

NewsReader

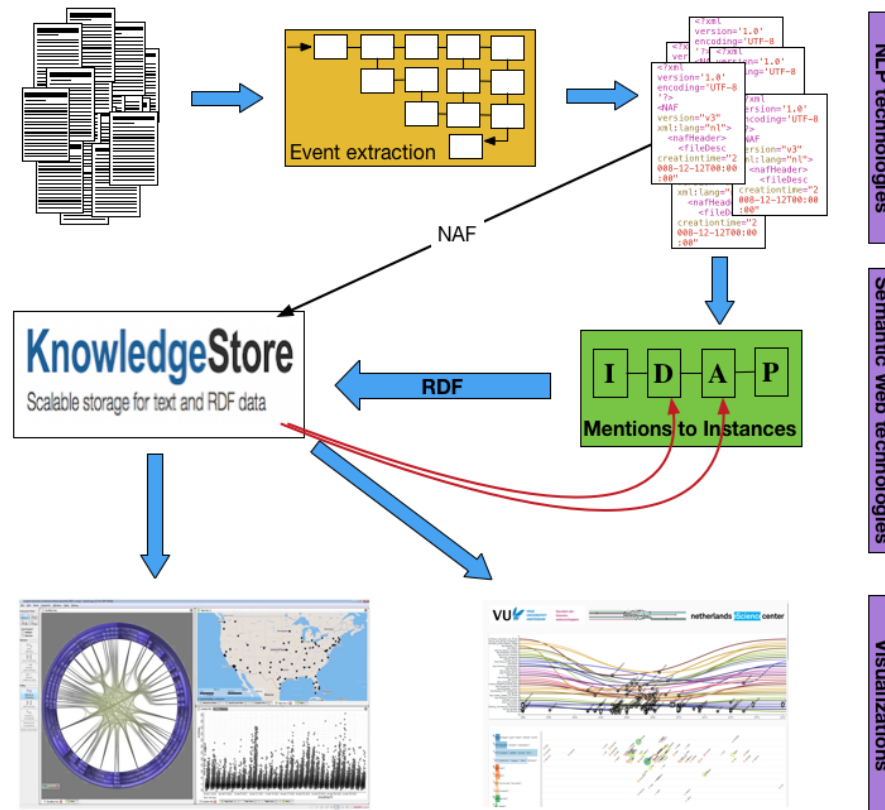
Building structured event indexes
of large volumes of financial and economic data
for making decisions
FP7-2012-ICT-315404

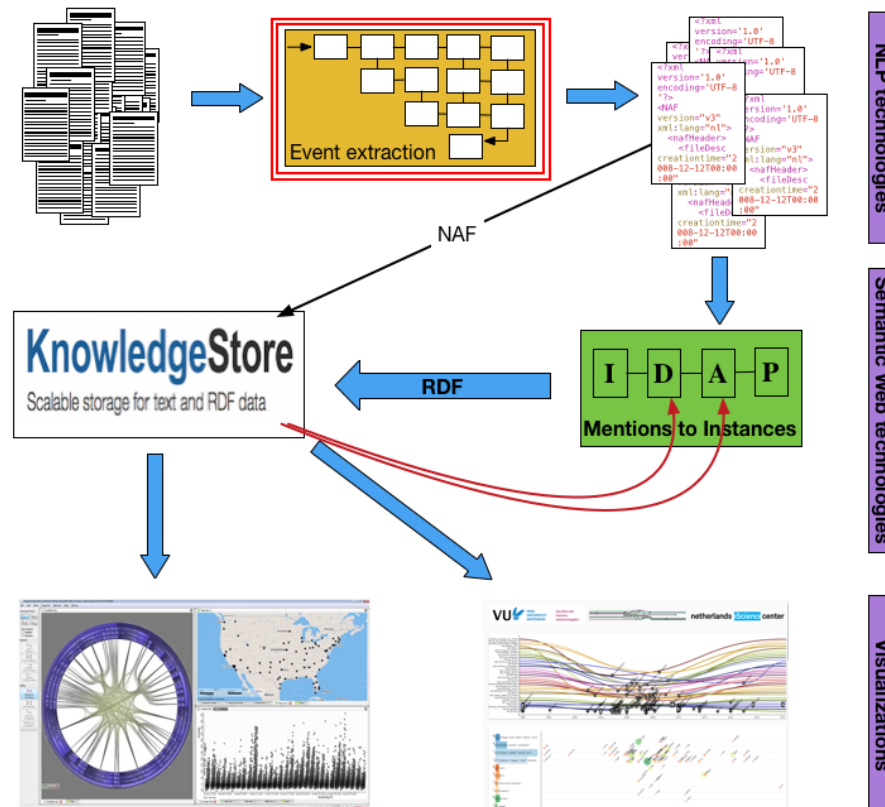
Event Detection

Itziar Aldabe / German Rigau / Egoitz Laparra

IXA group, UPV/EHU

<http://ixa.si.ehu.es>





Project Objectives

- Event Detection in English, Dutch, Spanish and Italian
 - events in terms of *who* did *what* *when* and *where*
 - relations between events
 - factual / non-factual or speculative
 - provenance: who tells what and when
- Narrative schemas and storylines over longer periods of time
- Event reasoning: richness, coherence, relevance
- Large-scale processing, storage and retrieval of events (integrate new with old)

Project Objectives

- **Event Detection in English, Dutch, Spanish and Italian**
 - **events in terms of *who did what when and where***
 - **relations between events**
 - **factual / non-factual or speculative**
 - **provenance: who tells what and when**
- Narrative schemas and storylines over longer periods of time
- Event reasoning: richness, coherence, relevance
- Large-scale processing, storage and retrieval of events (integrate new with old)

English news



June 6, 2005

Apple Computer CEO and co-founder Steve Jobs gave his annual opening speech to the World Wide Developers Conference (WWDC) at Moscone Center in San Francisco, California on Monday

NewsReader: Event detection

Spanish news



6 de junio de 2005

Steve Jobs, cofundador y CEO de Apple Computer, ofreció el lunes su conferencia inaugural anual de la World Wide Developers Conference (WWDC) celebrada en el Moscone Center de San Francisco (California)

06-Jun-05

Apple Computer CEO en medeoprichter Steve Jobs gaf afgelopen maandag zijn jaarlijkse opening keynote tijdens de World Wide Developers Conference (WWDC), die wordt gehouden in het Moscone Center in San Francisco, Californië.

6 giugno 2005

Lunedì, il CEO di Apple Computer e cofondatore Steve Jobs ha tenuto il suo discorso di apertura annuale alla World Wide Developers Conference (WWDC), presso il Moscone Center di San Francisco, California.

