**An Experimental Investigation of Ne-cliticization as a Split-intransitivity Diagnostic in Italian**

Pietro Cerrone and Jon Sprouse
Program in Psychology, New York University Abu Dhabi

**Abstract:** We report two acceptability judgment experiments investigating *ne-cliticization* as a split-intransitivity diagnostic in Italian. We test 20 verbs spanning 5 lexical semantic categories, and compare *ne* to another split-intransitivity diagnostic, the Absolute Small Clause. We use hierarchical clustering and linear mixed effects models to explore the behavior of *ne* and ASC from the perspectives of both the binary Unaccusative Hypothesis and the gradient Lexico-Semantic Hypothesis. Our results suggest that *ne-cliticization* does not behave as a split-intransitivity diagnostic under either the binary or the gradient approach to split-intransitivity, whereas the ASC shows a binary split consistent with the Unaccusative Hypothesis. (99 words)

**Keywords:** ne-cliticization, split-intransitivity, unaccusativity, experimental syntax, acceptability judgments

## 1 Introduction

A critical empirical claim in generative syntactic theory, first introduced by Perlmutter (1978), is that intransitive verbs can be divided into subcategories based on syntactic and/or semantic properties – a phenomenon known as *split-intransitivity* (henceforth SI). This claim crucially hinges upon identifying robust diagnostics of these subcategories; for a discussion of various diagnostics cross-linguistically, see, among others: Perlmutter 1978, 1989, Burzio 1981, 1986, Levin & Rappaport Hovav 1995, Rosen 1984, Zaenen 1993, Sorace 2000, Alexiadou et al. 2004, Bentley 2006. In this paper, we present an empirical evaluation of one of the most famous SI diagnostics – *ne*-cliticization in Italian (Belletti & Rizzi 1981, Rosen 1984, Burzio 1986, and Perlmutter 1989, et seq.; see Bentley 2006 for an extensive overview of Italian SI). The traditional claim is that *ne*-cliticization divides intransitive verbs into two classes: *unaccusative* verbs, which allow *ne*, and *unergative* verbs, which do not as in (1-2):

(1) $Ne_i$ arrivano [molti $t_i$]. UNACCUSATIVE
NE arrive.3.PL many
"There arrive many of them."

(2) *$Ne_i$ suonano [molti $t_i$]. UNERGATIVE
NE play.3.PL many
"Many of them play music."

Our goal is to experimentally test to what extent *ne*-cliticization is a robust diagnostic of SI.

We focus on *ne*-cliticization because it is one of the most frequently cited SI diagnostics, appearing across a wide range of work in generative grammar: e.g., it is offered as evidence for SI in Haegeman's 1994 textbook on government and binding; it appears in many of the most cited works on SI (e.g., Levin & Rapport Hovav 1995 and Alexiadou et al. 2004); it appears in monographs on lexical categories (e.g., Baker 2003); it appears in work dealing with Italian dialects (e.g., Suñer 1992, Parry 2000); and it is presented as a paradigmatic diagnostic in work on non-Romance languages (e.g. Harves 2009). However, there are studies that have challenged the judgments reported in (2), instead reporting that *ne* can appear with both unaccusative and unergative verbs (perhaps across the board, or perhaps only under certain circumstances, see: Lonzi 1986, Saccon 1992, Bentley 2004, Calabrese and Maling 2009, Glushan and Calabrese 2014). In this study, we test 20 verbs in Italian (spanning 5 lexical-semantic classes that instantiate both the binary unaccusative/unergative distinction and a gradient lexical-semantic distinction) in two basic declarative sentences (with and without a preposed prepositional phrase) in two acceptability judgment experiments (with 41 and 45 participants, respectively) to evaluate the basic claim that *ne-cliticization* can diagnose SI. Anticipating our results slightly, we find no evidence that *ne*-cliticization is sensitive to subclasses of intransitive verbs. We describe the logic of our experimental designs and results in more detail below.

