Subject Inversion in Romance: A Corpus-Based Study

Lisa Brunetti[a], and Stefan Bott[b],
[a]Laboratoire Dynamique du Langage / Université Lumière Lyon2;
[b]Universitat Politècnica de Catalunya / Universitat Pompeu Fabra, Barcelona

Introduction

- Romance Subject Inversion (RSI)

(1) a. Salió la ranita pequeña a través de la ventana
came-out the frog small across of the window
'And there came in the small frog through the window' (Spanish)

b. Fins i tot els hi cau el cafè
even to-them there falls the coffee
'He even drops the coffee' (Catalan)

c. A un bambino un giorno arriva un regalo
to a boy one day arrives a present
'One day a boy receives a present' (Italian)

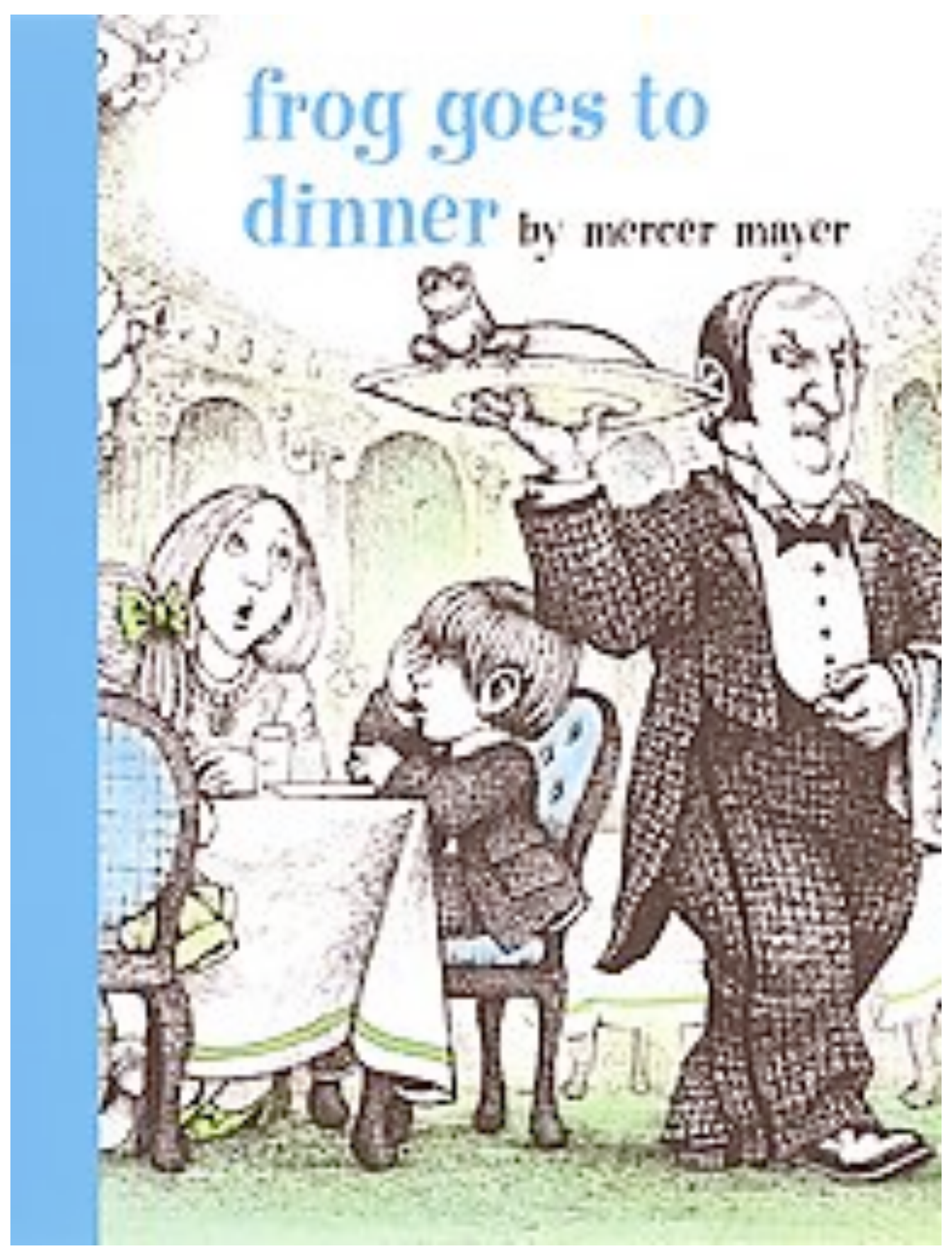
- Properties associated to RSI this are either semantic, syntactic, or pragmatic.
  - Syntactic properties:
    - ∗ Unaccusative verbs → subject in object position ([Burzio 1986])
    - ∗ Type of clause (relative, interrogative, exclamative) ([Torrego 1984].)
  - Semantic properties:
    - ∗ Verbs of appearance, existence, presentation, subject with unidentifiable reference ([Hatcher 1956], [Lambrecht 1994])
    - ∗ Unaccusative verbs → Non agentive subject ([Lambrecht 1995], [Lambrecht 2000], [Kennedy 1999])
  - Pragmatic properties:
    - ∗ Focused subject / discourse new subject ([Contreras 1976], [Zubizarreta 1998], [Zubizarreta 1999], [Burzio 1986])
    - ∗ Given predicate ([Marandin 2003])