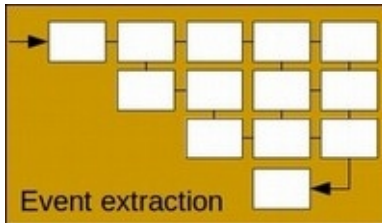


Dutch news



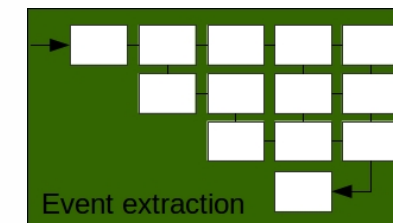
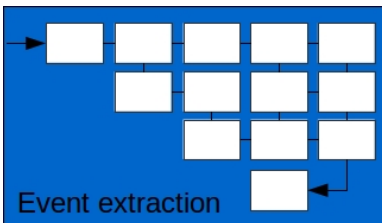
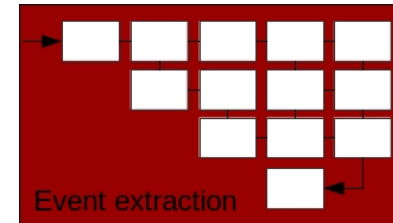
Italian news

English news



NewsReader: Event detection

Spanish news



Dutch news



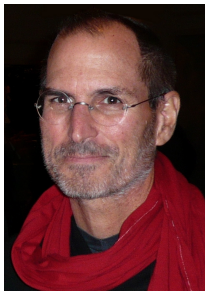
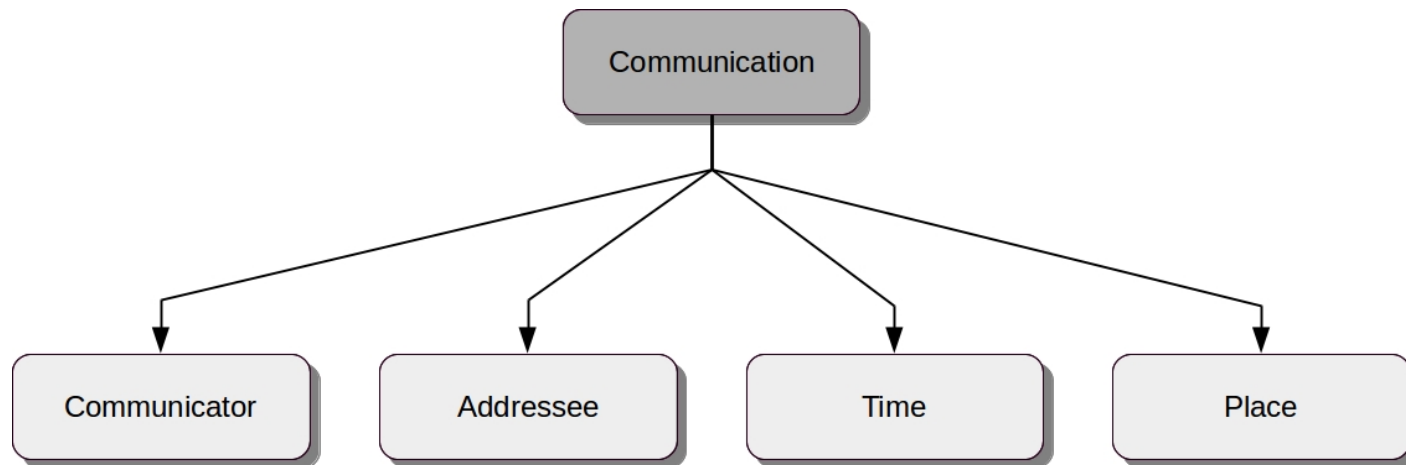
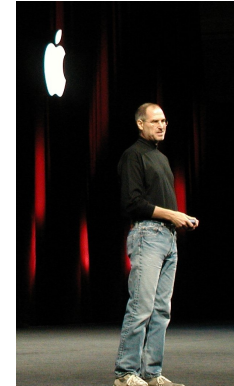
Italian news

Event Detection: Main goals

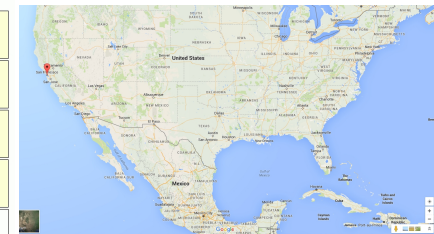
- **Events**, **participants**, their **roles** in text
- **Time** and **place** expressions
- **Attribution** of the events (aka *Authority* and *Factuality*)
- Standard **output** and **scaling**

June 6, 2005

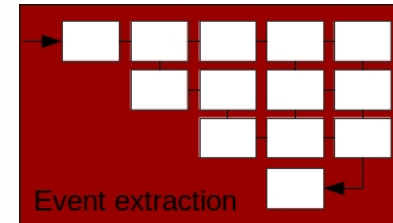
Apple Computer CEO and co-founder Steve Jobs gave his annual **opening speech** to the World Wide Developers Conference (WWDC) at Moscone Center in San Francisco, California on Monday



June 2005						
Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		



Spanish news



Event extraction



Italian news

Event Detection: Example

David Cameron announced yesterday in London that the budget cuts will continue next year.

Event Detection: Example

David Cameron **announced** yesterday in London that the budget cuts will continue next year.

event1: announced

Event Detection: Example

David Cameron **announced** yesterday in London that the budget cuts will continue next year.

event1: announced

who: ?

what: ?

when: ?

where: ?

Event Detection: Example

David Cameron **announced** yesterday in London that the budget cuts will continue next year.

event1: announced

who: David Cameron

what: that the budget cuts will continue next year

when: yesterday

where: in London

Event Detection: Example

David Cameron announced yesterday in London that the budget cuts will **continue** next year.

event1: announced

who: David Cameron

what: that the budget cuts will continue next year

when: yesterday

where: in London

event3: continue

what: ?

when: ?

