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
Elliphant

Outline

— **What**
  - objective

— **Why**
  - motivation

— **How**
  - methodology
  - analysis

— **Others?**
  - literature review
  - learning analysis
  - genre analysis
  - feature analysis
  - comparative evaluation

Luz Rello
Elliphant: A Machine Learning Method for Identifying Subject Ellipsis and Impersonal Constructions
Elliphant: A Machine Learning Method for Identifying Subject Ellipsis and Impersonal Constructions

Objective

What Elliphant

Elliphant:

fully automatic system

identify

machine learning

explicit subjects

zero pronouns

impersonal constructions

all possible constituents in subject position

complementary distribution
**1 - Frequent**

*Spanish pro-drop* language (Chomsky, 1981)

2. Chinese (Zhao & Ng, 2007)
4. Russian (Kibrik, 2004)
5. Spanish (Ferrández & Peral, 2000)

**zero pronouns**

1. English (Evans, 2001)
2. French (Danlos, 2005)
   *no Spanish*

**impersonal constructions**
Anaphora resolution  (Mitkov, 2010)

1º — Identification zero pronouns

2º — Filter non-referential impersonal constructions

2- Crucial

Motivation

1 - Frequent
2 - Crucial
3 - Useful
4 - Currently needed

Why
Elliphant
Elliphant

Motivation

- Why
  - Frequent
  - Crucial
  - Useful
  - Currently needed

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)
“[...] 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).
Luz Rello
Elliphant: A Machine Learning Method for Identifying Subject Ellipsis and Impersonal Constructions

How Elliphant

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

Elliphant
Elliphant

1 - Classification
2 - Training data
3 - Analysis
4 - Evaluation

Linguistic Criteria
(Rello, 2010)

Custom-built Software
Annotation

1 — Classification
2 — Training data
3 — Analysis
4 — Comparative evaluation

(Rello, Baeza-Yates & Mitkov, 2010)

(Rello, Suárez & Mitkov, 2010)
Linguistically motivated classes

1 — Explicit subject $\rightarrow [-\text{elliptic, } +\text{ referential}]

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

2 — Zero pronoun $\rightarrow [+\text{elliptic, } +\text{ referential}]

Las leyes no tendrán efecto retroactivo si $\emptyset$ no dispusieren lo contrario.
The law will not have a retroactive effect unless (\textit{they}) specify otherwise.

3 — Impersonal construction $\rightarrow [+\text{elliptic, } -\text{ referential}]

Cuando hay un diagnóstico.
When (\textit{there}) is a diagnosis.

(Rello, 2010)
**Linguistically motivated classes**

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

<table>
<thead>
<tr>
<th>CLASS</th>
<th>TYPE OF INFORMATION:</th>
<th>SYNTACTIC CATEGORY</th>
<th>VERBAL DIATHESIS</th>
<th>PHONETIC REALIZATION</th>
<th>SEMANTIC INTERPRETATION</th>
<th>DISCOURSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LINGUISTIC FEATURES:</td>
<td>[±] Nominal subject</td>
<td>[±] Active</td>
<td>[±] Elliptic noun phrase</td>
<td>[±] Elliptic noun phrase head</td>
<td>[±] Active participant</td>
</tr>
<tr>
<td>CLASS 1: <strong>EXPLICIT SUBJECT</strong></td>
<td>Explicit subject</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Reflex passive subject</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Passive subject</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CLASS 2: <strong>ZERO PRONOUN</strong></td>
<td>Omitted subject</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Omitted subject head</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Reflex passive omitted subject</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Reflex passive omitted subject head</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Passive omitted subject</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Non nominal subject</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Reflex passive non nominal subject</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Passive non nominal subject</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CLASS 3: <strong>NON-REFERENTIAL IMPERSONAL SUBJECT</strong></td>
<td>Impersonal with se</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Impersonal</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
Corpus compilation

**ESZIC corpus:** Explicit Subjects, Zero-pronouns and Impersonal Constructions

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

- 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.
Elliphant

ESZIC Corpus parsing

Connexor’s Functional Dependency Grammar Parser

(Tapanainen and Järvinen, 1997)

