# IXA pipes: Efficient and Ready to Use Multilingual NLP tools

Rodrigo Agerri

IXA NLP Group, UPV/EHU
OpenNLP project, Apache Software Foundation

# Outline

# Overview

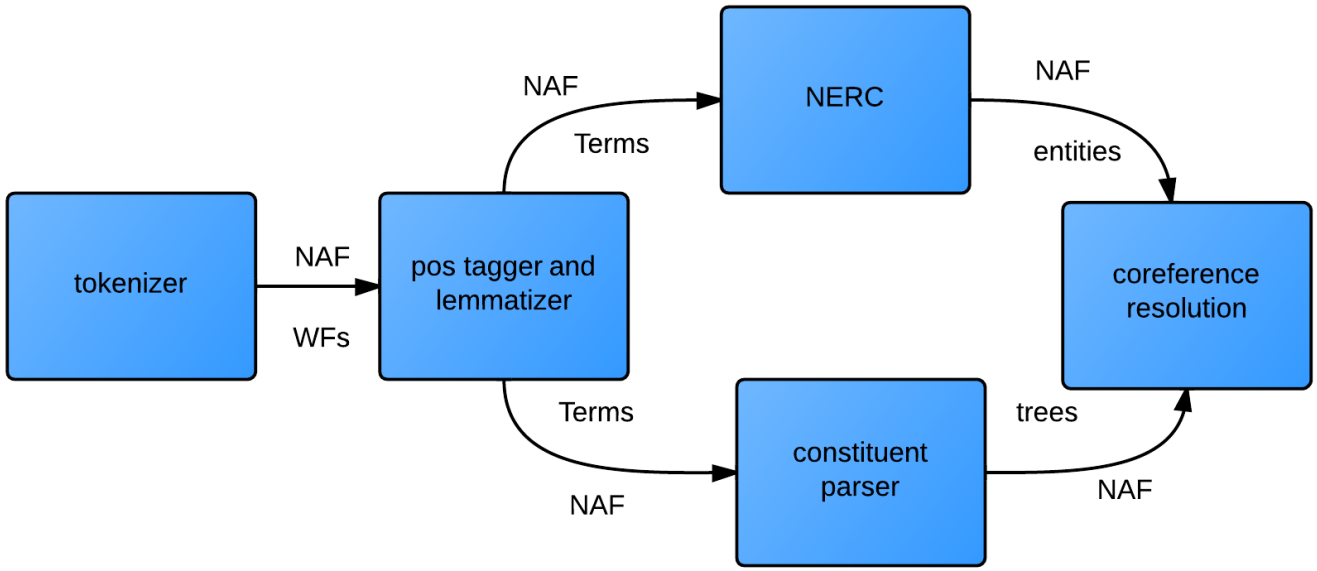
http://ixa2.si.ehu.es/ixa-pipes

# Motivation

Lowering the barriers of using NLP technology and allow researchers and SMEs to focus on their central/primary interests:

- Simple: Two simple steps: One if you get the binaries!
- Portable: Only a JVM 1.7+ and Maven is required.
- Modular data-centric architecture: The tools behave like Unix pipes; easily replaceable and extensible architecture.
- Multilingual: 8 languages and more coming soon!!
- Accurate: State of the art results.
- APL 2.0: To facilitate integration also with commercial applications.

# Architecture (or lack thereof)



# Basics

- NAF: Natural Language annotation Format
  https://github.com/newsreader/NAF.
- kaflib: https://github.com/ixa-ehu/kaflib
- Apache Maven: http://maven.apache.org.
- github and git: https://github.com/ixa-ehu/
- Apache OpenNLP Machine Learning Library:
  http://opennlp.apache.org.

