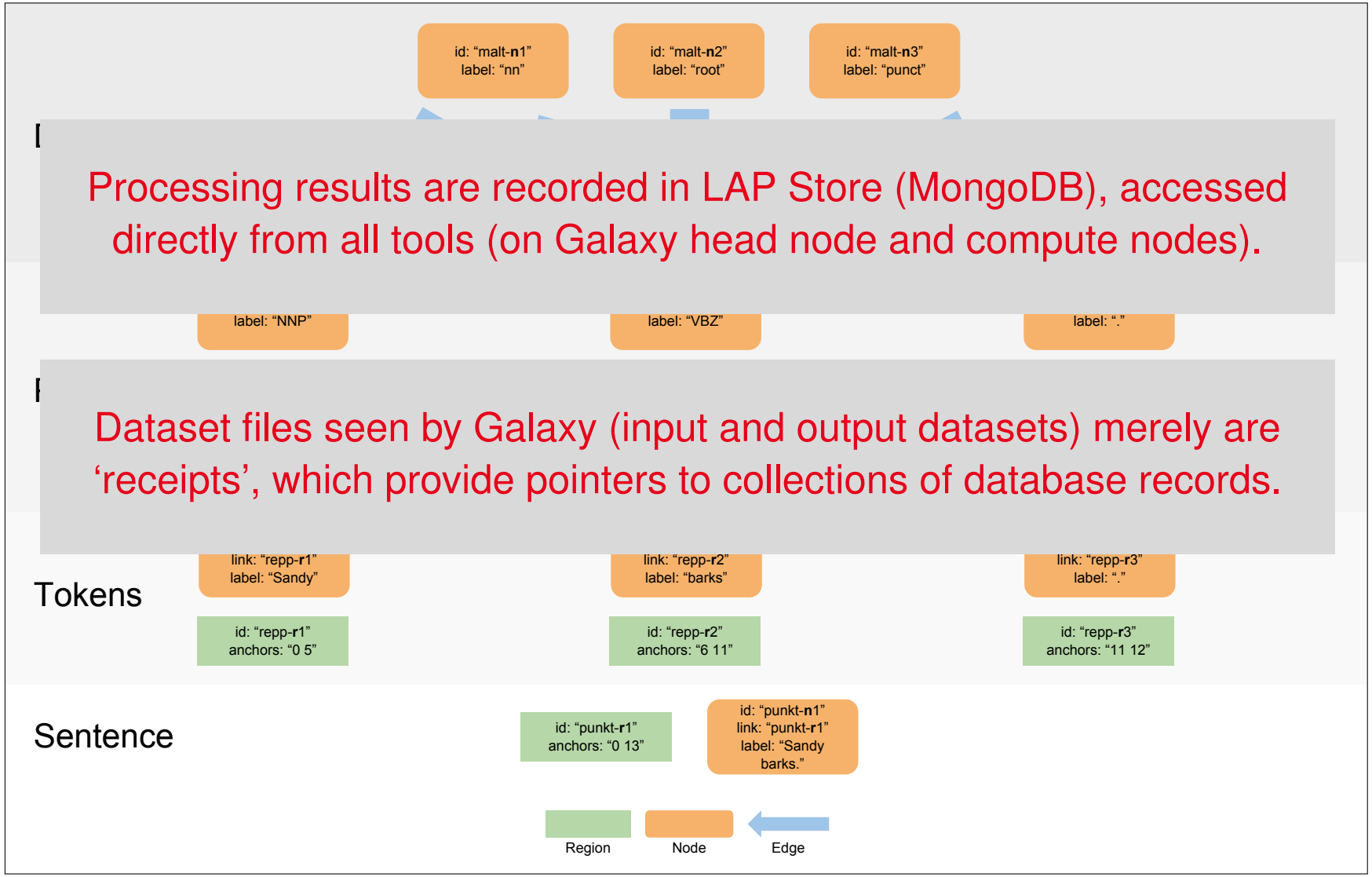
LAF: The Linguistic Annotation Framework