## 2 The Logic of the Present Study

There is an active debate in the SI literature between at least two prominent theories: the Unaccusative hypothesis (UH) (Perlmutter 1978, Burzio 1986) and the Lexico-Semantic hypothesis (LSH) (Sorace 2000). The UH proposes two classes of verbs based on an underlying syntactic difference (that may be encoding a semantic difference; see Levin and Rappaport-Hovav 1995), while the LSH proposes several categories (up to 7) based on underlying lexical semantic differences like agentivity and telicity. Resolving this debate is not our primary concern. That said, it is critical for us to test the full range of possible categories to ensure that our experiments have the best chance to detect SI, regardless of the form that SI takes. To that end, for both experiments, we selected a set of 20 verbs based on 5 putative lexical-semantic categories (4 verbs per category) based on the lexical-semantic categories from Sorace 2000: change of location, change of state, state (a category that combines continuation of a pre-existing state and existence of a state from Sorace 2000), controlled motional process, and controlled non-motional process. From the perspective of the UH, these 20 verbs would be split between 8 unaccusative verbs (encompassing change of location and change of state), 8 unergative verbs (encompassing controlled motional and controlled non-motional processes), and 4 that are typically categorized as unaccusative, but may also be unergative (state). Table 1 lists the 20 verbs, divided into the 5 lexical-semantic categories, that we selected for the experiments. We use these categories a priori in order to ensure a representative selection of verbs, but we will be conducting verb-level cluster analyses to empirically determine the number of categories and the verbs within them. In this way, we will avoid losing information due to unknowingly averaging different verb types together.

**Table 1**: The 20 verbs in our study divided into 5 lexical-semantic categories.

| **Verb Class** | **Verbs** | | | |
|---|---|---|---|---|
| Change of location | *venire* | *arrivare* | *cadere* | *entrare* |
| | 'come' | 'arrive' | 'fall' | 'come-in' |
| Change of state | *morire* | *nascere* | *fiorire* | *marcire* |
| | 'die' | 'be born' | 'bloom' | 'rot' |
| State | *rimanere* | *sopravvivere* | *bastare* | *apparire* |
| | 'stay' | 'survive' | 'be enough' | 'appear' |
| Controlled motional process | *ballare* | *nuotare* | *volare* | *correre* |
| | 'dance' | 'swim' | 'fly' | 'run' |
| Controlled non-motional process | *ridere* | *lavorare* | *suonare* | *telefonare* |
| | 'laugh' | 'work' | 'play' | 'call' |

In the first experiment, we test the *ne*-cliticization diagnostic through two conditions, with and without *ne* as in (3a-b). We built the items with a sentence-initial prepositional phrase to maximize the felicity of the sentences, particularly with *ne*.

(3) a. Alla festa, ne arrivano molte, di amiche.
to.the party NE arrive.3.PL many.F.PL of friend.F.PL
"There arrive many friends to the party."

b. Alla festa, arrivano molte amiche.
to.the party arrive.3.PL many.F.PL friend.F.PL
"There arrive many friends to the party."

In the second experiment, we test two SI diagnostics. The first is the *ne*-cliticization diagnostic again, but this time without the sentence-initial PP. Bentley 2004 claims that the presence of locational phrases license *ne* (but cf. Saccon 1992 for a different view). If that were the case, then a failure to find SI in the first experiment could be because of the presence of the sentence-initial PPs. If we find the same lack of SI without sentence-initial PPs, we can be more confident that *ne* is not a diagnostic of SI. The second diagnostic is the absolute small clause (henceforth ASC) as in (4a-b) (Perlmutter 1989, Belletti 1981, 1990, 1992, 1999, Egerland, 1996, Cinque 1990, Dini 1994; see Loporcaro 2003 for a review). We included the ASC as a baseline comparison to show what a successful diagnostic might look like under our statistical analyses.

(4) a. Arrivato Gianni, Mario ha cominciato a mangiare.
arrived.M.SG Gianni Mario has started.M.SG to eat.INF
"Once Gianni arrived, Mario has started to eat."

b. Dopo che è arrivato Gianni, Mario ha cominciato a mangiare

after that is arrived.M.SG Gianni Mario has started.M.SG to eat.INF
"Once Gianni arrived, Mario has started to eat."