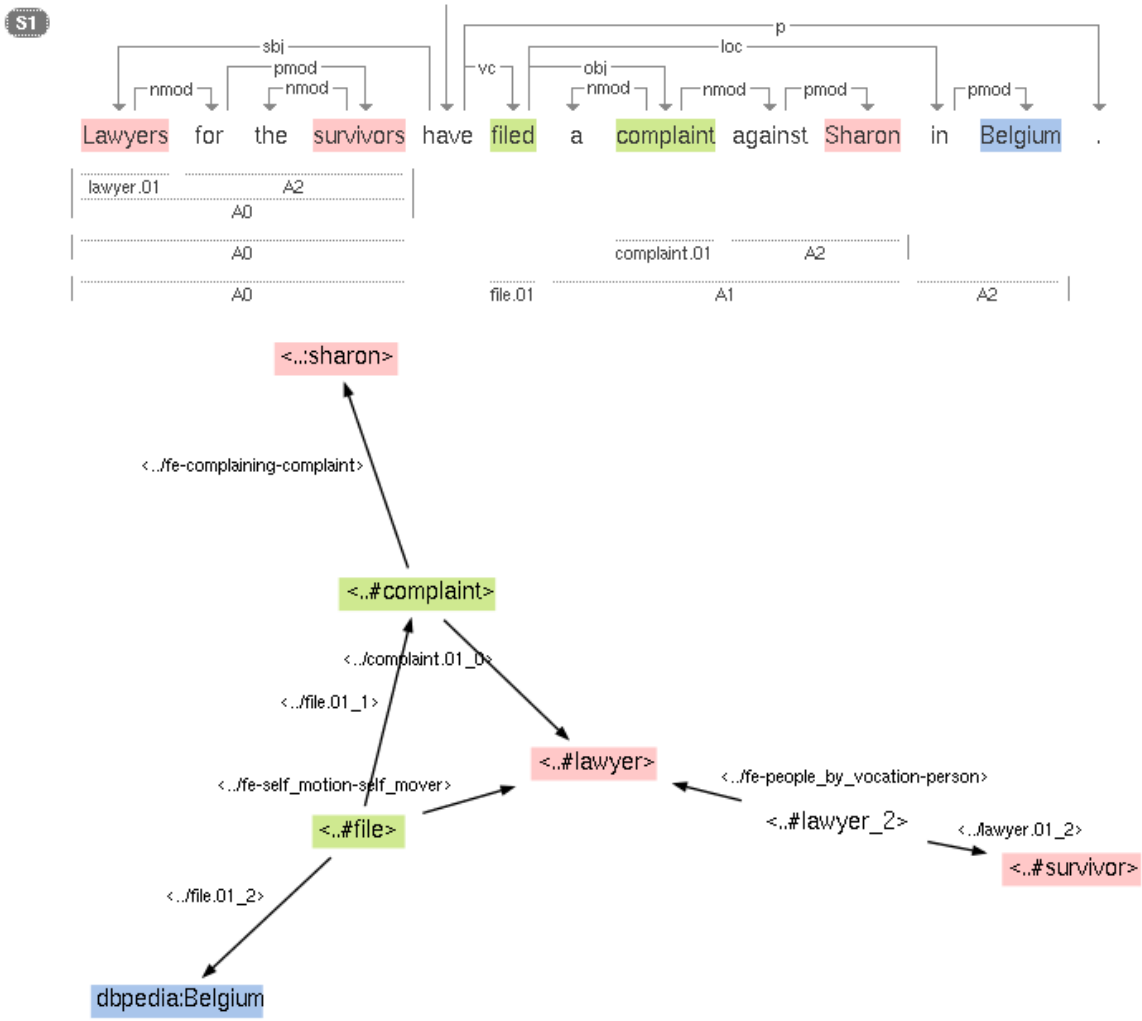
# NLP Annotation example


http://www.opener-project.eu/project/demos/

# NLP Annotation example


http://ber2tekdemo-opener.rhcloud.com/welcome.action

# NLP annotation example




http://pikes.fbk.eu/

# Pipes

| task | languages | ixa-pipe |
|---|---|---|
| tok | en, es, eu, fr, gl, it, nl | ixa-pipe-tok |
| pos | en, es, eu, fr, gl, it | ixa-pipe-pos |
| lemmatizer | en, es, eu, fr, gl, it | ixa-pipe-pos |
| nerc | de, en, es, eu, gl, it, nl | ixa-pipe-nerc |
| ote | en, es, fr, ru, tr, nl | ixa-pipe-nerc |
| sst | en | ixa-pipe-nerc |
| parse | en, es | ixa-pipe-parse |

# ixa-pipe-tok

```
<wf id="w69" sent="4" para="4" offset="354"
length="9">announced</wf>
```

- Tested for many languages, apostrophe treatment, etc.
- Treebank normalization: Ancora and Penn Treebank, Universal dependencies normalized tokenization...
- Paragraph treatment, whitespace tokenizer...
- Rule-based: regular expressions.