Goal and Motivation

- An exhaustive quantitative analysis of naturally occurring data is missing in the literature.
- We want to fill this empirical gap, namely:
  - quantitatively determine the burden of different factors in predicting RSI
  - understand how far RSI can be attributed to purely **syntactic/lexico-semantic** rather than **pragmatic** features.

Data

- Corpus.
  - Multilingual oral corpus Nocando http://nocando.barcelonamedia.org/ [Brunetti et al 2010] , transcribed from the recordings of free narrations of three children picture books (Frog goes to dinner, One frog too many, A frog on his own, Mayer 1969)
  - About 90000 words of speech.
  - Inverted subjects over the total of overt subejcts:
    - ∗ 259 over a total 1251 subjects for Spanish
    - ∗ 200 over 1437 for Italian,
    - ∗ 345 over 1034 for Catalan.
- Annotation Features
  - Properties of the verb:
    - ∗ verb of appearance
    - ∗ verb of directed movement
    - ∗ verb of occurrence
    - ∗ verb of stance
    - ∗ verb of commencement
    - ∗ transitive verb
    - ∗ intransitive verb
    - ∗ copula verb
    - ∗ reflexive verb decausative
    - ∗ reflexive verb psychological
    - ∗ reflexive verb autocausative
    - ∗ obj experiencer psych verb
  - 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.
- decistion trees (as an additional tool for manual error analysis)

Results

	Spanish		Catalan		Italian	
feature	χ^2	p-value	χ^2	p-value	χ^2	p-value
Non-agentive subject	129,79	<0,001	95,61	<0,001	34,14	<0,001
Unaccusative Verb	97,70	<0,001	63,34	<0,001	89,69	<0,001
rel	71,01	<0,001	117,50	<0,001	0,80	non-sig
Verb of directed movement	67,18	<0,001	6,72	<0,01	22,28	<0,001
Verb of appearance	46,84	<0,001	100,24	<0,001	96,86	<0,001
Indefinite Subject	36,57	<0,001	45,48	<0,001	48,91	<0,001
Verb of occurence	36,16	<0,001	24,64	<0,001	31,92	<0,001
Discourse new subject	34,16	<0,001	-	-	63,59	<0,001
Vtrans	32,52	<0,001	15,93	<0,001	19,02	<0,001
Subject meaning ”all” (todo)	30,90	<0,001	41,86	<0,001	38,26	<0,001
Discourse given predicate	25,53	<0,001	-	-	8,02	<0,005
Intransitive Verb	23,06	<0,001	2,22	non-sig	10,62	<0,005
sentential subject	22,92	<0,001	81,17	<0,001	19,54	<0,001
Decausative reflexive Verb	22,44	<0,001	17,84	<0,001	2,90	non-sig
Verb of existence or presence	14,29	<0,001	0,73	non-sig	1,12	non-sig
Not classified reflexive verb	10,79	<0,005	10,76	<0,005	3,52	non-sig
Copula verb	9,33	<0,005	0,60	non-sig	2,02	non-sig
Vest	5,95	<0,05	0,02	non-sig	0,69	non-sig
Psychological reflexive verb	4,80	<0,05	12,14	<0,001	4,71	<0,05
Reflexive verb	1,62	non-sig	10,92	<0,001	3,61	non-sig
Psychological verb	1,57	non-sig	5,01	<0,05	3,41	non-sig
Lexicalized reflexive verbs	0,29	non-sig	1,26	non-sig	3,75	non-sig

Differences between languages are marked in red. Features with a frequency below 10 are marked in gray.