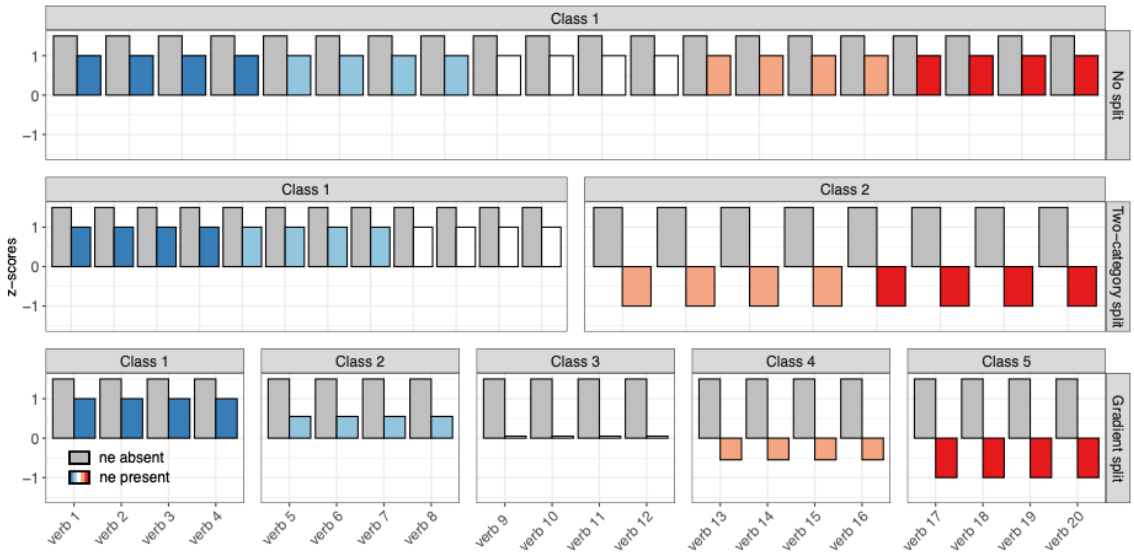
In the ASC construction in (4a), a participial clause precedes the matrix clause. Perlmutter first reported that only unaccusative verbs can be used as participial in (4a), but not unergative verbs. As a control condition, we used the complex form auxiliary + past participle (4b), which is grammatical with all intransitives, and has the same semantic relationship with the matrix clause as the ASC construction.

We divided the 20 verbs into 4 sub-experiments per each experiment. Each sub-experiment contains 5 of the verbs, one from each lexical-semantic category. This division into sub-experiments is to keep the length of the experiment reasonable for participants, and therefore to minimize satiation and/or fatigue effects. Because we wanted to look for individual or regional differences in the acceptability of *ne*-cl, we gave each participant all 4 sub-experiments, with at least 1 week between each sub-experiment (counterbalancing the order of the sub-experiments). This allows a within-participants analysis of the verbs. We recorded demographic information about participants' age and the region of Italy that they grew up in. We present a brief analysis of individual variation in section 4.4. We do not find any evidence of dialect variability. That said, the data for both experiments [will be] publicly available on the authors' websites for other researchers to analyze in additional ways.

There are three possible patterns that we will look for in the results. The first pattern is that *ne*-cliticization is not a diagnostic for SI. This would yield no significant difference in the acceptability of *ne* across the verbs as if they are all one class. This pattern does not make a specific prediction of the acceptability of *ne*-cliticization relative to the control condition, just that it would be identical across the lexical-semantic categories. This pattern is illustrated in the top row of Figure 1. The second pattern is that *ne*-cliticization is a diagnostic of SI, and that SI entails two categories as predicted by the UH. This would yield two classes of verbs with one showing acceptability and one showing unacceptability of *ne*-cl. This is illustrated in the second row of Figure 1. The third pattern is that *ne*-cliticization is a diagnostic of SI, and that SI entails multiple categories as predicted by the LSH. This would yield a gradient in acceptability: the acceptability of *ne*-cliticization will gradually decline across some number of classes. This is illustrated in the third row of Figure 1.

It is important to note that the empirical classes that arise in our results could either be aligned with the theoretical lexical-semantic classes that we used to construct the materials or misaligned with the categories. This is why we have labeled the classes in the columns and the verbs along the x-axis generically. Our cluster analyses below will empirically identify the number of classes and the verbs that are contained within them. This ensures that we can detect any form of SI, regardless of how well existing theories predict the behavior of specific verbs. We will use color to track the theoretical lexical-semantic class of each verb (i.e., the verb will always have the same color in all plots) for readers who may be interested in the alignment/misalignment.

**Figure 1:** The three possible outcomes of the experiment: no SI (top), two categories (UH; middle), and five categories (LSH; bottom). Color indicates the theoretical lexical-semantic category (though this is not part of our analysis).