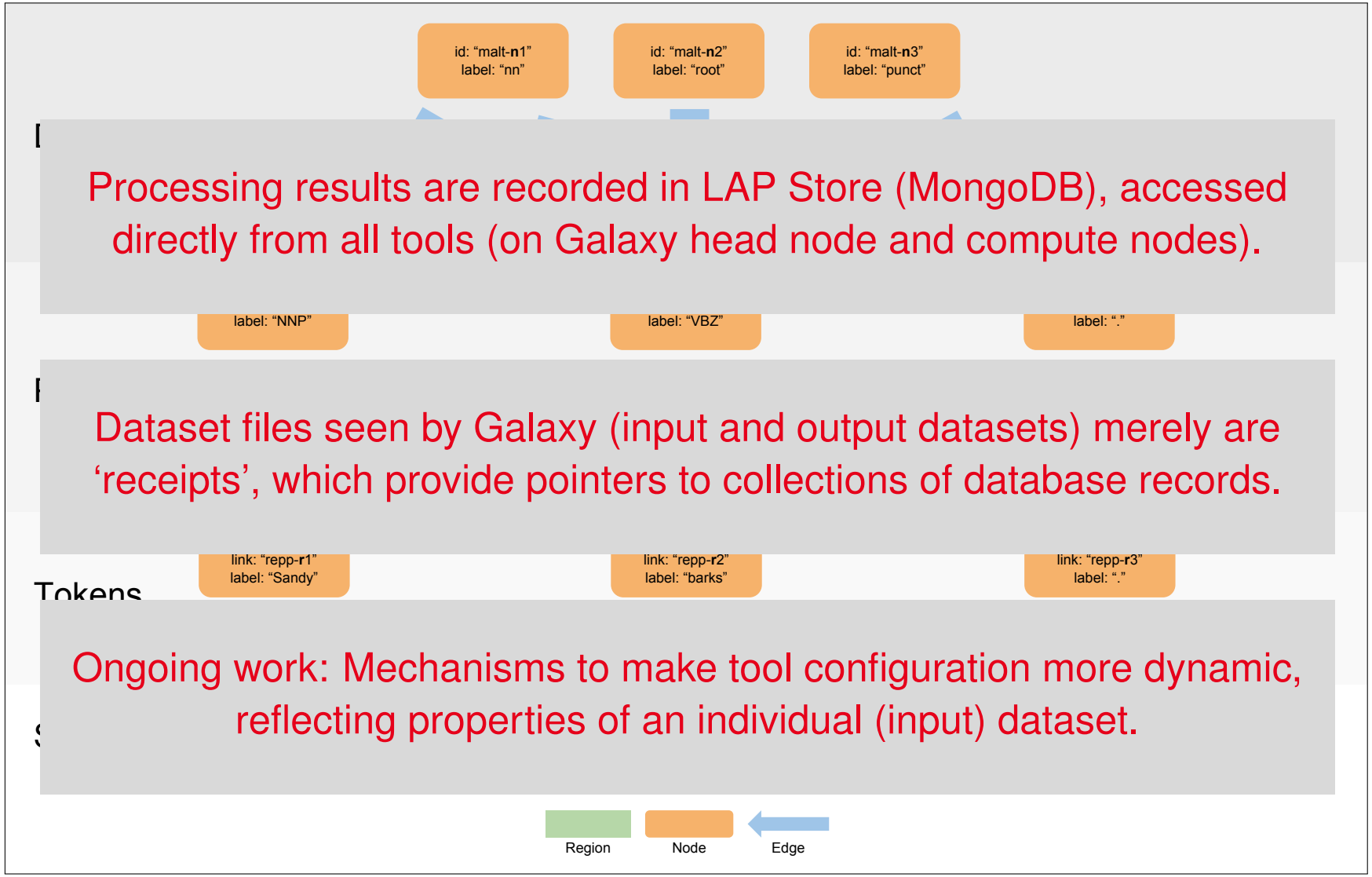
LAF: The Linguistic Annotation Framework



LAF: The Linguistic Annotation Framework



LAF: The Linguistic Annotation Framework



Go, Experiment Yourselves (Now in Production)!

http://lap.clarino.uio.no/

Go, Experiment Yourselves (Now in Production)!

http://lap.clarino.uio.no/

If we build it ... Will you come?

Go, Experiment Yourselves (Now in Production)!

http://lap.clarino.uio.no/

If we build it ... Will you come?

Go, Experiment Yourselves (Now in Production)!

http://lap.clarino.uio.no/

If we build it ... Will you come?