Event Detection: Example

David Cameron announced yesterday in London that the budget cuts will **continue** next year.

event1: announced

who: David Cameron

what: that the budget cuts will continue next year

when: yesterday

where: in London

event3: continue

what: the budget cuts

when: next year

Event Detection: Example

David Cameron announced yesterday in London that the budget **cuts** will continue next year.

event1: announced

who: David Cameron

what: that the budget cuts will continue next year

when: yesterday

where: in London

event2: cuts

what: ?

event3: continue

what: the budget cuts

when: next year

Event Detection: Example

David Cameron announced yesterday in London that the budget **cuts** will continue next year.

event1: announced

who: David Cameron

what: that the budget cuts will continue next year

when: yesterday

where: in London

event2: cuts

what: budget

event3: continue

what: the budget cuts

when: next year

Event Detection: Example

David Cameron **announced** yesterday in London that the budget cuts will continue next year.

event1: announced

who: David Cameron

what: that the budget cuts will continue next year

when: yesterday

where: in London

Event Detection: Example

David Cameron **announced** yesterday in London that the budget cuts will continue next year.

event1: announced

who: **David Cameron** NERC (PER)

what: that the budget cuts will continue next year

when: yesterday

where: in **London** NERC (LOC)

Event Detection: Example

David Cameron **announced** yesterday in London that the budget cuts will continue next year.

event1: announced

who: David Cameron

- http://dbpedia.org/resource/David_Cameron NED

- what: that the budget cuts will continue next year

when: yesterday

where: in London

- <http://dbpedia.org/resource/London> NED

Event Detection: Example

Bush announced yesterday in London that the budget cuts will continue next year.

- http://dbpedia.org/page/George_H._W._Bush
- http://dbpedia.org/page/George_W._Bush
- http://dbpedia.org/page/Jeb_Bush

Event Detection: Example

David Cameron **announced** yesterday in London that the budget cuts will continue next year.

event1: announced

who: David Cameron

what: that the budget cuts will continue next year

when: yesterday

▪ 2016-03-09 TIMEX3

where: in London

Event Detection: Example

David Cameron **announced** yesterday in London that the budget cuts will continue next year.

event1: announced => **announce.01**

who: Mariano Rajoy

- **Arg0**-PAG: announcer (**vnrole**: 37.7-1-Agent, 48.1.2-Agent)

what: that the budget cuts will continue next year

- **Arg1**-PPT: utterance (**vnrole**: 37.7-1-Topic, 48.1.2-Theme)

when: yesterday

- **AM-TMP**

where: in Madrid

- **AM-LOC**

SRL + Predicate Matrix

Event Detection: NLP tools

David Cameron announced yesterday in London that the budget cuts will continue next year.

NERC NED COREF EVENTS PARTICIPANTS TIME ...
FACTUALITY ATTRIBUTION OPINION ...

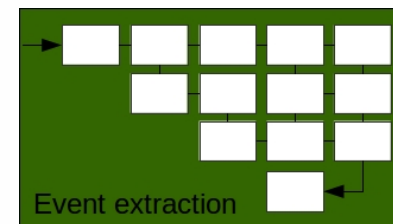
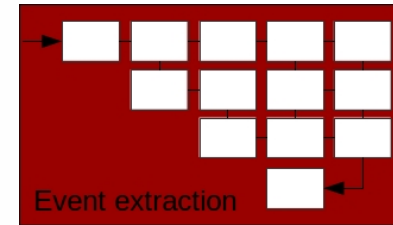
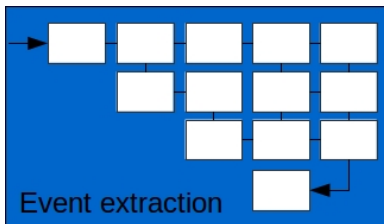
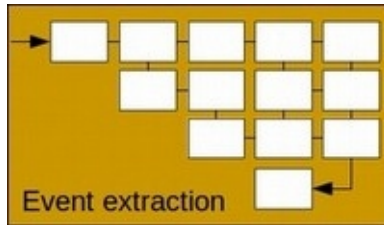
Outline

- EN, SP, IT and NL pipelines
- Cross-lingual interoperability
- Benchmarking

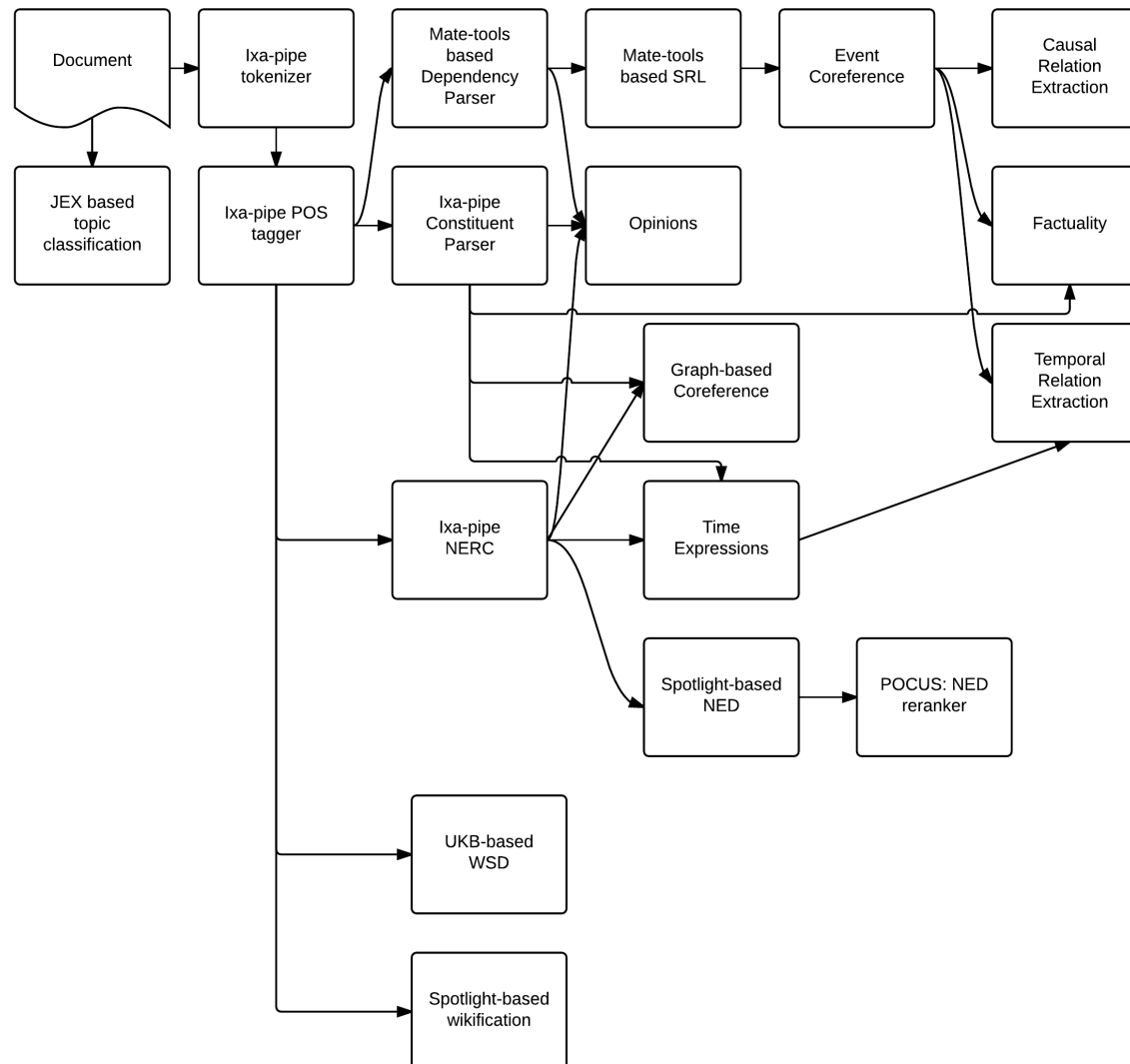
Standards for inter-operability

- NAF
 - Basic format for inter-document NLP analysis
 - Stand-off XML, multi-layered annotation format
 - Allows parallel processing
 - Covers many linguistic levels
 - All NLP modules read and write NAF

