Elliphant: A Machine Learning Method for Identifying Subject Ellipsis and Impersonal Constructions in Spanish

Luz Rello

This presentation is based on my Master's thesis and on-going work partly co-authored with Ricardo Baeza-Yates, Ruslan Mitkov and Pablo Suárez

Research Group in Computational Linguistics
University of Wolverhampton

Laboratori fLexSem
Universitat Autònoma de Barcelona

Natural Language Processing Group
Universitat Pompeu Fabra

Outline

— **What**
objective

— **Why**
motivation

— **How**

— **Others?**
literature review

methodology
analysis → learning analysis, genre analysis, feature analysis

results → comparative evaluation

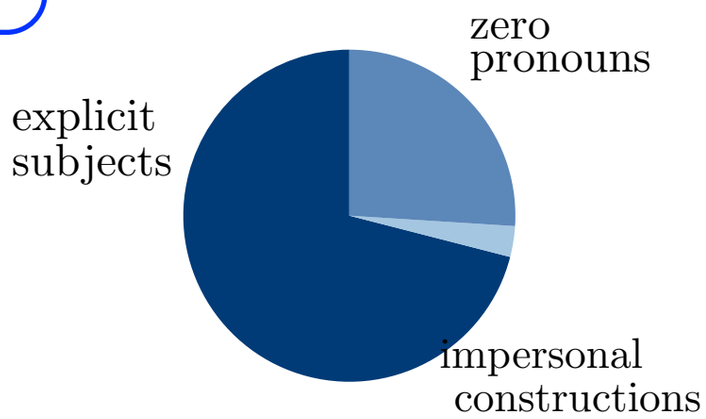
Objective

Elliphant:
fully automatic system → **identify**

identify → machine learning

identify → explicit subjects
identify → zero pronouns
identify → impersonal constructions

explicit subjects → **all** possible constituents in **subject** position → complementary distribution



Motivation

1- Frequent

- Spanish **pro-drop** language (Chomsky, 1981)
- zero pronouns
  - Japanese (Okumura & Tamura, 1996)
  - Chinese (Zhao & Ng, 2007)
  - Korean (Lee & Byron, 2004)
  - Russian (Kibrik, 2004)
  - Spanish (Ferrández & Peral, 2000)
- impersonal constructions
  - English (Evans, 2001)
  - French (Danlos, 2005)
  - **no Spanish**

Motivation

Anaphora resolution (Mitkov, 2010)

2- Crucial

1º — Identification zero pronouns

2º — Filter non-referential impersonal constructions

Motivation

3- Useful

Used in:

— information extraction (Chinchor & Hirschman, 1997)
— machine translation (Peral, 2002)
— automatic summarization (Steinberger et al., 2007)
— text categorization (Yeh & Chen, 2003a)
— topic identification (Yeh & Chen, 2007)
— salience identification (Iida et al., 2009)
— word sense disambiguation (Kawahara & Kurohashi, 2004)
— named entity recognition (Hirano et al., 2007)

Beneficial in:

— parser performance investigation (Foster, 2010)
— discriminative predicate-argument structure analysis (Imamura et al., 2009)
— centering theory (Matsui, 1999) (Takada & Doi, 1994)
— universal convergence of translation (Corpas et al.,2008)

Motivation

4 - Currently needed

"[...] whereas only 80% was achieved for verbs whose subjects were not. This **lower success** rate is justified, however, for several reasons. One important reason is the **non-detection of impersonal verbs** by the POS tagger. "

(Ferrández & Peral, 2000, p.170)

"In contrast with previous work, many of the features relied on gold standard annotations, pointing out the **need for automatic tools for ellipticals detection** and deep parsing"

(Recasens & Hovy, 2009, p.41)

1.a - NLP Approaches to Zero Pronouns

- **Japanese** (Hirano et al., 2007; Iida et al., 2006, 2009; Imamura et al., 2009; etc.)
- **Chinese** (Hu, 2008; Peng & Araki, 2007a,b; Yeh & Chen, 2003a,b, 2007; etc.);
- **Korean** (Han, 2004; Lee & Byron, 2004; Lee et al., 2005);
- **Spanish** (Barreras, 1993; Ferrández & Peral, 2000; Rello & Illisei, 2009a,b, etc.)
- **Russian** (Kibrik, 2004).

- Zero pronoun **classification** or **annotation**:
  (Han, 2004; Kibrik, 2004; Lee & Byron, 2004; Lee et al., 2005, etc.)
- Zero pronoun **identification**:
  (Nakaiwa, 1997; Rello & Illisei, 2009b; Yoshimoto, 1988, etc.)
- **Resolution** of zero pronouns, including their prior **identification**
  (Hu, 2008; Isozaki & Hirao, 2003; Kawahara & Kurohashi, 2004; Zhao & Ng, 2007, etc.)
- Zero pronoun **generation**
  (Peral, 2002; Peral & Ferrández, 2000; Theune et al., 2006; Yeh & Mellish, 1997)

1.a - NLP Approaches to Zero Pronouns

– **Rule-based** approaches
(Barreras, 1993; Corpas Pastor et al., 2008; Ferrández & Peral, 2000; Hu, 2008; Kawahara & Kurohashi, 2004; Kibrik, 2004; Matsui, 1999; Mori & Nakagawa, 1996; Murata et al., 1999; Nakagawa, 1992; Nakaiwa & Ikehara, 1992; Nakaiwa & Shirai, 1996; Nomoto & Yoshihiko, 1993; Peral, 2002; Peral & Ferrández, 2000; Rello & Illisei, 2009b; Yeh & Chen, 2003a,b, 2007; Yeh & Mellish, 1997; Yoshimoto, 1988);

– **Machine-learned** approaches
(Hirano et al., 2007; Iida et al., 2006, 2009; Kawahara & Kurohashi, 2004; Peng & Araki, 2007b; Zhao & Ng, 2007);

– **Hybrid methods** combining rules and learning algorithms (Isozaki & Hirao, 2003);

– **Probabilistic models** (Sasano et al., 2008; Seki et al., 2002); and

– **Other** techniques, such as the exploitation of **parallel corpora** (Nakaiwa, 1997).

