# IXA Pipes: Ready to Use Multilingual NLP tools

Rodrigo Agerri and German Rigau

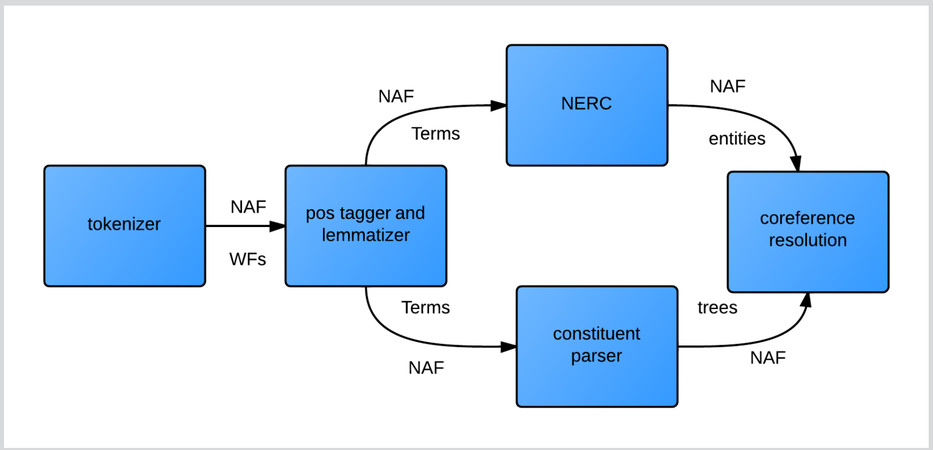
IXA NLP Group, Univ. of the Basque Country (UPV/EHU), Donostia-San Sebastián

## 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 JVM 1.8+ is required.
- Modular data-centric architecture: The tools behave like Unix pipes; easily replaceable and extensible architecture.
- Multilingual and more languages more coming soon!!
- Accurate: State of the art results.
- APL 2.0: To facilitate integration also with commercial applications.

## Architecture




- 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`.

## Linguistic Processors

| ixa-pipe | de | en | es | eu | fr | gl | it | nl |
|---|---|---|---|---|---|---|---|---|
| tokenizer | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| POS tagger | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| lemmatizer | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| NERC | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ |
| Chunker | | ✓ | | ✓ | | | | |
| parsing | | ✓ | ✓ | | | | | |
| coreference | | ✓ | ✓ | | | | | |
| Opinions | | ✓ | | | | | | |
| SST | | ✓ | | | | | | |

- Version 2.0: ca, pl, pt, ru.
- Version 2.0: Aspect-based Sentiment Analysis, Semantic Role Labeling.
- Common local + distributional semantic features.
- Server mode, Maven Central, Common API ...

## ixa-pipe-tok

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

- Tested for ca, de, en, es, eu, fr, gl, it, nl.
- Treebank normalization: Ancora, Penn Treebank, Universal Dependencies normalized tokenization ...
- Paragraph treatment, character offsets, whitespace tokenizer.
- Rule-based, regular expressions.

## ixa-pipe-pos

```
<term id="t69" type="open" lemma="announce" pos="V" morphofeat="VBN">
```

| | | |
|---|---|---|
| rosa | rosa | AQ0CS0 |
| rosa | rosa | NCFS000 |
| rosado | rosa | NCMS000 |
| ... | ... | ... |

- Perceptron models Collins (2002).
- State of the art results.
- Distributional semantic features for morphological analysis in highly-inflected languages.

## ixa-pipe-chunk

[NP He ] [VP reckons ] [NP the current account deficit ] [VP will narrow ] ...

- Available for Basque and English (94.50 F1 on CoNLL 2000).

## 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 |

R. Agerri and G. Rigau (2016). Robust multilingual Named Entity Recognition with shallow semi-supervised features. *Artificial Intelligence*, 238, 63-82.

## ixa-pipe-opinion

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

| ABSA 2015 | 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 |

I. San Vicente, X. Saralegi and R. Agerri (2015) EliXa: A modular and flexible ABSA platform. In SemEval 2015, pp. 748-752.

## 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 |

- Bottom-up shift reduced parser (Apache OpenNLP based).
- * Trained on different subsets of the Ancora corpus.

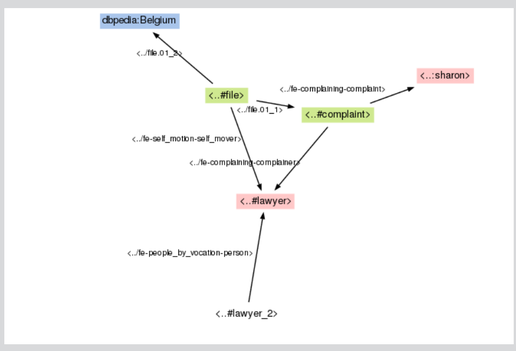
## Third-party tools

- Rule-based multilingual coreference system with Corefgraph (https://bitbucket.org/Josu/corefgraph).
- Unsupervised WSD with UKB (ixa2.si.ehu.es/ukb).
- SRL+Dependencies with Mate tools (code.google.com/p/mate-tools/).
- NED with DBpedia Spotlight (spotlight.dbpedia.org).

## Used by

- European projects: OpeNER: http://www.opener-project.eu/, Newsreader: http://www.newsreader-project.eu/, QTLeap: http://qtleap.eu/, Limousine: http://limosine-project.eu/
- Industry: Trivago, Olery, Vicomtech-IK4, Elhuyar...
- Administration: SETSI...

## Annotation Example




- Lawyers have filed a complaint against Sharon in Belgium.
- Entities: Belgium and Sharon.
- *lawyers* → *file* → *against Sharon*
- *lawyers* → *file* → *in Belgium*
- ...

## Acknowledgements

Presented at enetCollect CA16105 (H2020 funded). Supported by: OpeNER FP7 project under Grant No. 296451, NewsReader FP7 project, Grant No. 316404, and by MINECO SKATER, TIN2012-38584-C06-01 and TUNER, TIN2015-65308-C5-1-R projects.

**Contact: rodrigo.agerri@ehu.eus**

**ixa2.si.ehu.eus/ixa-pipes**