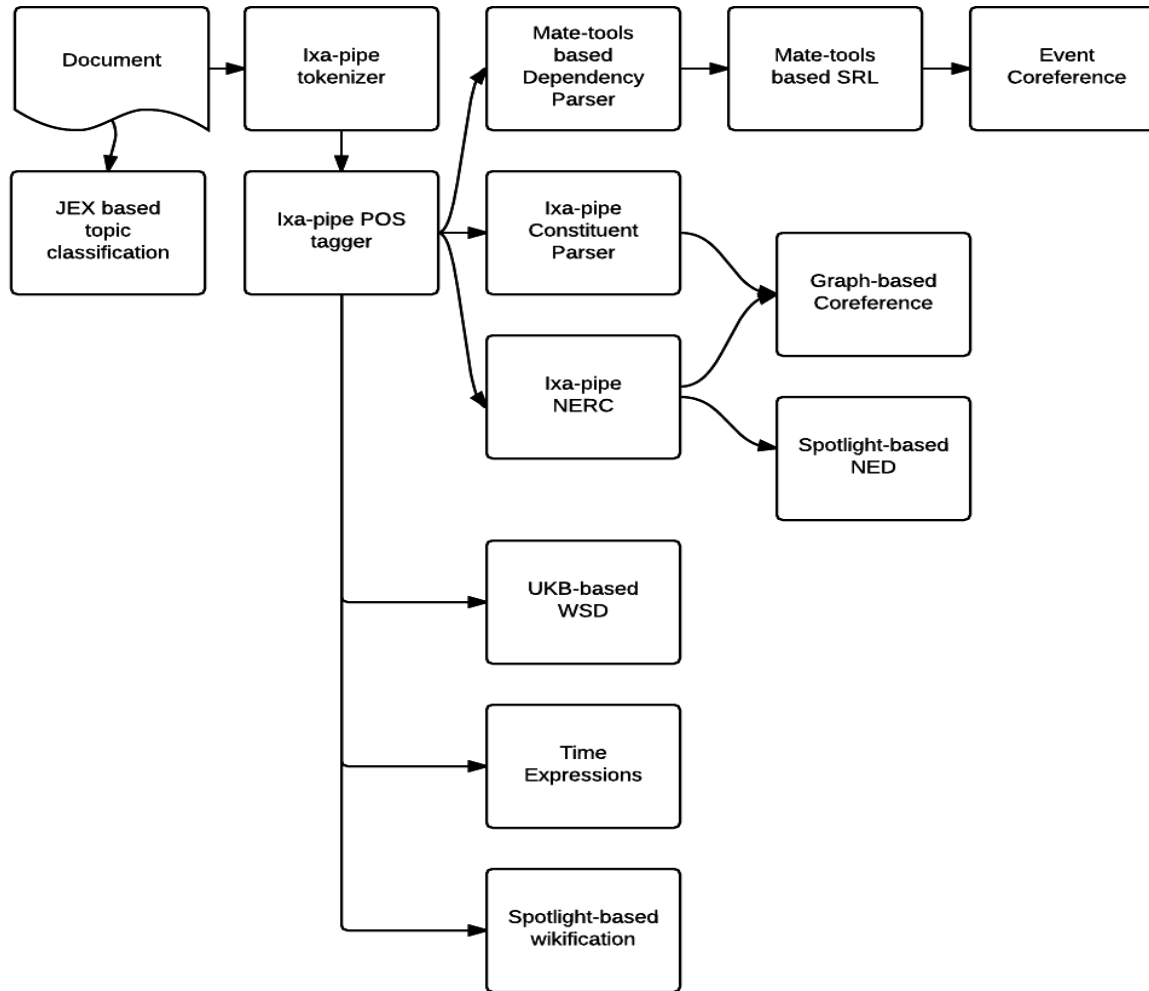
English news



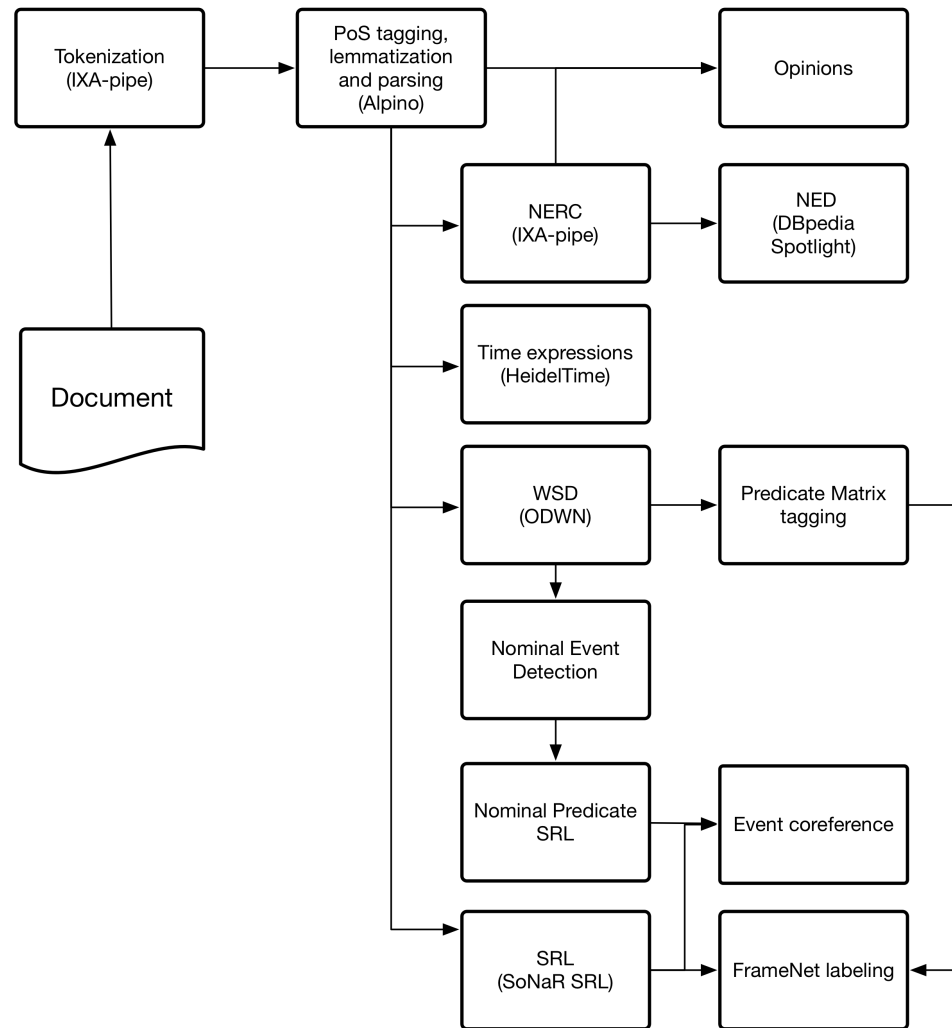
English generic pipeline v3.0



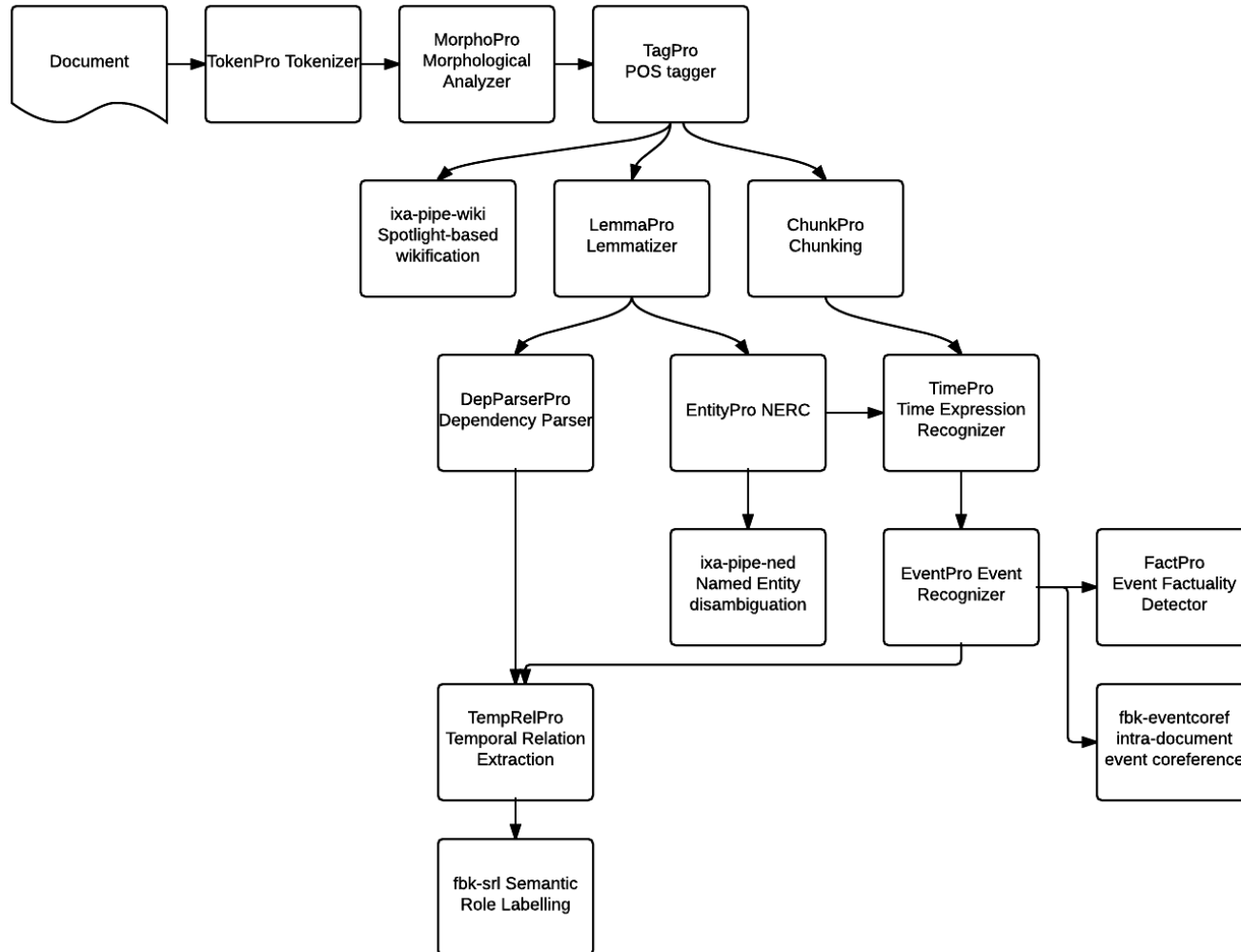
Spanish generic pipeline v3.0



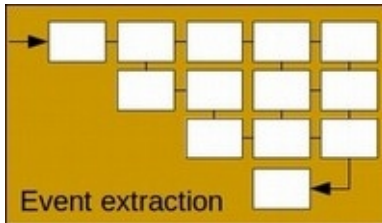
Dutch generic pipeline v3.0



Italian generic pipeline v3.0

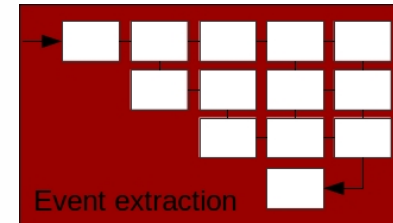


English news

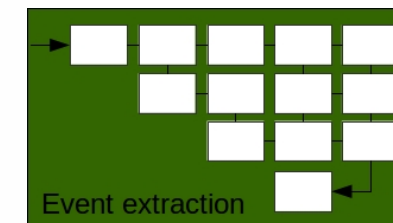
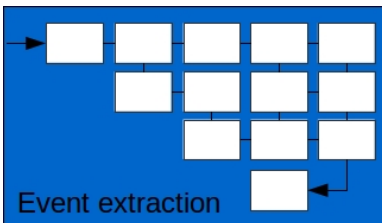


NewsReader: Event detection

Spanish news



??????