1.b - NLP Approaches to Identifying Non-referential Constructions

- **Identification** of pleonastic *it* in **English** (Denber, 1998; Lappin & Leass, 1994; Paice & Husk, 1987). Work by Evans (2000, 2001) is exploited by an anaphora resolution system in Mitkov et al. (2002). Also (Bergsma et al., 2008; Boyd et al., 2005; Clemente et al., 2004; Gundel et al., 2005; Lambrecht, 2001; Li et al., 2009; Müller, 2006; Ng & Cardie, 2002); and

– **Identification** of expletive pronouns in **French** (Danlos, 2005).

1.b - NLP Approaches to Identifying Non-referential Constructions

– **Rule-based** approaches (Danlos, 2005; Denber, 1998; Lappin & Leass, 1994; Paice & Husk, 1987; Rello & Illisei, 2009b;);

– **Machine-learned** approaches (Bergsma et al., 2008; Boyd et al., 2005; Clemente et al., 2004; Evans, 2000, 2001; Mitkov et al., 2002; Müller, 2006; Ng & Cardie, 2002);

– **A web-based** approach (Li et al., 2009); and

– Descriptive studies from contextual (Lambrecht, 2001) and **intonational points** of view (Gundel et al., 2005).

2 - Linguistic Approaches to Subject Ellipsis

– **Semantic**: (Bello, [1847] 1981) and **prescriptive** (Real Academia Española, 2001);

– **Descriptive** and **explicative**: (Brucart, 1999);

– **Distributional**: (Francis, 1958; Fries, 1940);

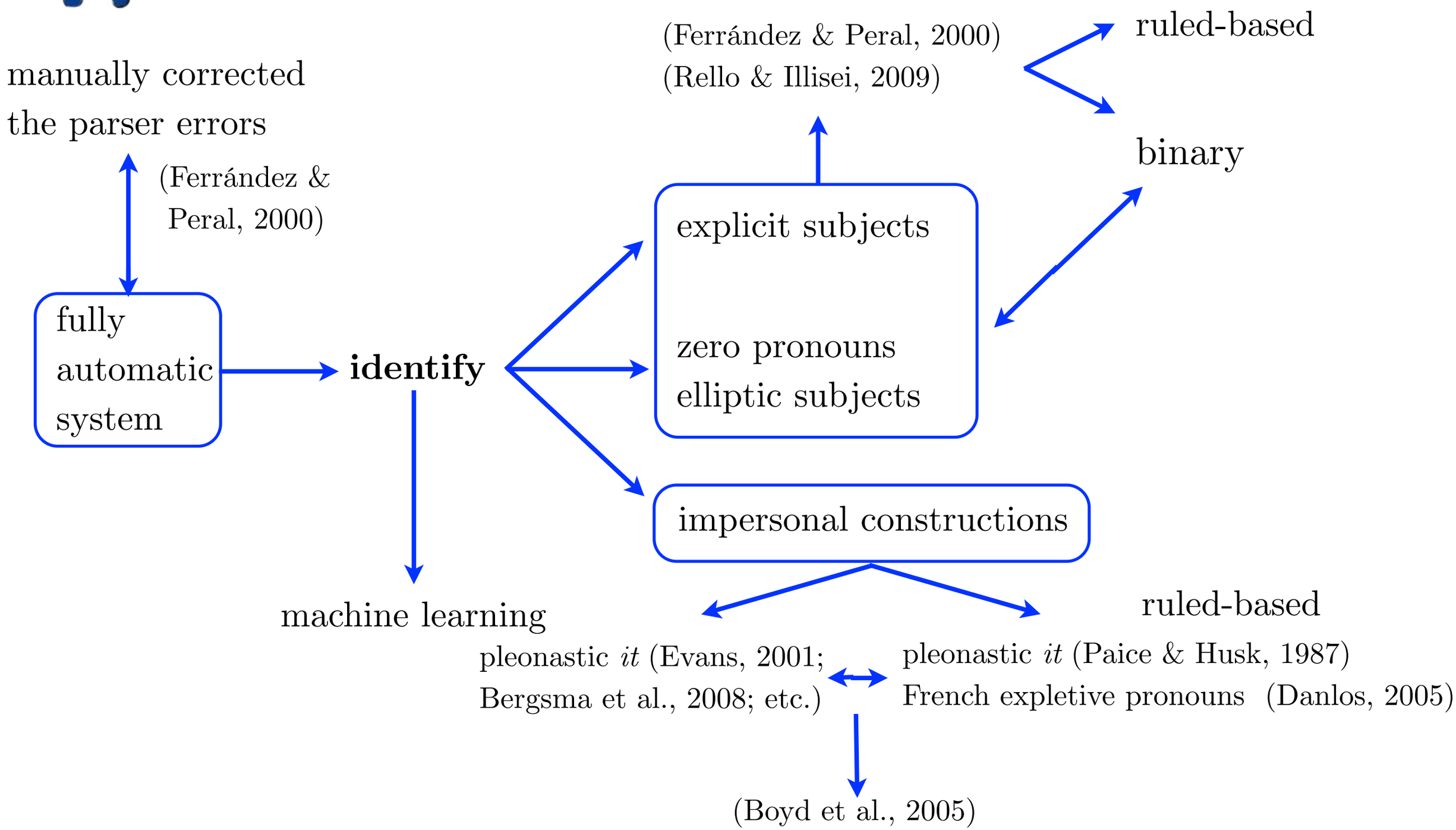
– **Pragmatic**: in diverse pragmatic paradigms, such as implications (Grice, 1975), textual coherence (Halliday & Hasan, 1976), or restrictive code (Shopen, 1973); and

– **Cognitive**: in terms of ellipsis processing by the brain (Streb et al., 2004, p. 175):

> "Ellipses and pronouns/proper names are processed by distinct mechanisms being implemented in distinct cortical cell assemblies."

or as part of the explanation of the **language faculty** (Chomsky, 1965).