# ixa-pipe-pos

rosa rosa AQ0CS0
rosa rosa NCFS000
rosa rosa NCMS000

```
<term id="t69" type="open" lemma="rosa" pos="R"
                morphofeat="NCFS000">
```

| POS tagger | Basque | English | Spanish | Italian |
|---|---|---|---|---|
| ixa-pipe-pos | 94.28 | 97.07 | 98.88 | 95.00 |
| SVMTool | | 97.16 | 98.86* | |
| Stanford POS | | 97.24 | | |
| Freeling | | 97** | 97** | |
| Felice 2009 | | | | 96.34 |

# ixa-pipe-nerc

Morras munduko txapeldun izan zen juniorretan 1994an, Ekuadorko hiriburuan, Quiton.

| NERC | eu | en | es | nl | de |
|---|---|---|---|---|---|
| ixa-pipe-nerc | 75.70 | 91.36 | 84.16 | 85.04 | 76.48 |
| Passos et al. 2014 | – | 90.90 | – | – | – |
| Ratinov and Roth 2009 | – | 90.57 | – | – | – |
| Stanford NER | – | 88.65 | – | – | – |
| CMP (2002-03) | – | 85.00 | 81.39 | 77.05 | – |
| C&C | – | – | – | 79.63 | – |
| Eihera | 71.31 | – | – | – | – |
| ExB (2014) | – | – | – | – | 76.38 |

# Out-of-domain: Wikinews

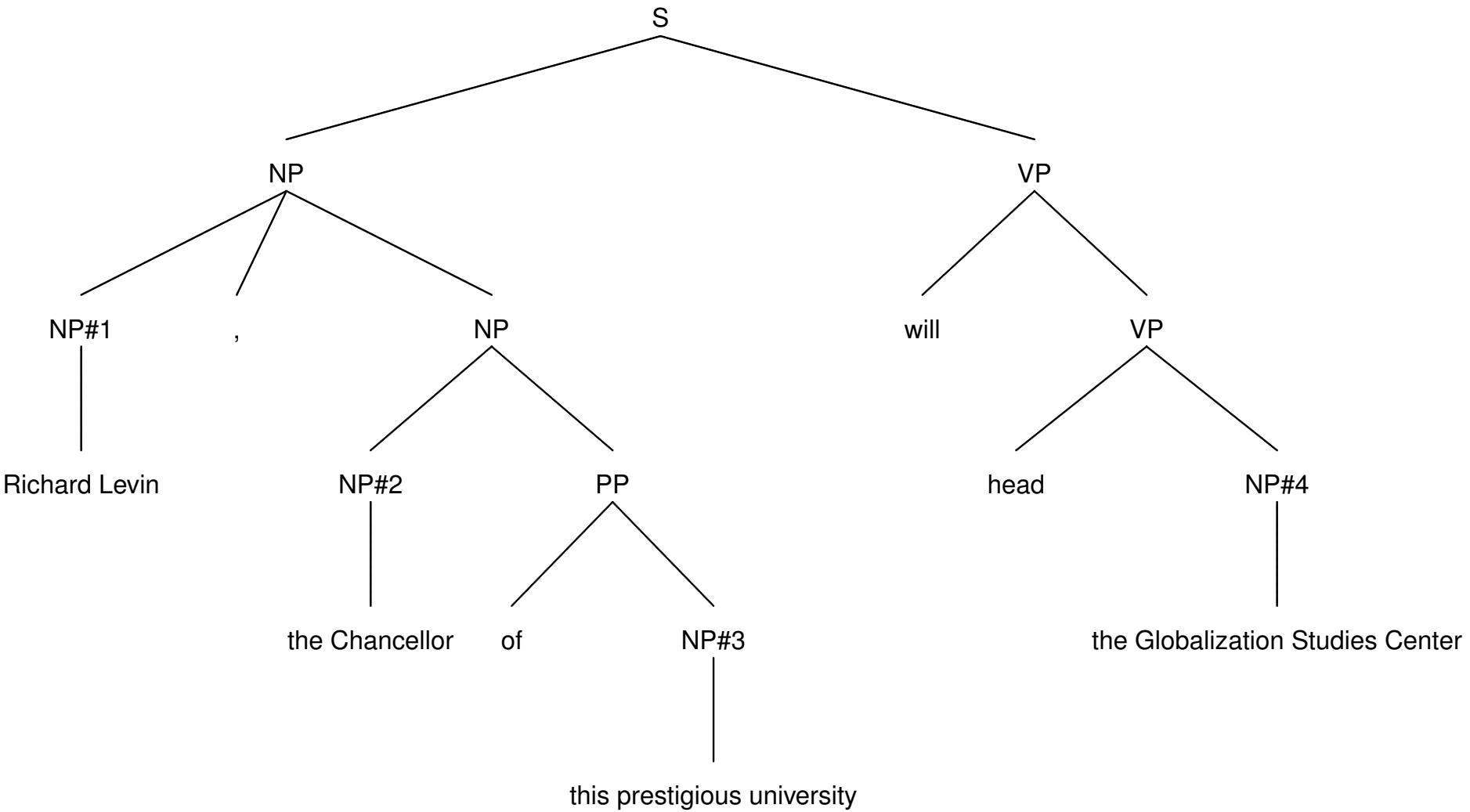