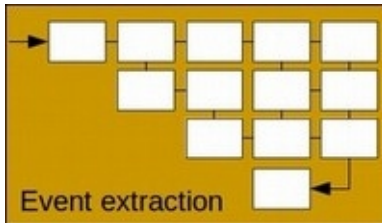


Dutch news



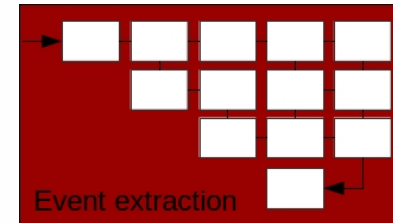
Italian news

English news

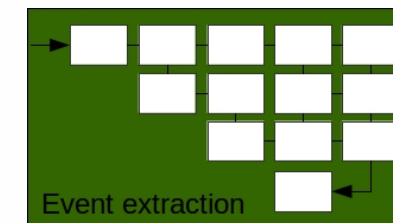
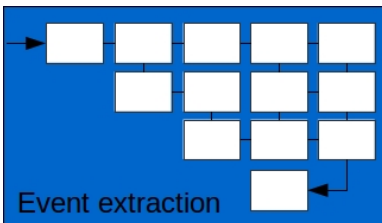


NewsReader: Event detection

Spanish news



ISO-TimeML



Dutch news



Italian news

Cross-lingual interoperability: Time

David Cameron announced **yesterday** in London that the budget cuts will continue **next year**.

```
<timeExpressions>
  <timex3 id="tmx0" type="DATE" functionInDocument="CREATION_TIME" value="2016-03-10"/>
  <timex3 id="tmx1" type="DATE" value="2016-03-09">
    <!--yesterday-->
    <span>
      <target id="w4"/>
    </span>
  </timex3>
  <timex3 id="tmx2" type="DATE" value="2017">
    <!--next year-->
    <span>
      <target id="w13"/>
      <target id="w14"/>
    </span>
  </timex3>
```

Cross-lingual interoperability: Time

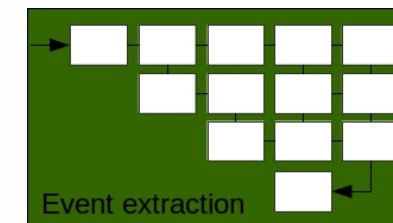
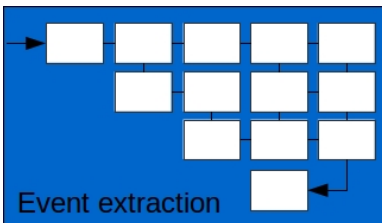
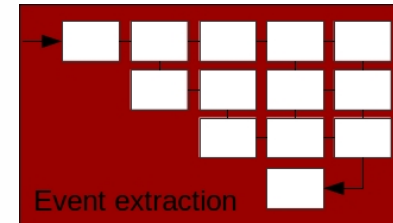
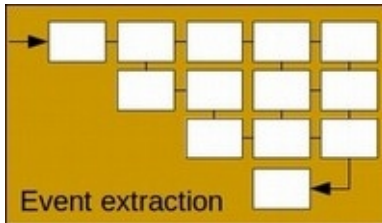
David Cameron anunció **ayer** en Londres que los recortes continuarán **el próximo año**.

```
<timeExpressions>
  <timex3 id="tmx1" type="DATE" functionInDocument="CREATION_TIME" value="2016-03-10" />
  <timex3 id="tmx2" type="DATE" value="2016-03-09">
    <!--ayer-->
    <span>
      <target id="w4" />
    </span>
  </timex3>
  <timex3 id="tmx3" type="DATE" value="2017">
    <!--el pr?ximo a?o-->
    <span>
      <target id="w11" />
      <target id="w12" />
      <target id="w13" />
    </span>
  </timex3>
</timeExpressions>
```

English news

NewsReader: Event detection

Spanish news



Dutch news

Italian news

Cross-lingual interoperability: Named entities

David Cameron announced yesterday in London that the budget cuts will continue next year.

```
<entity id="e1" type="PERSON">
  <references>
    <!--David Cameron-->
    <span>
      <target id="t1"/>
      <target id="t2"/>
    </span>
  </references>
  <externalReferences>
    <externalRef resource="spotlight_v1" reference="http://dbpedia.org/resource/David_Cameron"
confidence="0.9999989" reftype="en" source="en"/>
    <externalRef resource="spotlight_v1"
reference="http://dbpedia.org/resource/Premiership_of_David_Cameron" confidence="1.0534758E-6"
reftype="en" source="en"/>
    <externalRef resource="spotlight_v1"
reference="http://dbpedia.org/resource/David_Cameron_(footballer)" confidence="4.0770318E-11"
reftype="en" source="en"/>
    <externalRef resource="spotlight_v1"
reference="http://dbpedia.org/resource/Dave_Cameron_(footballer)" confidence="3.4414578E-12"
reftype="en" source="en"/>
  </externalReferences>
</entity>
```

Cross-lingual interoperability: Named entities

David Cameron anunció ayer en Londres que los recortes continuarán el próximo año.

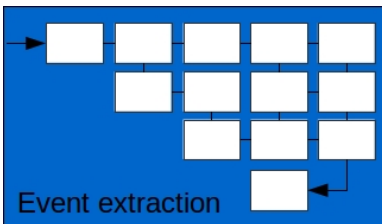
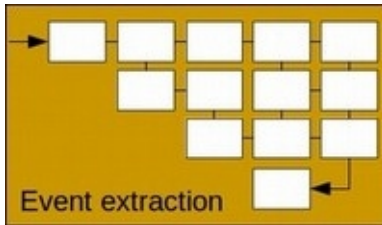
```
<entities>
  <entity id="e1" type="PER">
    <references>
      <!--David Cameron-->
      <span>
        <target id="t1" />
        <target id="t2" />
      </span>
    </references>
    <externalReferences>
      <externalRef resource="spotlight_v1" reference="http://es.dbpedia.org/resource/David_Cameron"
confidence="0.9999967" reftype="es" source="es">
      <externalRef resource="wikipedia-db-esEn"
reference="http://dbpedia.org/resource/David_Cameron" confidence="0.9999967" reftype="en" source="es" />
    </externalRef>
    </externalReferences>
  </entity>
```

Cross-lingual interoperability: Named entities

David Cameron anunció ayer en **Londres** que los recortes continuarán el próximo año.

```
<entity id="e2" type="LOC">
  <references>
    <!-- Londres -->
    <span>
      <target id="t6" />
    </span>
  </references>
  <externalReferences>
    <externalRef resource="spotlight_v1" reference="http://es.dbpedia.org/resource/Londres"
confidence="0.99987996" reftype="es" source="es">
      <externalRef resource="wikipedia-db-esEn" reference="http://dbpedia.org/resource/London"
confidence="0.99987996" reftype="en" source="es" />
    </externalRef>
```

English news



Dutch news

NewsReader: Event detection

English analysis

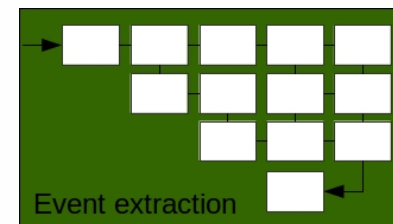
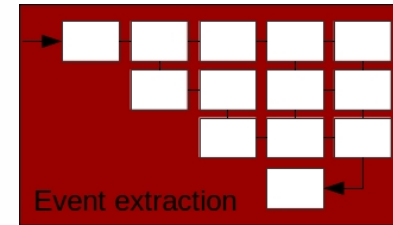
David Cameron announced yesterday in London that the budget cuts will continue next year.