Elliphant


manually corrected
the parser errors
(Ferrández & Peral, 2000)
fully automatic system
identify
machine learning
explicit subjects
zero pronouns
elliptic subjects
(Ferrández & Peral, 2000)
(Rello & Illisei, 2009)
ruled-based
binary
impersonal constructions
pleonastic *it* (Evans, 2001; Bergsma et al., 2008; etc.)
(Boyd et al., 2005)
ruled-based
pleonastic *it* (Paice & Husk, 1987)
French expletive pronouns (Danlos, 2005)

Elliphant

1 — Classification

Linguistic Criteria
(Rello, 2010)

Custom-built Software

Annotation

2 — Training data

Features

Corpus Compilation

3 — Analysis

Learning Analysis
Feature Selection
Genre Analysis

(Rello, Baeza-Yates & Mitkov, 2010)

4 — Comparative evaluation
(Rello, Suárez & Mitkov, 2010)

Elliphant

Linguistically motivated classes

1 — Explicit subject → [– elliptic, + referential]

Las fuentes son ***la ley, la costumbre y los principios generales del derecho.***
The sources are ***the law, the judicial costume and the general principles of law.***

2 — Zero pronoun → [+ elliptic, + referential]

Las leyes no tendrán efecto retroactivo si **Ø** no dispusieren lo contrario.
The law will not have a retroactive effect unless ***(they)*** specify otherwise.

3 — Impersonal construction → [+ elliptic, – referential]

Cuando hay un diagnóstico.
When ***(there)*** is a diagnosis.

(Rello, 2010)

Elliphant

Linguistically motivated classes

(Bosque 1989; Real Academia Española, 2009; Brucart 1999)

CLASS	TYPE OF INFORMATION:	SYNTACTIC CATEGORY	VERBAL DIATHESIS	PHONETIC REALIZATION		SEMANTIC INTERPRETATION	DISCOURSE
	LINGUISTIC FEATURES:	[±] Nominal subject	[±] Active	[±] Elliptic noun phrase	[±] Elliptic noun phrase head	[±] Active participant	[±] Referential subject
CLASS 1: **EXPLICIT SUBJECT**	Explicit subject	+	+	–	–	+	+
	Reflex passive subject	+	+	–	–	–	+
	Passive subject	+	–	–	–	–	+
CLASS 2: **ZERO PRONOUN**	Omitted subject	+	+	+	–	+	+
	Omitted subject head	+	+	–	+	+	+
	Reflex passive omitted subject	+	+	+	–	–	+
	Reflex passive omitted subject head	+	+	–	+	–	+
	Passive omitted subject	+	–	+	–	–	+
	Non nominal subject	–	+	–	–	+	+
	Reflex passive non nominal subject	–	+	–	–	–	+
	Passive non nominal subject	–	–	–	–	–	+
CLASS 3: **NON-REFERENTIAL IMPERSONAL SUBJECT**	Impersonal with se	——					–
	Impersonal	——					–

Corpus compilation

ESZIC corpus: **E**xplicit **S**ubjects, **Z**ero-pronouns and **I**mpersonal **C**onstructions

- Two genres:
  - **legal** (laws)
  - **health** (psychiatric papers)
- **Originally** written in **Spanish**

- Avg. 2.3 clauses/sentence
- Avg. 11.7 words/clause
- Avg. 26.9 words/sentence.

ESZIC Corpus	Number of Tokens	Number of Sentences	Number of Clauses
Legal text 1	9,972	941	600
Legal text 2	1,147	47	56
Legal text 3	17,960	1,035	1,181
Legal text 4	3,578	189	191
Legal text 5	12,456	746	891
Legal text 6	3,962	130	219
Legal text 7	2,159	131	136
Legal text 8	5,219	291	282
Health text 1	2,753	110	270
Health text 2	11,339	658	1,028
Health text 3	1,854	47	140
Health text 4	1,937	84	124
Health text 5	2,183	93	148
Health text 6	1,568	63	210
Health text 7	1,296	69	89
Health text 8	1,687	53	127
Health text 9	12,441	525	1,394
Total	**93,511**	**5,212**	**7,086**

ESZIC Corpus parsing

Connexor's Functional Dependency Grammar Parser

(Tapanainen and Järvinen, 1997)

- morphological tags
- syntactic tags
- syntactic relations

(Connexor Oy, 2006a, b)

