Introduction

- Romance Subject Inversion (RSI)
  - Example: "Salió la ranita pequeña a través de la ventana."
  - Properties associated to RSI this are either semantic, syntactic, or pragmatic.

Goal and Motivation

- An exhaustive quantitative analysis of naturally occurring data is missing in the literature.
- We want to fill this empirical gap, namely:
  - An exhaustive quantitative analysis of naturally occurring data is missing in the literature.
  - Pragmatic properties:
    - Focused subject / discourse new subject
  - Properties of the subject:
    - non agent subject
    - indefinite subject
    - quantified subject
    - sentential subject
    - subject todo
  - Properties of the clause:
    - relative clause
    - direct interrogative clause
    - indirect interrogative clause
    - exclamative clause
  - Pragmatic properties:
    - discourse new subject
    - discourse given predicate
    - Relation with each story and each speaker.

Method

- chi-square test for the correlation of each feature to RSI. The test is carried out for features which have been claimed to trigger SI in the literature we revisited.
- decision trees (as an additional tool for manual error analysis)

Results

<table>
<thead>
<tr>
<th>Type of clause</th>
<th>Number of tokens</th>
<th>RSI</th>
<th>Non-RSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>relative clause</td>
<td>1234</td>
<td>0.78</td>
<td>0.22</td>
</tr>
<tr>
<td>direct clause</td>
<td>5678</td>
<td>0.89</td>
<td>0.11</td>
</tr>
<tr>
<td>exclamative clause</td>
<td>9012</td>
<td>0.65</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Decision tree for the Spanish data

Although the data we have are too sparse for pure classification purposes, we used C4.5 (348) as a tool to find cases which are hard to classify and hence give us good material for error analysis.

- C4.5 decision tree classifiers (in the 348 implementation of Weka, [Witten and Frank 2005]).
  - Tree-based classification procedure: training phase
  - Rule-based classification procedure: search for rules that describe the data
- Error analysis: 36.5% of the false negative cases (wrongly classified as -RSI) would also be acceptable with a postverbal subject and 64% of the false positive cases with a preverbal subject.
  - This explains the low recall for RSI. In many cases SI is simply not obligatory.
  - Interestingly, inverted subjects are more predictable then preverbal ones when the cues for one particular construction are fewer. In other words, inversion appears to be the default case, while preverbal subjects are required under more specific circumstances.

Discussion

- Lexico-syntactic factors, related to argument structure show the strongest correlation to RSI: subjects lacking volition/control on the event favour inversion. Also verbs of appearance, occurrence, and decussative-reflexive verbs, which are all select a non-volitional volitional subject / subject favour inversion.
- Also some syntactic features show a strong correlation: SI in Spanish and Catalan is highly favoured within a relative clause. The same does not hold for Italian, where this correlation is not significant.
- Inversion is more frequent in narrations with frequent topic shifts. This suggests that the organization of discourse influences the subject position.
- RSI varies very much among speakers: 10% to 37% in Spanish, 7% to 37% in Catalan and 7% to 24% in Italian.
  - Stylistic choices are crucial for RSI selection.
  - The upper bound for the performance of any automatic binary classifier is necessarily low.

Conclusion and further work

- Romance Languages behave similarly with respect to SI: Only few features show a different behaviour (relatives clauses and decussative-reflexive verbs in Italian, intransitive and reflexive verbs in Catalan and Copula verbs in Spanish).
- Some differences receive a theoretical explanation (decussative in Italian), others may reveal flaws in the statistical methodology.
- Questions for future work:
  - Are relative clauses in Italian syntactically different from Spanish and Catalan ones or is it the rules of inversion that vary?
  - Why do copula verbs in Spanish favor SI more than in the other two languages?
  - Decussative reflexive verbs (e.g. rompere to break, ITI) are much more limited in number in Italian than in the other two languages (in particular, Italian does not have causer to fall, SP).
  - Why should this affect significance with SI in Italian?

References


Goal and Motivation

- An exhaustive quantitative analysis of naturally occurring data is missing in the literature.
- We want to fill this empirical gap, namely:
  - An exhaustive quantitative analysis of naturally occurring data is missing in the literature.
  - Pragmatic properties:
    - Focused subject / discourse new subject
  - Properties of the subject:
    - non agent subject
    - indefinite subject
    - quantified subject
    - sentential subject
    - subject todo
  - Properties of the clause:
    - relative clause
    - direct interrogative clause
    - indirect interrogative clause
    - exclamative clause
  - Pragmatic properties:
    - discourse new subject
    - discourse given predicate
    - Relation with each story and each speaker.

Method

- chi-square test for the correlation of each feature to RSI. The test is carried out for features which have been claimed to trigger SI in the literature we revisited.
- decision trees (as an additional tool for manual error analysis)