Predicate Matrix	A0	announce-01	AM-TMP	AM-LOC	A1	PropBank
	A0	announcement-01	AM-TMP	AM-LOC	A1	NonBank
Agent	say-37 7-1-1				Topic	VerbNet
Speaker	Statement	Time	Place	Message	FrameNet	
	IntentionalEvent					ESD
	B-30-00974367-v					WordNet
arg0	anuncio-01	arg-imp	arg-loc	arg1	AnCoro-Non	
arg0	anuncio-01	arg-imp	arg-loc	arg1	AnCoro-Verb	

Spanish analysis

David Cameron anunció ayer en Londres que los recortes continuarán el próximo año.

Spanish news



Italian news

Cross-lingual interoperability: Events

- Predicate Matrix is a new lexical resource resulting from the integration of multiple sources of predicate information including FrameNet, VerbNet, PropBank and WordNet.
- <http://adimen.si.ehu.es/web/PredicateMatrix>

Predicate Matrix

English analysis

David Cameron announced yesterday in London that the budget cuts will continue next year.

Predicate Matrix

A0	announce.01	AM-TMP	AM-LOC	A1	PropBank
A0	announcement.01	AM-TMP	AM-LOC	A1	NomBank
Agent	say-37.7-1-1			Topic	VerbNet
Speaker	Statement	Time	Place	Message	FrameNet
	IntentionalEvent				ESO
	ili-30-00974367-v				WordNet
arg0	anuncio.01	arg-tmp	arg-loc	arg1	AnCora-Nom
arg0	anunciar.01	arg-tmp	arg-loc	arg1	AnCora-Verb

David Cameron anunció ayer en Londres que los recortes continuarán el próximo año.

Spanish analysis

Cross-lingual interoperability: Events

David Cameron **announced** yesterday in London that the budget cuts will continue next year.

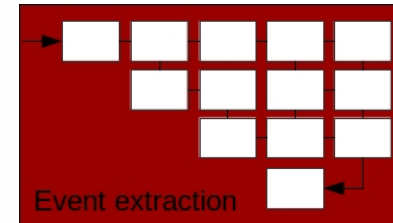
```
<srl>
  <!--t3 announced : A0[t1 David] AM-TMP[t4 yesterday] AM-LOC[t5 in] A1[t7 that]-->
  <predicate id="pr1">
    <!--announced-->
    <span>
      <target id="t3"/>
    </span>
    <externalReferences>
      <externalRef resource="PropBank" reference="announce.01"/>
      <externalRef resource="VerbNet" reference="reflexive_appearance-48.1.2"/>
      <externalRef resource="VerbNet" reference="say-37.7"/>
      <externalRef resource="VerbNet" reference="say-37.7-1"/>
      <externalRef resource="FrameNet" reference="Statement"/>
      <externalRef resource="PropBank" reference="announce.01"/>
      <externalRef resource="EventType" reference="communication"/>
      <externalRef resource="WordNet" reference="ili-30-00974367-v"/>
      <externalRef resource="WordNet" reference="ili-30-00975427-v"/>
    </externalReferences>
    <role id="r11" semRole="A0">
      <!--David Cameron-->
      <span>
        <target id="t1"/>
        <target id="t2" head="yes"/>
      </span>
      <externalReferences>
        <externalRef resource="VerbNet" reference="reflexive_appearance-48.1.2@Agent"/>
        <externalRef resource="VerbNet" reference="say-37.7@Agent"/>
        <externalRef resource="FrameNet" reference="Statement@Speaker"/>
        <externalRef resource="PropBank" reference="announce.01@0"/>
      </externalReferences>
    </role>
  </srl>
```

Cross-lingual interoperability: Events

David Cameron **anunció** ayer en Londres que los recortes continuarán el próximo año.

```
<srl>
  <!--t3 anunci? : arg0[t1 David] argM[t4 ayer] argM[t5 en] arg1[t7 que]-->
  <predicate id="pr1">
    <!--anunci?-->
    <span>
      <target id="t3" />
    </span>
    <externalReferences>
      <externalRef resource="AnCorra" reference="anunciar.1.benefactive" />
      <externalRef resource="VerbNet" reference="reflexive_appearance-48.1.2" />
      <externalRef resource="VerbNet" reference="say-37.7" />
      <externalRef resource="VerbNet" reference="say-37.7-1" />
      <externalRef resource="FrameNet" reference="Statement" />
      <externalRef resource="PropBank" reference="announce.01" />
      <externalRef resource="EventType" reference="communication" />
      <externalRef resource="WordNet" reference="ili-30-00974367-v" />
      <externalRef resource="WordNet" reference="ili-30-00975427-v" />
    </externalReferences>
    <role id="r11" semRole="arg0">
      <!--David Cameron-->
      <span>
        <target id="t1" head="yes" />
        <target id="t2" />
      </span>
      <externalReferences>
        <externalRef resource="VerbNet" reference="reflexive_appearance-48.1.2@Agent" />
        <externalRef resource="VerbNet" reference="say-37.7@Agent" />
        <externalRef resource="FrameNet" reference="Statement@Speaker" />
        <externalRef resource="PropBank" reference="announce.01@0" />
      </externalReferences>
    </role>
  </predicate>
</srl>
```

Spanish news



Event extraction



Italian news

Cross-lingual interoperability: Concepts

A militant **Korean** nationalist has slashed the face of the US ambassador to South Korea at a breakfast meeting in Seoul.

```
<!--Korean-->
<term id="t3" type="open" lemma="korean" pos="G" morphofeat="JJ">
  <span>
    <target id="w3" />
  </span>
  <externalReferences>
    <externalRef resource="wn30g.bin64" reference="ili-30-02967791-a" confidence="1.0" />
    <externalRef resource="WordNet-3.0" reference="ili-30-02967791-a" confidence="1.0" />
  </externalReferences>
</term>
```