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE analysis SYSTEM "http://www.connexor.com/dtds/4.0/fdg3.dtd">
<analysis><sentence id="w1">
<token id="w2"> <text>LA</text> <lemma>la</lemma>        <depend head="w3">det</depend>  <tags><syntax>#PREMOD</syntax> <morpho>DET FEM SG</morpho>
<token id="w3"> <text>CONSTITUCIÓN</text>        <lemma>constitución</lemma>      <depend head="w1">main</depend> <tags><syntax>#NH</syntax> <morpho
<token id="w4"> <text>ESPAÑOLA</text>   <lemma>español</lemma>  <depend head="w3">ads</depend>  <tags><syntax>#POSTMOD</syntax> <morpho>A FEM SG</
<token id="w5"> <text>DE</text> <lemma>de</lemma>        <depend head="w6">pm</depend>   <tags><syntax>#POSTMOD</syntax> <morpho>PREP</morpho></tag
<token id="w6"> <text>1978</text>        <lemma>1978</lemma>      <depend head="w3">mod</depend>  <tags><syntax>#NH</syntax> <morpho>NUM CARD</morp
<token id="w7"> <text>.</text>  <lemma>.</lemma></token>
</sentence>
<paragraph/><sentence id="w9">
<token id="w10">        <text>Aprobada</text>   <lemma>aprobada</lemma> <depend head="w22">subj</depend>        <tags><syntax>#NH</syntax> <morpho
<token id="w11">        <text>por</text>        <lemma>por</lemma>      <depend head="w13">pm</depend>  <tags><syntax>#POSTMOD</syntax> <morpho>PR
<token id="w12">        <text>las</text>        <lemma>las</lemma>      <depend head="w13">det</depend> <tags><syntax>#PREMOD</syntax> <morpho>DET
<token id="w13">        <text>Cortes</text>     <lemma>corte</lemma>    <depend head="w10">mod</depend> <tags><syntax>#NH</syntax> <morpho>N FEM P
<token id="w14">        <text>Generales</text>  <lemma>general</lemma>  <depend head="w13">ads</depend> <tags><syntax>#POSTMOD</syntax> <morpho>A
<token id="w15">        <text>el</text> <lemma>el</lemma>        <depend head="w16">det</depend> <tags><syntax>#PREMOD</syntax> <morpho>DET MSC SG
<token id="w16">        <text>31</text> <lemma>31</lemma>                <tags><syntax>#NH</syntax> <morpho>NUM ORD</morpho></tags></token>
<token id="w17">        <text>de</text> <lemma>de</lemma>        <depend head="w18">pm</depend>  <tags><syntax>#POSTMOD</syntax> <morpho>PREP</mor
<token id="w18">        <text>octubre</text>    <lemma>octubre</lemma>  <depend head="w16">mod</depend> <tags><syntax>#NH</syntax> <morpho>N MSC S
<token id="w19">        <text>de</text> <lemma>de</lemma>        <depend head="w20">pm</depend>  <tags><syntax>#POSTMOD</syntax> <morpho>PREP</mor
<token id="w20">        <text>1978</text>       <lemma>1978</lemma>     <depend head="w18">mod</depend> <tags><syntax>#NH</syntax> <morpho>NUM CA
<token id="w21">        <text>,</text>  <lemma>,</lemma></token>
<token id="w22">        <text>fue</text>        <lemma>ser</lemma>      <depend head="w23">v-ch</depend>        <tags><syntax>#AUX</syntax> <morp
<token id="w23">        <text>refrendada</text> <lemma>refrendar</lemma>        <depend head="w9">main</depend> <tags><syntax>#MAIN</syntax> <mor
<token id="w24">        <text>por</text>        <lemma>por</lemma>      <depend head="w26">pm</depend>  <tags><syntax>#PREMARK</syntax> <morpho>PR
<token id="w25">        <text>el</text> <lemma>el</lemma>        <depend head="w26">det</depend> <tags><syntax>#PREMOD</syntax> <morpho>DET MSC SG
<token id="w26">        <text>pueblo</text>     <lemma>pueblo</lemma>   <depend head="w23">agt</depend> <tags><syntax>#NH</syntax> <morpho>N MSC S
<token id="w27">        <text>español</text>    <lemma>español</lemma>  <depend head="w26">ads</depend> <tags><syntax>#POSTMOD</syntax> <morpho>A
```

ESZIC Corpus annotation

Annotation software → *finite verb* with the *clause* ← purpose built clause splitter module ↔ (Rello & Ilisei, 2009b)

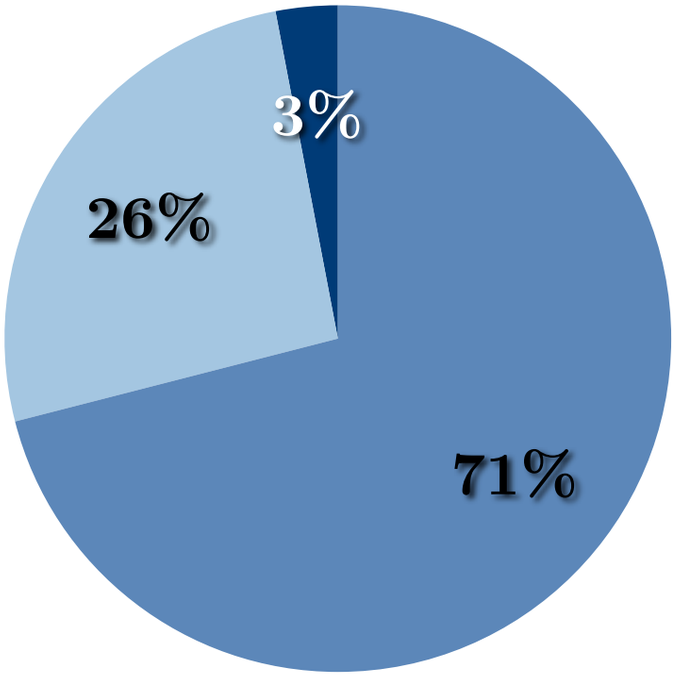
3 human annotators

```
IDLE 1.2
>>> ================================ RESTART ================================
>>>
current file is: Corpus/p2008-vol12-n1-7-v3--89 copia.xml

[A]notar [S]alir [O]mitir a
Longitud texto: 31 parrafos
Con 69 frases
Y 89 verbos.

Resumen

89 w81 Resumen [S|SNN|Z|ZH|SPR|ZPR|SNNPR|ZHPR|SP|ZP|SNNP|I|SE|A|E]:E

Justificación : Pretendemos estudiar un elemento muy presente en nuestro trabajo dia
rio con el paciente psiquiátrico : las emociones .

88 w86 Pretendemos [S|SNN|Z|ZH|SPR|ZPR|SNNPR|ZHPR|SP|ZP|SNNP|I|SE|A|E]:Z
87 w91 presente [S|SNN|Z|ZH|SPR|ZPR|SNNPR|ZHPR|SP|ZP|SNNP|I|SE|A|E]:E

Intentamos determinar las diferencias entre profesionales por sexo , edad , años de
profesión y ámbito de trabajo .