## 3 The Design of the Experiments

The surveys for each experiment had generally the same composition: 3 anchor items in the instructions to illustrate a rating for the two endpoints and the midpoint of the scale (1, 4, 7), 6 "burn in" items that span the range of acceptability (presented in the same order at the start of the survey) to help participants flesh out their scale, and then the target items (10 for experiment 1; 20 for experiment 2) mixed in a pseudorandom order with filler items (17 for experiment 1; 20 for experiment 2). In the remainder of this section, we describe the construction of the experiments in detail.

### *3.1 Participants*

For experiment 1, we recruited 41 participants; for experiment 2 we recruited a different sample of 45 participants. All are self-reported native speakers of Italian who reside in Italy. (Though we see no evidence of dialectal variation in our samples, we list each anonymous participant's age and geographic region in the publicly available data file for researchers interested in potential dialectal variation.) Each participant was asked to complete all 4 sub-experiments for their assigned experiment, with each sub-experiment separated by at least one week's time, and the order counterbalanced across participants. Participants were paid 2 Euros for completing each sub-experiment, and a 2 Euro bonus for completing all 4 sub-experiments. Given the length of our experiments, this is a rate of roughly 15 Euros per hour.

*3.2 Materials*

For the target conditions for experiment 1 (*ne*-cliticization only), we created 8 lexically matched pairs for each verb for a total of 320 items (20 verbs x 2 *ne* conditions x 8 tokens). For the target conditions for experiment 2 (*ne*-cliticization and ASC) we created 8 lexically matched pairs per verb for a total of 640 items (20 verbs x 2 *ne* conditions x 8 tokens + 20 verbs x 2 ASC conditions x 8 tokens). For the filler items for experiment 1, we included 8 items from an independent island effects experiment and then constructed an additional 9 novel items that are unrelated structurally to both *ne*-cliticization and islands. For experiment 2, the filler items consisted of 8 items from an unrelated experiment about island effects, and 12 entirely unrelated items. Combined with the target items, we expected roughly half of the items in each experiment to be in the acceptable range of the scale and half of the items to be in the unacceptable range of the scale.

*3.3 Survey Construction*

We distributed the target items for each sub-experiment into 8 lists using a Latin Square procedure, such that participants did not see the same lexicalization either within or across verbs. We then combined each list with the filler items, pseudorandomized the order so that related target conditions did not follow one another, and added the 6 burn-in items to the beginning of each survey in a fixed order. The surveys were coded using IBEX (Drummond 2013). The task was rating acceptability on a 1-7 scale, where 1 was labeled as *molto brutta* 'very bad' and 7 was labeled as *molto buona* 'very good'. Items were presented one per screen, with no ability to go back after an item was rated. Each participant was sent a link and completed the experiment online at their own pace.

## 4 Results

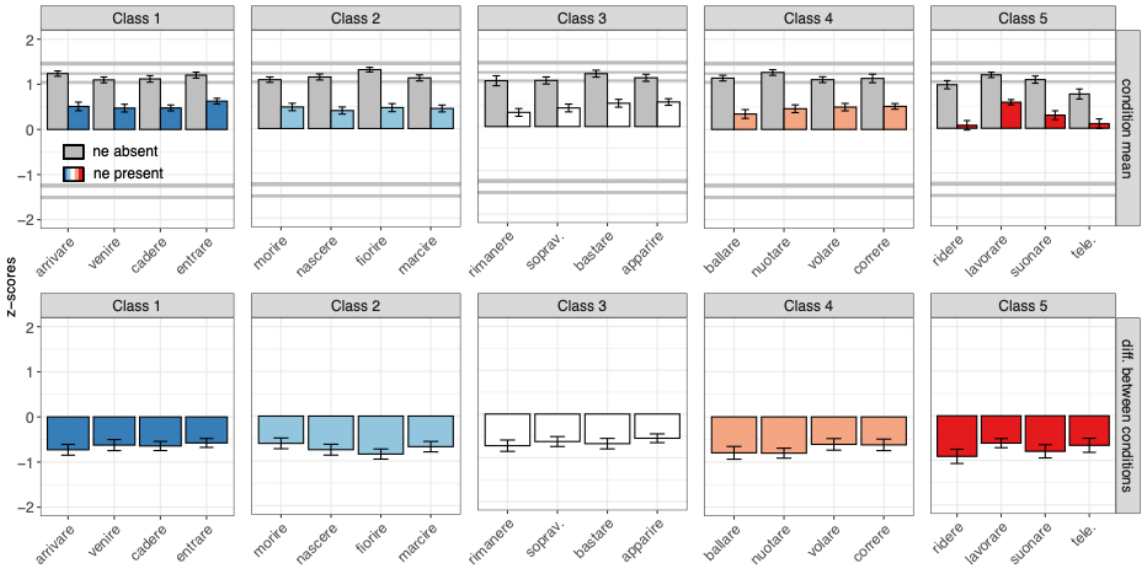
We first z-score transformed the raw judgments for each participant to eliminate certain common types of scale biases that could arise with Likert-like scaling tasks. We believe this is the most appropriate way to report judgment results (see Schütze & Sprouse 2014), however we note that there is no difference between the pattern of results with raw judgments and z-scores (and the data [will be] available for download).

