

Lisa Brunetti, Stefan Bott
Laboratoire Dynamique du Langage / Université Lumière Lyon2, Departament de
Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya
lisa.brunetti@univ-lyon2.fr, sbott@lsi.upc.edu

Subject inversion in Romance: a corpus-based study

Introduction

We present a corpus based study that investigates which factors trigger a postverbal position of the subject in Spanish utterances, as in (1). This study is the first part of a larger study which includes other Romance languages (Catalan and Italian).

Spanish

- (1) *y voilà, salió la ranita pequeña a través de la ventana*
and there-he-is came-out the frog small across of the window
'And there came the small frog through the window'

The literature on Romance Subject Inversion (RSI) associates this phenomenon either with semantic, syntactic, or pragmatic properties. According to several authors, RSI depends on the semantic and thematic properties of the verb. In particular, RSI has been associated with intransitive presentational verbs or verbs of existence or appearance (cf. Hatcher 1956, Lambrecht 1994). This fact would explain the preference for inverted subjects with unidentifiable reference. Other scholars have emphasized the relation between inversion and unaccusative verbs. While most part of the literature has focused on the syntactic aspects of such a relation, a few works have focused on the relation between inversion and the fact that these verbs select a non-agent subject (cf. Lambrecht 1995, 2000, Kennedy 1999). From a pragmatic point of view, an inverted subject is considered as either focused or part of the sentence focus (for Spanish, see Contreras 1976, Zubizarreta 1998, 1999). A different proposal comes from Marandin 2003, who argues that the crucial pragmatic factor triggering RSI is that the predicate is given. Finally, from the point of view of clause type, inversion in Spanish obligatorily occurs in interrogative clauses, and frequently – but not obligatorily – in relative clauses (Torrego 1984).

Although RSI has been the object of much interest in the literature, research based on an exhaustive quantitative analysis of naturally occurring data is, to the best of our knowledge, still scarce. On the other hand, the complexity of this phenomenon and the fact that it is often claimed to be strongly related to discourse factors demand an account that is not solely based on constructed sentences and introspective judgments. The present work wants to be a first step towards the filling of this empirical gap. Our aim is to quantitatively determine the burden of different factors in predicting RSI and more precisely, to understand how far RSI can be attributed to purely syntactic/lexico-semantic rather than pragmatic features.

Data

A Spanish oral corpus of 25000 words was used, transcribed from the recordings of free narrations of textless stories for children (Meyer 1969). This corpus is part of a larger multilingual corpus which also includes Catalan, Italian, German, and English narrations with the same characteristics. A total of 1221 pre-and postverbal subjects

were found. These occurrences were annotated for 34 different features. There were 13 features we could find less than 10 times in the corpus, so we excluded them and we used the remaining 21 features for the study.

As for the lexical-semantic properties of the subject, we annotated indefinite, generic, and quantified subjects. We also annotated whether subjects lacked agentive properties. From a syntactic point of view, we signaled whether the subject was sentential or modified by a restrictive relative clause. Verbs were divided by 11 classes according to their argument properties: transitive, intransitive, unaccusative, (object experiencer) psychological verbs, and different types of verbs with reflexive morphology, divided into: pure reflexives, lexicalized reflexives, reciprocals, de-causatives, auto-causatives, antipassives, psychological reflexives (Creissels 2006). As for the class of unaccusatives, its existence is not uncontroversial in the literature. For this reason, we also classified verbs according to those semantic properties that are generally, but not always, associated with the unaccusative class, namely: verbs of movement toward a point, of existence, absence, commencement, continuation, appearance, occurrence, and stance (Hatcher 1956). A third set of features concerns clause types: relative, temporal, concessive, final, conditional, causal, and declarative subordinate clauses, *wh*-and *y/n*-direct and indirect questions, exclamatives. We finally added two pragmatic features: *givP*, for a predicates that are discourse given, and *sbj_new*, for subjects that are discourse new.

Analysis

A standard chi-square test was applied for the correlation of each feature with the occurrence of RSI. The results show that the correlation is significant for 18 features (see Table 1). The three top-scoring features are related to argument structure, hence to lexico-semantic factors. The crucial factor seems to be the subject lacking volition/control on the event (see also the high score of appearance, occurrence, and decausative-reflexive verbs, which all select a non-volitional subject). The fourth top-scoring feature is syntactic: SI in Spanish is highly predictable with a relative clause. A comparison with Catalan and Italian data shows that the same holds for Catalan but not for Italian, where this correlation is not significant.