86 w127 Intentamos [S|SNN|Z|ZH|SPR|ZPR|SNNPR|ZHPR|SP|ZP|SNNP|I|SE|A|E]:
```

ESZIC Corpus

- 1- Explicit subject
- 2- Zero pronoun
- 3- Impersonal construction

— 6,827 annotated instances



Number of instances per class	Legal ESZIC Corpus	Health ESZIC Corpus	ESZIC Corpus
Explicit subjects	2,739	2,116	**4,855**
Zero pronouns	619	1,174	**1,793**
Impersonal constructions	71	108	**179**
Total	3,429	3,398	**6,827**

Elliphant

Purpose built tools ⟷ Features

parser **extrinsic** *vs.* parser **intrinsic**

clause splitter/ classifier module

agreement module

noun phrases phrases module

verb lists (11,000) (Real Academia Española, 2001)

	Feature	Definition	Value
intrinsic	**1** PARSER	Parsed subject	True, False
	2 CLAUSE	Clause type	Main, Rel, Imp, Prop, Punct
intrinsic	**3** LEMMA	Verb lemma	Parser's lemma tag
intrinsic	**4** NUMBER	Verb morphological number	SG, PL
intrinsic	**5** PERSON	Verb morphological person	P1, P2, P3
	6 AGREE	Agreement in person, number, tense and mood	FTFF, TTTT, FFFF, TFTF, TTFF, FTFT, FTTF, TFTT, FFFT, TTTF, FFTF, TFFT, FFTT, FTTT, TFFF TTFT
	7 NHPREV	Previous noun phrases	Number of noun phrases previous to the verb
	8 NHTOT	Total noun phrases	Number of noun phrases in the clause
intrinsic	**9** INF	Infinitive	Number of infinitives in the clause
intrinsic	**10** SE	Particle *se*	se, no
intrinsic	**11** A	Preposition *a*	True, False
intrinsic	**12** POS_{pre}	Four parts of the speech previous to the verb	292 different values combining the parser's POS tags, *i.e.:* @HN, @CC, @MAIN, etc.
intrinsic	**13** POS_{pos}	Four parts of the speech speech following the verb	280 different values combining the parser's POS tags, *i.e.:* @HN, @CC, @MAIN, etc.
	14 $VERB_{type}$	Type of verb: copulative, impersonal, pronominal, transitive and intransitive	CIPX, XIXX, XXXT, XXPX, XXXI, CIXX, XXPT, XIPX, XIPT, XXXX, XIXI, CXPI, XXPI, XIPI, XXEX

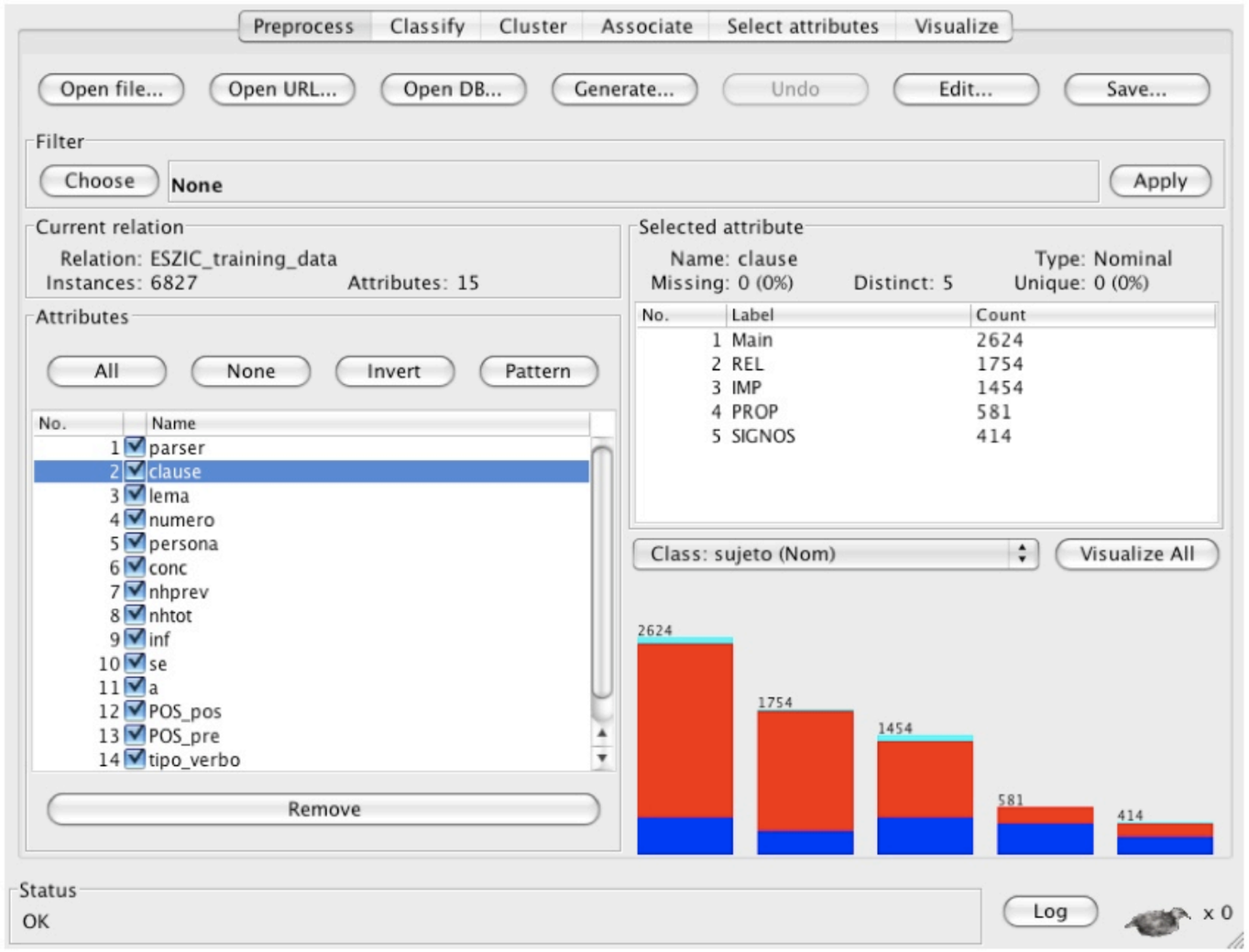
Elliphant

Training data → **Weka 3-6-3**

— training vector for each instance (6,827)

— standard cvs format

state-of-the-art machine learning algorithms (Hall et al., 2009; Witten & Frank, 2005)



Elliphant

Analysis → **Questions**

10-fold cross-validation

1 — Algorithm Selection Parameter Optimization → *Which **method** and **parameter values work best** for our problem?*

2 — Learning Analysis → *How many **instances** are **needed** to train the algorithm?*