(Connexor Oy, 2006a, b)
ESZIC Corpus annotation

Annotation software

finite verb
with the clause

purpose built
clause splitter module

(Rello & Ilisei, 2009b)
ESZIC Corpus

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

<table>
<thead>
<tr>
<th>Number of instances per class</th>
<th>Legal ESZIC Corpus</th>
<th>Health ESZIC Corpus</th>
<th>ESZIC Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicit subjects</td>
<td>2,739</td>
<td>2,116</td>
<td>4,855</td>
</tr>
<tr>
<td>Zero pronouns</td>
<td>619</td>
<td>1,174</td>
<td>1,793</td>
</tr>
<tr>
<td>Impersonal constructions</td>
<td>71</td>
<td>108</td>
<td>179</td>
</tr>
<tr>
<td>Total</td>
<td>3,429</td>
<td>3,398</td>
<td>6,827</td>
</tr>
</tbody>
</table>
Elliphant: A Machine Learning Method for Identifying Subject Ellipsis and Impersonal Constructions

- **How Elliphant**
  - **Purpose built tools**
    - parser *extrinsic* vs. parser *intrinsic*
      - clause splitter/classifier module
      - agreement module
      - noun phrases/phrases module
  - verb lists (11,000)
    - (Real Academia Española, 2001)

### Features

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

How

Elliphant

Training data

- training vector for each instance (6,827)
- standard csv format

Weka 3-6-3

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

1 - Classification
2 - Training data
3 - Analysis
4 - Evaluation

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

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%

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicit subjects</td>
<td>0.900</td>
<td>0.923</td>
<td>0.911</td>
</tr>
<tr>
<td>Zero pronouns</td>
<td>0.772</td>
<td>0.740</td>
<td>0.756</td>
</tr>
<tr>
<td>Impersonal constructions</td>
<td>0.889</td>
<td>0.626</td>
<td>0.734</td>
</tr>
</tbody>
</table>

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

— Most effective groups of features?

<table>
<thead>
<tr>
<th>Attribute Selection</th>
<th>Features selected</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CfsSubsetEval</td>
<td>PARSER, NUMBER, NHPREV, NHTOT, VERB\textsubscript{type}, PERSON</td>
<td>0.824</td>
</tr>
<tr>
<td>ChiSquaredAttributeEval</td>
<td>LEMMA, POS\textsubscript{pos}, NHTOT, NHPREV, POS\textsubscript{pre}, PARSER</td>
<td>0.848</td>
</tr>
<tr>
<td>ConsistencySubsetEval</td>
<td>PARSER, LEMMA, NUMBER, AGREE, NHTOT, POS\textsubscript{pos}, POS\textsubscript{pre}</td>
<td>0.843</td>
</tr>
<tr>
<td>FilteredAttributeEval</td>
<td>POS\textsubscript{pos}, LEMMA, NHPREV, NHTOT, PARSER, POS\textsubscript{pre}</td>
<td>0.848</td>
</tr>
<tr>
<td>FilteredSubsetEval</td>
<td>PARSER, NHPREV, NHTOT</td>
<td>0.819</td>
</tr>
<tr>
<td>GainRatioAttributeEval</td>
<td>NHPREV, PARSER, PERSON, NHTOT, POS\textsubscript{pos}, CLAUSE</td>
<td>0.833</td>
</tr>
<tr>
<td>InfoGainAttributeEval</td>
<td>POS\textsubscript{pos}, LEMMA, NHPREV, NHTOT, PARSER, POS\textsubscript{pre}</td>
<td>0.848</td>
</tr>
<tr>
<td>OneRAttributeEval</td>
<td>NHTOT, POS\textsubscript{pos}, CLAUSE, PERSON, NHPREV, PARSER</td>
<td>0.833</td>
</tr>
<tr>
<td>ReliefFAttributeEval</td>
<td>POS\textsubscript{pos}, VERB\textsubscript{type}, LEMMA, PARSER, CLAUSE, POS\textsubscript{pre}</td>
<td>0.825</td>
</tr>
<tr>
<td>SymmetricalUncertAttributeEval</td>
<td>NHPREV, PARSER, NHTOT, POS\textsubscript{pos}, PERSON, LEMMA</td>
<td>0.851</td>
</tr>
</tbody>
</table>

Accuracy of the classification using selected groups of features
Feature Selection

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

<table>
<thead>
<tr>
<th>Classes</th>
<th>Extrinsic features</th>
<th>Intrinsic features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Explicit subjects</td>
<td>0.654</td>
<td>0.664</td>
</tr>
<tr>
<td>Zero pronouns</td>
<td>0.865</td>
<td>0.891</td>
</tr>
<tr>
<td>Impersonal constructions</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Extrinsic and intrinsic features classification results
Feature Selection

— How informative are the individual features?