The two pragmatic features – predicate givenness and subject newness in discourse – show both a significant correlation with inversion, but the latter appears to be more significant than the former. If we look at these results in the Italian corpus, we see an even greater distance in terms of predictability between the two features: although both are significantly correlated with inversion, predicate-givenness has a higher p-value than subject-newness. From these results we cannot conclude with Marandín 2003 that the pragmatic trigger of SI is predicate-givenness. The traditional assumption that inversion is related to the information status of the subject seems to be confirmed by our data.

The percentage of inversion per speaker was also calculated. The variation widely ranges from 3% to over 37%, by which we conclude that stylistic choices are crucial for SI selection. A significant correlation is also observed with the story *Frog goes to dinner*, whose plot favors a narration with frequent topic shifts. This fact supports the idea that the organization of discourse strongly influences the subject position. However, we must say that no such correlation was found in Catalan and Italian data.

In addition to pure descriptive statistics we carried out a small experiment with C4.5 decision tree classifiers (in the J48 implementation of Weka, Witten & Frank 2005). We used a tenfold cross-classification to remedy the sparseness of data. Although the amount of data was not sufficient to train a strongly reliable classifier, its overall accuracy and the precision of predicting +RSI are fairly acceptable (83,8% and 73,6%). However, the recall of +RSI prediction is poor (34,9%).

The decision tree algorithm allowed us for an analysis of the misclassified examples. Error analysis showed that 36,5% of the false negative cases (wrongly classified as -RSI) and 64% of the false positive cases would also be acceptable with a preverbal subject, which explains the low recall for +RSI: in many cases SI is simply not obligatory. Interestingly, inverted subjects are more predictable than 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.

TABLE 1.

	Spanish	chi-square	p-value
1	non-agent sbj (no volition)	129,785933	<0,001
2	Unaccusative V	96,391268	<0,001
3	V of Appearance	72,9405773	<0,001
4	relative clause	71,0134903	<0,001
5	V of directed movement	47,7637428	<0,001
6	indefinite sbj	36,5741623	<0,001
7	V of Occurrence	36,1585354	<0,001
8	discourse new sbj	34,1611428	<0,001
9	Transitive V	32,518726	<0,001
10	sbj todo	30,9036948	<0,001
11	Given Predicate	25,5316137	<0,001
12	Intransitive V	23,0607253	<0,001
13	Decausative reflexive V	22,4422042	<0,001
14	Copula V	9,3327433	<0,005
15	V of Stance	5,94830663	<0,05
16	« Frog goes to dinner »	5,0754379	<0,05
17	quantified subject	4,82319875	<0,05
18	Psychological reflexive V	4,80499178	<0,05

References

- Contreras, Heles 1976. *A Theory of Word Order with Special Reference to Spanish*. Amsterdam: North Holland.
- Creissels, Denis 2006. *Syntaxe générale, une introduction typologique*. Paris: Hermès.
- Hatcher, Anne G. 1956. Theme and underlying question. Two studies of Spanish word order. *Word* 12, 14-31.
- Kennedy, Becky 1999. Focus constituency. *Journal of Pragmatics* 31, 1203-1230.
- Witten, Ian H. and Frank, Eibe 2005. *Data Mining: Practical Machine Learning Tools and Techniques* (Second Edition). San Francisco: Morgan Kaufmann

- 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, Knud 1995. The pragmatics of case: On the relationship between semantic, grammatical, and pragmatic roles in English and French. In Shibatani, Masayoshi and Sandra A. Thompson (eds.), *Essays in Semantics and Pragmatics. In Honor of Charles J. Fillmore*, 145-190. Amsterdam: Benjamins.
- 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 Jean-Marie 2003. Inversion du sujet et structure de l'information dans les langues romanes. In Godard, Danièle (ed.), *Langues romanes. Problèmes de la phrase simple*. Paris: Editions du CNRS.
- Torrego, Esther 1984. On inversion in Spanish and some of its effects. *Linguistic Inquiry* 15, 102-129.
- Zubizarreta, María Luísa 1998. *Prosody, Focus and Word Order*. Cambridge, Mass.: MIT Press.
- Zubizarreta, María Luísa 1999. Las funciones informativas: tema y foco. In Bosque, Ignacio and Demonte, Violeta (eds.), *Gramática descriptiva de la lengua española vol. 3*, 4215-4244. Madrid: Espasa Calpe.
-