3 — Feature Selection → *Which are the most **significant features** and what are the most **effective** combinations of features?*

4 — Genre Analysis → *Does the **genre matter**?*

Algorithm Selection and Parameter Optimization

1 — test 20% training data
2 — test 100% training data
3 — test parameter values

(Cleary & Trigg, 1995)

K* instance-based classifier (lazy)
— **blending parameter** of 40%

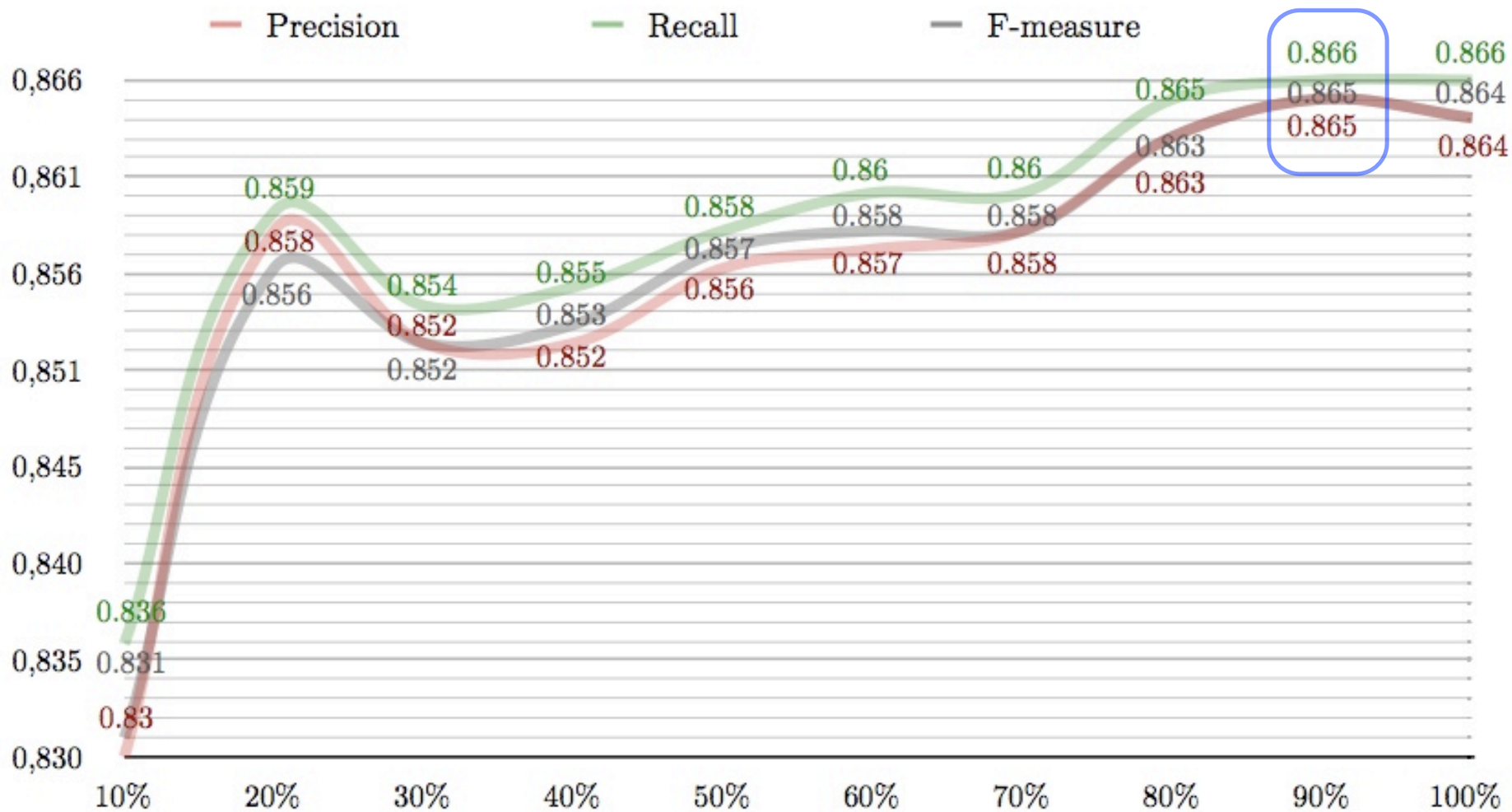
Class	Precision	Recall	F-measure
Explicit subjects	0.900	0.923	0.911
Zero pronouns	0.772	0.740	0.756
Impersonal constructions	0.889	0.626	0.734

ESZIC **training data Accuracy**: 0.867 (ten-fold cross-validation)

WEKA classifiers	Accuracy	WEKA classifiers	Accuracy
Bayes: BayesNet	*0.848*	Meta: RacedIncrementalLogitBoost	0.717
Bayes: NaiveBayes	*0.848*	Meta: RandomSubSpace	0.731
Bayes: NaiveBayesSimple	0.842	Meta: Stacking	0.717
Bayes: NaiveBayesUpdateable	*0.848*	Meta: StackingC	0.717
Functions: RBFNetwork	*0.848*	Meta: Vote	0.717
Lazy: IB1	0.804	Misc: HyperPipes	0.715
Lazy: IBk	0.810	Misc: VFI	0.704
Lazy: K*	**0.850**	Rules: ConjunctiveRule	0.809
Lazy: LWL	0.809	Rules: DecisionTable	0.834
Meta: AdaBoostM1	0.81	Rules: DTNB	0.834
Meta: AttributeSelectedClassifier	0.836	*Rules: JRip*	*0.845*
Meta: ClassificationViaClustering	0.66	Rules: NNge	0.740
Meta: CVParameterSelection	0.717	Rules: OneR	0.762
Meta: Decorate	0.795	Rules: PART	0.795
Meta: END	0.809	Rules: Ridor	0.821
Meta: EnsembleSelection	0.762	Rules: ZeroR	0.717
Meta: FilteredClassifier	0.810	Trees: BFTree	0.760
Meta: Grading	0.717	Trees: DecisionStump	0.810
Meta: LogitBoost	0.841	Trees: J48	0.810
Meta: MultiBoostAB	0.810	Trees: J48graft	0.813
Meta: MultiClassClassifier	0.661	*Trees: LADTree*	*0.846*
Meta: MultiScheme	0.717	*Trees: NBTree*	*0.850*
NestedDichotomies: ClassBalancedND	0.809	Trees: RandomForest	0.793
NestedDichotomies: DataNearBalancedND	0.809	Trees: RandomTree	0.749
NestedDichotomies: ND	0.809	Trees: REPTree	0.723
Meta: OrdinalClassClassifier	0.810	Trees: SimpleCart	0.763

Learning Analysis

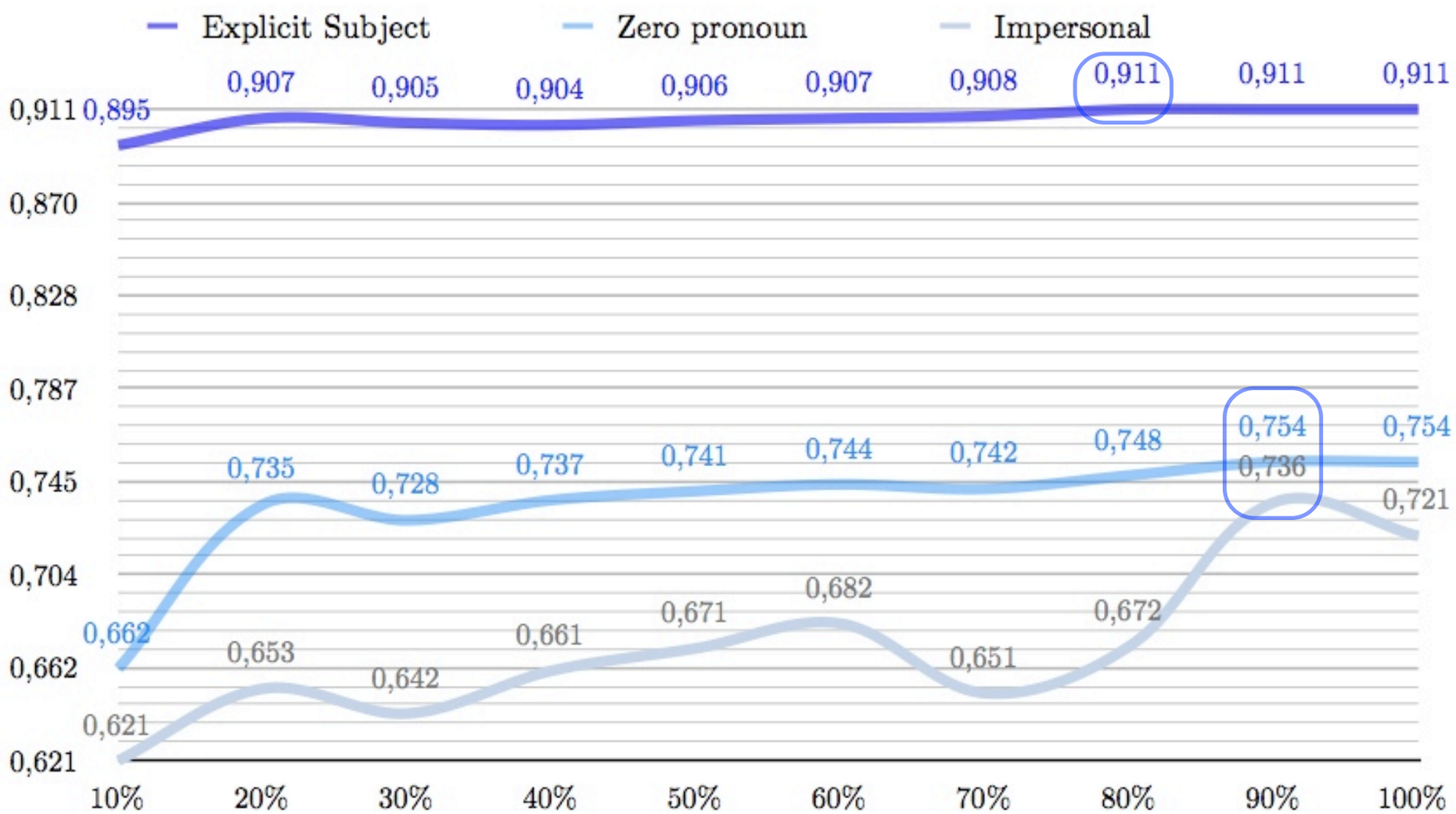
— When is the **plateau** is reached?



ESZIC training data learning curve for precision, recall and f-measure

Learning Analysis

— When is the plateau reached for **each class**?

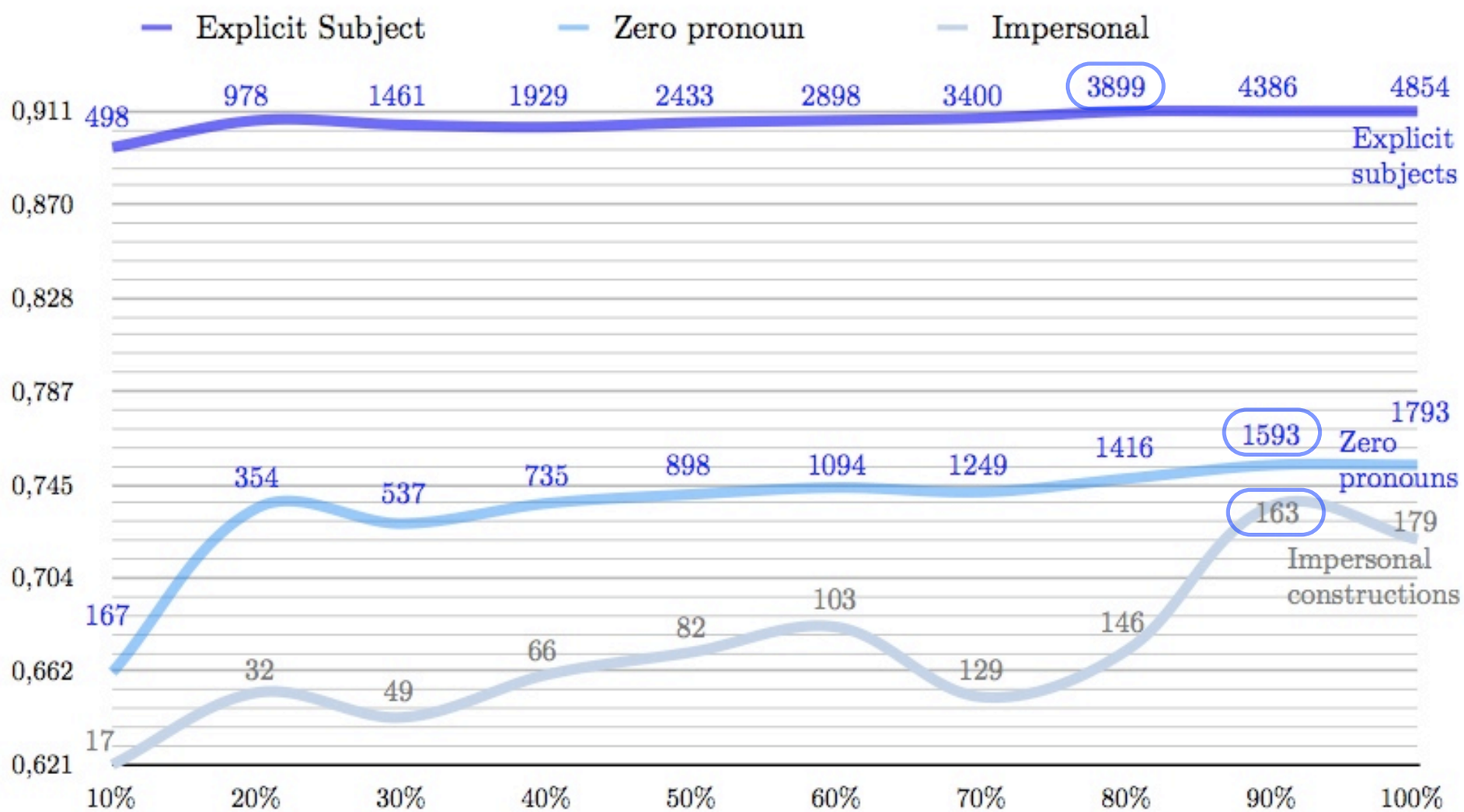


Learning curve for accuracy of the classes

Elliphant

Learning Analysis