Leave one feature out

<table>
<thead>
<tr>
<th>Feature omitted</th>
<th>Accuracy</th>
<th>Feature omitted</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARSER</td>
<td>0.854</td>
<td>VERB_{type}</td>
<td>0.863</td>
</tr>
<tr>
<td>NHTOT</td>
<td>0.860</td>
<td>NUMBER</td>
<td>0.864</td>
</tr>
<tr>
<td>LEMMA</td>
<td>0.861</td>
<td>INF</td>
<td>0.864</td>
</tr>
<tr>
<td>POS_{pos}</td>
<td>0.861</td>
<td>AGREE</td>
<td>0.865</td>
</tr>
<tr>
<td>NHPREV</td>
<td>0.862</td>
<td>POS_{pre}</td>
<td>0.866</td>
</tr>
<tr>
<td>PERSON</td>
<td>0.863</td>
<td>SE</td>
<td>0.866</td>
</tr>
<tr>
<td>CLAUSE</td>
<td>0.863</td>
<td>A</td>
<td>0.867</td>
</tr>
</tbody>
</table>

Leave-one-out classification results
Genre Analysis

— What is the impact of different genres?

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legal genre Explicit subjects</td>
<td>0.920</td>
<td>0.955</td>
<td>0.937</td>
</tr>
<tr>
<td>Health genre Explicit subjects</td>
<td>0.881</td>
<td>0.888</td>
<td>0.884</td>
</tr>
<tr>
<td>Legal genre Zero pronouns</td>
<td>0.761</td>
<td>0.649</td>
<td>0.701</td>
</tr>
<tr>
<td>Health genre Zero pronouns</td>
<td>0.784</td>
<td>0.796</td>
<td>0.790</td>
</tr>
<tr>
<td>Legal genre Impersonal constructions</td>
<td>0.786</td>
<td>0.620</td>
<td>0.693</td>
</tr>
<tr>
<td>Health genre Impersonal constructions</td>
<td>0.905</td>
<td>0.620</td>
<td>0.736</td>
</tr>
</tbody>
</table>

Comparative evaluation of legal and health genres.
Genre Analysis

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

<table>
<thead>
<tr>
<th>Training set</th>
<th>Legal</th>
<th>Health</th>
<th>ESZIC Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legal</td>
<td>0.895</td>
<td>0.859</td>
<td>0.885</td>
</tr>
<tr>
<td>Health</td>
<td>0.858</td>
<td>0.841</td>
<td>0.887</td>
</tr>
<tr>
<td>ESZIC Corpus (all)</td>
<td><strong>0.920</strong></td>
<td><strong>0.933</strong></td>
<td>0.869</td>
</tr>
</tbody>
</table>

Accuracy of cross-genre training and testing evaluation
## Comparative Evaluation

leave-one-out cross-validation

### The table below shows the accuracy results for Elliphant and Machine.

<table>
<thead>
<tr>
<th>ESZIC training data</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elliphant Explicit subjects</td>
<td>0.901</td>
<td>0.924</td>
<td>0.913</td>
</tr>
<tr>
<td>Elliphant Zero pronouns</td>
<td>0.774</td>
<td>0.743</td>
<td>0.758</td>
</tr>
<tr>
<td>Elliphant Impersonal constructions</td>
<td>0.889</td>
<td>0.626</td>
<td>0.734</td>
</tr>
</tbody>
</table>

**Elliphant ESZIC training data accuracy:** 0.869

<table>
<thead>
<tr>
<th>ESZIC training data</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Explicit subjects</td>
<td>0.911</td>
<td>0.716</td>
<td>0.802</td>
</tr>
<tr>
<td>Machine Zero pronouns</td>
<td>0.543</td>
<td>0.829</td>
<td>0.656</td>
</tr>
<tr>
<td>+ Impersonal constructions</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Machine ESZIC training data accuracy:** 0.749

---

Luz Rello
Elliphant: A Machine Learning Method for Identifying Subject Ellipsis and Impersonal Constructions
Comparative Evaluation

**Elliphant**

1. Classification
2. Training data
3. Analysis
4. Evaluation

**Issues**

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Approach</th>
<th>Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

**Explicit subject accuracy**: 0.80

**Zero pronouns accuracy**: 0.98
*(Ferrández & Peral, 2000)*

**Impersonal constructions accuracy**: 0.73
*(Rello, Suárez & Mitkov, 2010)*

**Explicit subject accuracy**: 0.912

**Zero pronouns accuracy**: 0.76

*(Rello, Suárez & Mitkov, 2010)*
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, VERB<sub>type</sub>, LEMMA, NUMBER, PERSON, POS<sub>pos</sub>, POS<sub>pre</sub>

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?