Cross-lingual interoperability: Concepts

Un militante nacionalista **coreano** ha rajado la cara del embajador estadounidense en Corea del Sur en un desayuno de trabajo en Seúl.

```
<!-- coreano -->
<term id="t4" type="open" lemma="coreano" pos="G" morphofeat="AQ0MS0">
  <span>
    <target id="w4" />
  </span>
  <externalReferences>
    <externalRef resource="wn30sp.bin64" reference="ili-30-02967791-a" confidence="1.0" />
  </externalReferences>
</term>
```

Cross-lingual interoperability

- **Named entities**
 - **Cross-lingual links** from DBpedia
- **Events**
 - **Predicate Matrix**
 - **Interoperable** event models
 - VN, FN, PB, WN (SUMO, etc.), ESO
 - Extending to **nominal** predicates (EN)
 - **Cross-lingual** PM (EN, ES, NL, IT)
- **Time**
 - **Cross-lingual** time normalization
- **Concepts**
 - **Cross-lingual** wordnets
 - MCR, MultiWordNet, OMW

Open source modules and VMs

- Open source code (>30 modules)
 - <http://github.com/newsreader>
- Virtual Machines for clusters (automatic deployment)
<http://github.com/ixa-ehu/vmc-from-scratch>

NLP Benchmarking

- Intra-document Benchmarking
 - On **standard datasets** (EN, ES, IT, NL)
 - CoNLLs, AIDA, TAC, TempEval, Evalita, etc.
 - **MEANTIME**
 - WikiNews (EN, ES, IT, NL) :
 - 19K English, 8K Italian, 7K Spanish and 1K Dutch
 - 120 documents (EN **manually translated** to ES, IT, NL)
 - 600 sentences per language **manually annotated** at multiple levels
 - Using CAT (intra-document) and CROMER (cross-document)
 - Timelines (SemEval-2015)

NLP benchmarking

- Intra-document Benchmarking
 - **No standard datasets** for all languages and tasks
 - All **advanced** components
 - NERC (EN, ES, IT, NL)
 - NED (EN, ES, IT, NL)
 - Nominal Coref (EN, ES)
 - SRL (EN, ES, IT, NL)
 - Factuality (EN, IT, *ES*)
 - Temporal Normalization (EN, ES, IT, NL)
 - Temporal Relation (EN, IT, *ES*)
 - Event Coref (EN, ES, IT, NL)

			English	Dutch	Italian	Spanish
NERC	Evaluation metric		CoNLL 2003	CoNLL 2003	CoNLL 2003	CoNLL 2003
	Standard	Dataset	CoNLL 2003	CoNLL 2002	EVALITA 2007	CoNLL 2002
		F1	91.18	85.04	82.10	84.16
	MEANTIME		77.18	70.24	56.77	65.54
Nominal coref.	Evaluation metric		CoNLL 2011	-	-	CoNLL 2011
	Standard	Dataset	CoNLL 2011	-	-	SemEval 2010
		F1	71.03	-	-	64.22
	MEANTIME		19.00	-	-	15.74
NED	Evaluation metric		Standard P&R	Standard P&R	Standard P&R	Standard P&R
	Standard	Dataset	AIDA	-	-	TAC 2012
		F1	77.66	-	-	65.11
		Dataset	TAC 2011	-	-	-
		F1	68.92	-	-	-
	MEANTIME		60.26	51.44	60.37	65.87
SRL	Evaluation metric		CoNLL 2009	CoNLL 2009	CoNLL 2009	CoNLL 2009
	Standard	Dataset	CoNLL 2009	-	-	CoNLL 2009
		F1	84.74	-	-	78.85
	MEANTIME		34.78	26.76	31.62	29.68
Time expr.	Evaluation metric		Tempeval3	Tempeval3	Tempeval3	Tempeval3
	Standard	Dataset	TempEval3	-	EVALITA 2014	-
		F1	79.61	-	82.7	-
	MEANTIME		80.50	58.70	85.7	78.30
Temporal relation	Evaluation metric		TempEval3	-	Tempeval3	-
	Standard	Dataset	-	-	EVALITA 2014	-
		F1	-	-	26.4	-
	MEANTIME		22.6	-	13.1	-
Factuality	Evaluation metric		Standard R	-	Standard R	-
	MEANTIME		55.45	-	71.9	-
Event coref.	Evaluation metric		F1	F1	F1	F1
	MEANTIME		41.57	27.32	49.36	30.37

Table 54: Evaluation results on Standard benchmark datasets and NewsReader MEANTIME in the 4 languages of the project

NLP benchmarking (EN)

- Most **complete** evaluation for EN
 - All standard datasets (except Time Relations)
 - All MEANTIME annotations
- **SOA** results in standard datasets
 - NERC (**new** SOA)
- Best results on all tasks, **except** in MEANTIME:
 - NED (ES)
 - Time Detection (IT)
 - Factuality (IT)
 - V-Coref (IT)
- Significant **drop** between standard and MEANTIME

NLP benchmarking (NL)

- Most **complete** evaluation for NL
 - Except N-coref, TempRel, Factuality
 - Standard datasets (only NERC)
- NERC (**new** SOA)
- Slightly lower results on some tasks in MEANTIME (except NERC which obtains the best results)
- **Less resources and datasets** than other languages
- Almost **no standard datasets** for advanced tasks
 - No comparison between standard and MEANTIME

NLP benchmarking (IT)

- Most **complete** evaluation for IT
 - Except N-coref
- NERC (**new** SOA)
- **Best** results in:
 - Time Detection
 - Factuality
 - V-coref
- **No standard datasets** for advanced tasks
 - N-coref
 - NED
 - SRL

NLP benchmarking (ES)

- Most **complete** evaluation for ES
 - Except Temporal Relation, Factuality
- NERC (**new** SOA)
- **Best** results in:
 - NED MEANTIME
- **No standard datasets** for advanced tasks
 - Temporal Relations (incomplete)
 - Factuality

NLP benchmarking

- Most modules perform as **SOA**
- NERC **best results**
 - all languages (EN, ES, IT, NL, DE, EU)
 - **in** and **out** of domain
- **Similar results** across languages
 - when having appropriate linguistic resources and annotation datasets
- **Very difficult** tasks:
 - Time relations, Coref
- Significant **drop** in performance between Standard and MEANTIME on most tasks
 - N-coref: singletons, guidelines, cascading errors
 - SRL: guidelines (TimeML vs. SRL annotation!)

NewsReader

Building structured event indexes
of large volumes of financial and economic data
for making decisions
FP7-2012-ICT-315404

WP4 Event Detection

The team

***Rodrigo Agerri, Itziar Aldabe, Zuhaitz Beloki, Egoitz Laparra,
German Rigau, Aitor Soroa, Mariek van Erp, Antske Fokkens,
Filip Ilievski, Ruben Izquierdo, Roser Morante, Chantal van Son,
Piek Vossen, Anne-Lyse Minard***