| | English | | | | Spanish | | | | Dutch | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Outer | | Inner | | Outer | | Inner | | Outer | | Inner | |
| Features | F1 | T-F1 | F1 | T-F1 | F1 | T-F1 | F1 | T-F1 | F1 | T-F1 | F1 | T-F1 |
| Local | 41.83 | 54.17 | 48.57 | 57.85 | 34.42 | 42.95 | 37.14 | 41.93 | 48.49 | 54.84 | 49.77 | 55.86 |
| best-clusters | 54.04 | 65.96 | 63.72 | 71.13 | 56.78 | 62.55 | 59.77 | 63.04 | 59.94 | 66.03 | 60.27 | 65.42 |
| best-overall | **55.48** | 67.36 | **64.95** | 71.98 | **58.94** | 65.63 | **62.14** | 65.54 | **63.40** | 70.68 | **63.93** | 70.94 |
| Stanford NER | 53.14 | 64.62 | 62.45 | 69.76 | 46.42 | 54.40 | 47.48 | 54.27 | - | - | - | - |
| Illinois NER | 53.24 | 65.68 | 62.72 | 71.04 | - | - | - | - | - | - | - | - |
| Freeling 3.1 | - | - | - | - | 38.27 | 48.06 | 40.93 | 46.52 | - | - | - | - |
| Sonar ner | - | - | - | - | - | - | - | - | 48.60 | 53.60 | 48.44 | 52.79 |