— How **many instances** are needed to reach the plateau?



Learning curve for accuracy, recall and f-measure in relation to the number of instances of each class

Elliphant

Feature Selection

— Most effective **groups of features**?

Attribute Selection	Features selected	Accuracy
CfsSubsetEval	PARSER, NUMBER, NHPREV, NHTOT, VERB_{type}, PERSON	0.824
ChiSquaredAttributeEval	LEMMA, POS_{pos}, NHTOT, NHPREV, POS_{pre}, PARSER	0.848
ConsistencySubsetEval	PARSER, LEMMA, NUMBER, AGREE, NHTOT, POS_{pos}, POS_{pre}	0.843
FilteredAttributeEval	POS_{pos}, LEMMA, NHPREV, NHTOT, PARSER, POS_{pre}	0.848
FilteredSubsetEval	PARSER, NHPREV, NHTOT	0.819
GainRatioAttributeEval	NHPREV, PARSER, PERSON, NHTOT, POS_{pos}, CLAUSE	0.833
InfoGainAttributeEval	POS_{pos}, LEMMA, NHPREV, NHTOT, PARSER, POS_{pre}	0.848
OneRAttributeEval	NHTOT, POS_{pos}, CLAUSE, PERSON, NHPREV, PARSER	0.833
ReliefFAttributeEval	POS_{pos}, VERB_{type}, LEMMA, PARSER, CLAUSE, POS_{pre}	0.825
SymmetricalUncertAttributeEval	NHPREV, PARSER, NHTOT, POS_{pos}, PERSON, LEMMA	0.851

Accuracy of the classification using selected groups of features

Elliphant

Feature Selection

— How much **quantity** and **type of knowledge** needed?

Extrinsic *vs.* **Intrinsic** features

PARSER
LEMMA
PERSON
POS_{pos}
POS_{pre}

CLAUSE,
AGREE
NHPREV
NHTOT
$VERB_{type}$

Classes	Extrinsic features			Intrinsic features		
Features	P	R	F	P	R	F
Explicit subjects	0.654	0.664	0.659	0.866	0.312	0.459
Zero pronouns	0.865	0.891	0.878	0.779	0.983	0.869
Impersonal constructions	0	0	0	0.944	0.285	0.438

Extrinsic and intrinsic features classification results

Elliphant

Feature Selection

— How **informative** are the **individual features**?

Leave one feature out

Feature omitted	Accuracy	Feature omitted	Accuracy
PARSER	0.854	VERB_{type}	0.863
NHTOT	0.860	NUMBER	0.864
LEMMA	0.861	INF	0.864
POS_{pos}	0.861	AGREE	0.865
NHPREV	0.862	POS_{pre}	0.866
PERSON	0.863	SE	0.866
CLAUSE	0.863	A	0.867

Leave-one-out classification results

Elliphant

Genre Analysis

— What is the impact of different **genres**?

Class	Precision	Recall	F-measure
Legal genre Explicit subjects	0.920	0.955	0.937
Health genre Explicit subjects	0.881	0.888	0.884
Legal genre Zero pronouns	0.761	0.649	0.701
Health genre Zero pronouns	0.784	0.796	0.790
Legal genre Impersonal constructions	0.786	0.620	0.693
Health genre Impersonal constructions	0.905	0.620	0.736

Comparative evaluation of legal and health genres.