*4.1 Acceptability ratings for the three diagnostics*

Figures 2-4 plots the acceptability ratings for each of the three diagnostics (*ne*+PP, *ne*, and ASC, respectively). The top row plots the means for the control and target conditions for each individual verb, organized by lexical-semantic category for convenience, along with error bars that estimate one standard error of the mean in each direction. The order of the verb classes reflects the order predicted by the LSH. The color indicates the lexical-semantic class (redundantly in this plot, but it will be useful for the cluster analyses). The horizontal gray bars mark the mean rating of the most acceptable and least acceptable filler items as an empirical estimate of the ceiling and floor

of the acceptability scale (averaged across all participants). The bottom row plots the difference between the control condition and target condition for each verb to highlight the effect size for each verb. From these plots, we can informally look for one of the three patterns discussed in section 2 and illustrated in Figure 1.

**Figure 2:** Means and difference scores for *ne*+PP.




For the results of the two *ne* experiments, we find that both of the *ne* conditions are rated on the acceptable side of the scale (zero is the midpoint after the z-transformation), except for perhaps *ridere* 'laugh' and *telefonare* 'call' in the *ne*+PP experiment, which, though numerically positive, have error bars that overlap zero (but there is no overlap in the *ne* without PP experiment). This suggests that *ne*-cliticization is not a SI diagnostic in the classic sense of creating a clearly ungrammatical sentence with unergative verbs either with or without a PP.

Though the classic conception of *ne* as a SI diagnostic is likely incorrect, we could potentially reconceptualize the diagnostic to be one that focuses on effect size such that the *ne* effect size varies by verb class. In the bottom row of Figures 2 and 3, we see that there is some minor variation in effect size across verbs, roughly between -0.5 and -0.9 in the z-score scale for *ne*+PP and between -0.3 and -0.8 for *ne* without a PP.

This variation is relatively small, and therefore likely not what either of the SI theories originally predicted. But we can nonetheless ask whether the variation leads to distinct subclasses of intransitive verbs. To evaluate this new question, we will use hierarchical cluster analysis and linear mixed effects model comparison (sections 4.2-4.3). Turning to results of the ASC experiment in Figure 4, we see a very different pattern: for some verbs, the target ASC condition is on the positive side of the scale, and for others it is on the negative side of the scale. This is in line with the logic of SI diagnostics, which predict that the construction should be

unacceptable for one or more classes of verbs. We also see variability in the control (full adjunct clause) condition.