# OTE at ABSA SemEval 2014 and 2015

This place is not good enough, especially the service is disgusting.

| System (type) | Precision | Recall | F1 score |
|---|---|---|---|
| Baseline | 55.42 | 43.4 | 48.68 |
| EliXa (u) | 68.93 | 71.22 | **70.05** |
| NLANGP (u) | 70.53 | 64.02 | 67.12 |
| EliXa (c) | 67.23 | 66.61 | **66.91** |
| IHS-RD-Belarus (c) | 67.58 | 59.23 | 63.13 |

# ixa-pipe-parse



# ixa-pipe-parse

| Constituent Parsing | English | Spanish |
|---|---|---|
| ixa-pipe-parse | 87.42 | 87.8* |
| Collins | 88.1 | 85.0* |
| Stanford PCFG | 85.5 | n/a |
| St. Factored | 86.6 | n/a |
| St. PCFG Factored | 89.4 | n/a |
| St. CVG (SURNN) | 90.4 | n/a |
| Berkeley | 90.1 | n/a |

# Third-party tools

- Corefgraph: Rule-based coreference resolution for English and Spanish.
- Unsupervised WSD with UKB.
- SRL + Dependencies with Mate tools.
- NED with DBpedia Spotlight.

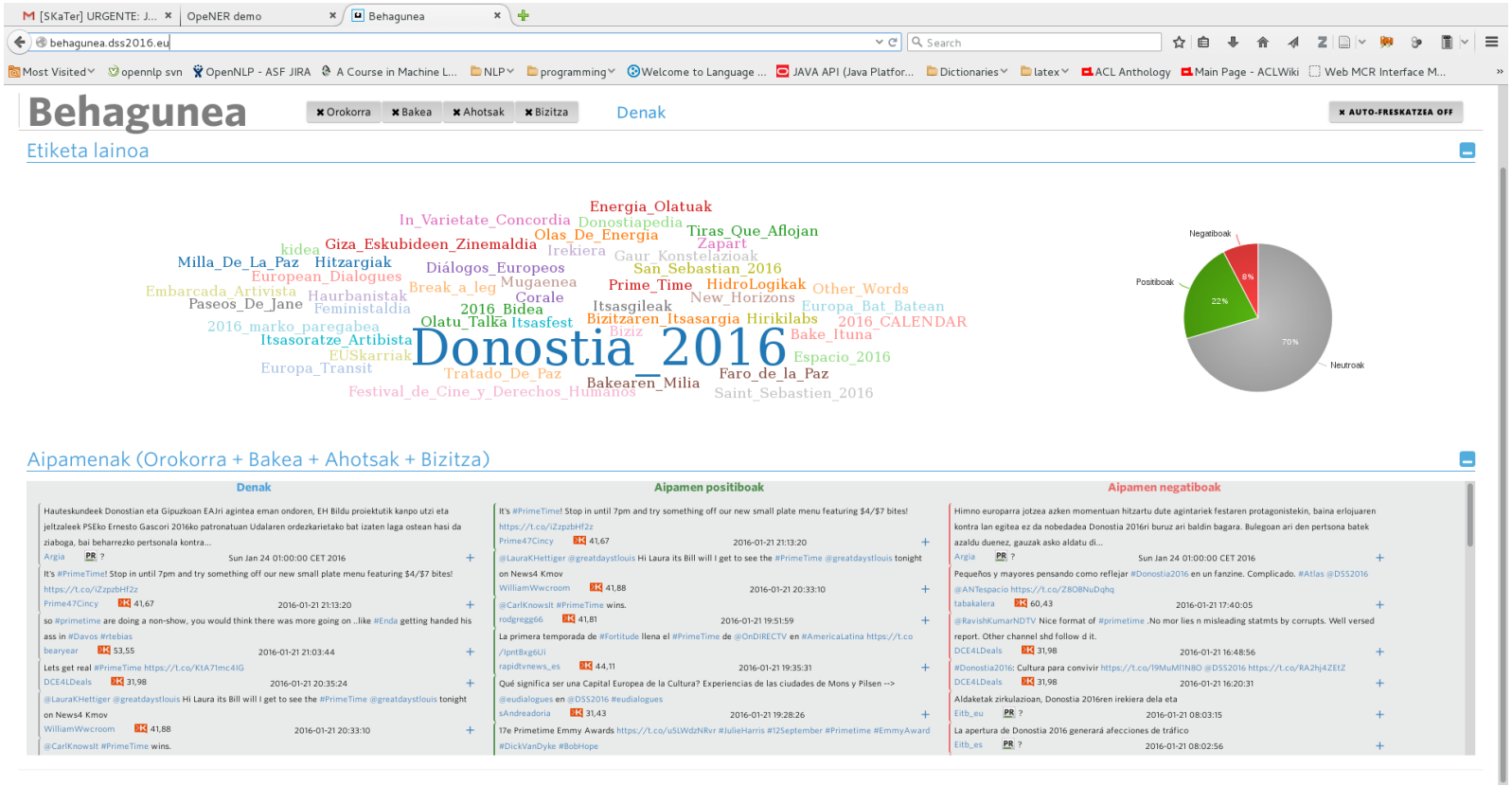
http://ixa2.si.ehu.es/ixa-pipes/third-party-tools.html

# Used in

- OpeNER: http://www.opener-project.eu/
- Newsreader: http://www.newsreader-project.eu/
- QTLeap: http://qtleap.eu/
- Limousine: http://limosine-project.eu/
- Spanish Administration
- Trivago, Olery, Vicomtech-IK4, Elhuyar...
- DSS2016 http://behagunea.dss2016.eu/

# DSS2016 Behagunea

http://behagunea.dss2016.eu/



# Conclusion

- More languages, more annotations.
- Easy to use, easy to train, easy to adapt, easy to deploy.
- Server mode.
- State of the art performance.
- Free software and industrial friendly.

Cross fertilization with Apache Software Foundation's OpenNLP project.