Elliphant

Genre Analysis

— What is the impact of different **genres** on classifier **training/testing**?

Training set \ Testing set	Legal	Health	ESZIC Corpus
Legal	0.895	0.859	0.885
Health	0.858	0.841	**0.887**
ESZIC Corpus (all)	**0.920**	**0.933**	0.869

Accuracy of cross-genre training and testing evaluation

Elliphant

Comparative Evaluation

leave-one-out cross-validation

ESZIC **training data**	**Precision**	**Recall**	**F-measure**
Elliphant Explicit subjects	0.901	0.924	0.913
Elliphant Zero pronouns	0.774	0.743	0.758
Elliphant Impersonal constructions	0.889	0.626	0.734

Elliphant ESZIC training data accuracy: 0.869

ESZIC **training data**	**Precision**	**Recall**	**F-measure**
Machinese Explicit subjects	0.911	0.716	0.802
Machinese Zero pronouns + Impersonal constructions	0.543	0.829	0.656

Machinese ESZIC training data accuracy: 0.749

Elliphant

Comparative Evaluation → **Issues**

— corpus

— approach

— concepts

Explicit subject accuracy: 0.912
Zero pronouns accuracy: 0.76
Impersonal constructions accuracy: 0.73
(Rello, Suárez & Mitkov, 2010)

Explicit subject accuracy: 0.80
Zero pronouns accuracy: 0.98
(Ferrández & Peral, 2000)

Zero pronouns accuracy: 0.57
(Rello & Illisei, 2009b)

Summary

1 — Performance tops out at when using **90%** of the available training data

2 — The **most effective group** comprises **just six** of the features: only **one feature** does **not** make any **meaningful contribution**

3 — **Genre variation** was observed

Summary

4 — Elliphant: **first machine learning** approach to the **identification** of zero pronouns, impersonal constructions, and explicit subjects

5 — **Outperforms** the **parser** and competitive with the **previous approaches**

6 — The **first attempt** to identify **impersonal sentences** in Spanish

On-going work

1 — **More efficient features selection** and **parameter optimization**

Economic features: SE, INF, A, $\text{VERB}_{\text{type}}$, LEMMA, NUMBER, PERSON, POS_{pos}, POS_{pre} Accuracy 0.788 (KStar, blend 20)

2 — **13 classes** classification

Explicit Subject, active clauses: Accuracy 0.88 (KStar, blend 40)

Zero pronoun, passive clauses: Accuracy 0.001 (KStar, blend 40)

3 — Unsupervised machine learning: **Clustering**

Algorithm k-means, 3 classes: 70 %

4 — **Elliphant_PT for Portuguese**

Muchísimas gracias :-)

¿Alguna pregunta? *¿Demo?*