Decision tree for the Spanish data

Although the data we have are too sparse for pure classification purposes, we used C4.5 (J48) 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 J48 implementation of Weka, [Witten and Frank 2005]).
  - Tenfold cross-classification to remedy the sparseness of data.
  - Overall accuracy and precision of predicting +RSI : 83,8% and 73,6%.
  - However, the recall of +RSI prediction is poor (34,9%).
- 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.
- An observation of the contexts of false positives further reveals that many misclassifications co-occur with discourse phenomena, like topic shift or contrast. This finding confirms us how discourse plays a crucial role in inversion, and that future research will have to focus on the addition of more, and more sophisticated, pragmatic features.

Discussion

- Lexico-semantics 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 decausative-reflexive verbs, which are all all select a non-volitional volitional subject favour inversion.
- Also some syntactic features show a stong 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 (relative clauses and decausative reflexive verbs in Italian, intransitive and reflexive verbs in Catalan and Copula verbs in Spanish). Some differences receive a theoretical explanation (decausatives 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?
  - Decausative reflexive verbs (e.g. *rompersi* to break, IT), are much more limited in number in Italian than in the other two languages (in particular, Italian does not have **caerse** to fall, SP). Why should this affect significance with SI in Italian?

References

[Brunetti et al 2010] Brunetti, Lisa, Bott, Stefan, Costa, Joan, and Vallduví, Enric 2010. A multilingual annotated corpus for the study of Information Structure. In Proceedings of the Third international Conference/ Dritte Internationale Konferenz, Mannheim, 22-24.09.2009, Konopka, Marek, Kubczak, Jacqueline, Mair, Christian, Šticha, Františekand Waßner, Ulrich H. (eds), Tübingen: Gunter Narr Verlag.

[Burzio 1986] Burzio, L., 1986. Italian syntax: A government-binding approach. Dordrecht: Reidel.

[Contreras 1976] Contreras, Heles 1976. A theory of word order with special reference to Spanish. Amsterdam: North Holland.

[Creissels 2006] Creissels, Denis 2006. Syntaxe générale, une introduction typologique. Paris : Hermès.

[Hatcher 1956] Hatcher, Anne G. 1956. Theme and underlying question. Two studies of Spanish word order. Word (12): 14-31.

[Kennedy 1999] Kennedy, Becky 1999. Focus constituency. Journal of Pragmatics (31): 1203-1230.

[Witten and Frank 2005] Witten, Ian H. and Frank, Eibe. 2005. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). San Francisco: Morgan Kaufmann

[Lambrecht 1994] Lambrecht, Knud 1994. Information structure and sentence form. A theory of topic, focus, and the mental representations of discourse referents, [Cambridge Studies in Linguistics (71)], Cambridge: Cambridge University Press.

[Lambrecht 1995] Lambrecht, Knud 1995. The pragmatics of case: On the relationship between semantic, grammatical, and pragmatic roles in English and French. In Essays in Semantics and Pragmatics. In Honor of Charles J. Fillmore, Shibatani, Masayoshi and Sandra A. Thompson (eds.), 145-190. Amsterdam: Benjamins.

[Lambrecht 2000] Lambrecht, Knud 2000. When subjects behave like objects: An analysis of the merging of S and O in sentence-focus constructions across languages. Studies in Language (24): 611-682.

[Marandin 2003] Marandin Jean-Marie 2003. Inversion du sujet et structure de l'information dans les langues romanes. In Langues romanes. Problèmes de la phrase simple, Godard, Danièle (ed). Paris: Editions du CNRS.

[Torrego 1984] Torrego, Esther 1984. On inversion in Spanish and some of its effects. Linguistic Inquiry (15):102-129.

[Zubizarreta 1998] Zubizarreta, María Luísa 1998. Prosody, focus and word order. Cambridge, Mass.: MIT Press.

[Zubizarreta 1999] Zubizarreta 1999 María Luísa 1999. Las funciones informativas: tema y foco. In: Gramática descriptiva de la lengua española vol. 3, Ignacio Bosque and Violeta Demonte (eds). Madrid: Espasa Calpe, 4215-4244.