**Figure 3:** Means and difference scores for *ne* (without a PP).

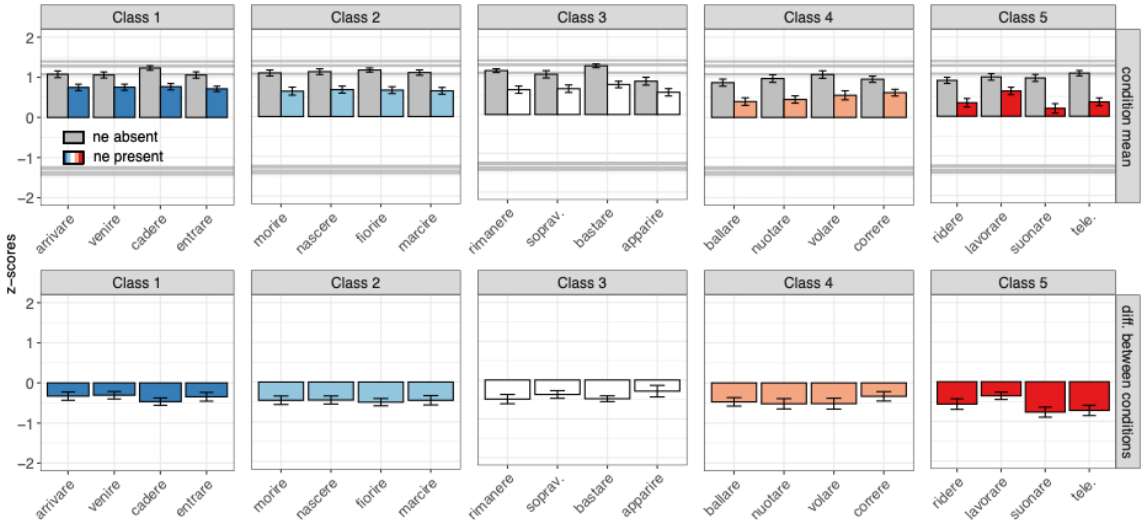
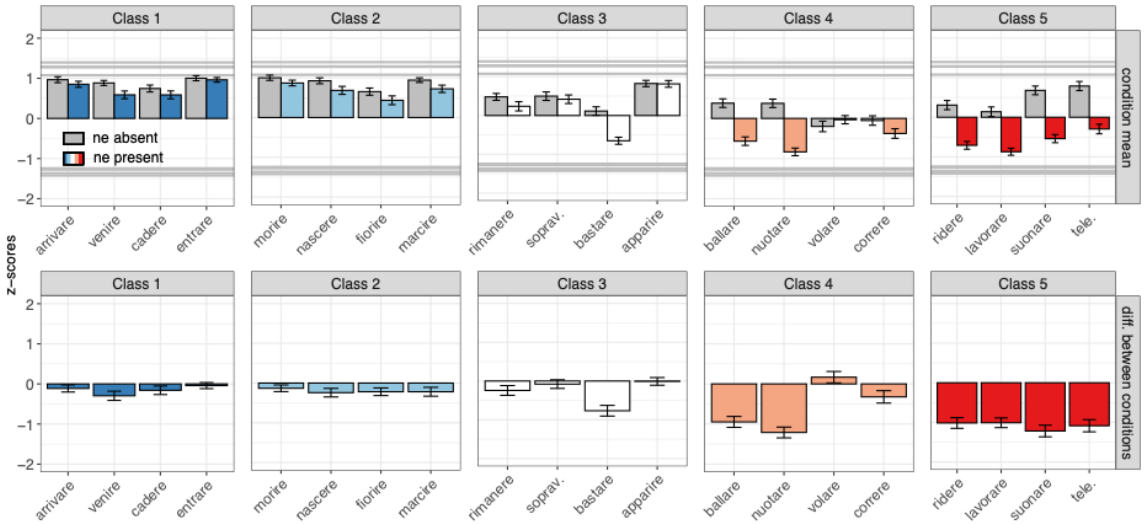



**Figure 4**: Means and difference scores for ASC.

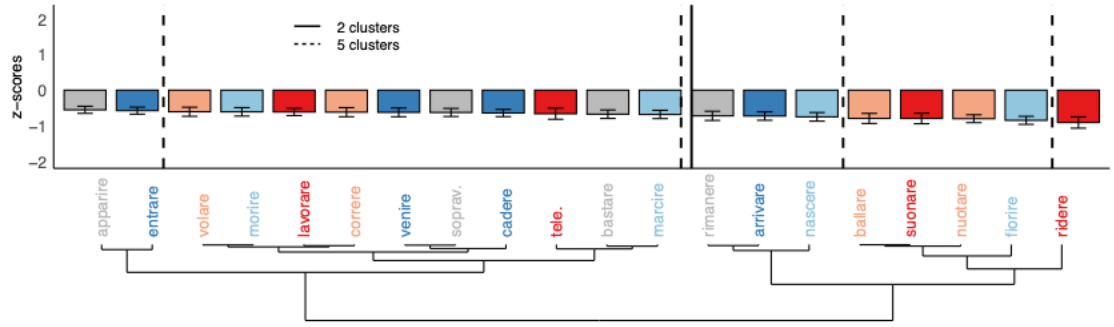



The difference scores in the bottom row of Figure 4 control for this variability, revealing two types of verbs: those that show a relatively small difference between the control condition and ASC, and those that show a relatively large difference. To explore the number of classes quantitatively, we will again use hierarchical cluster analysis and linear mixed effects model comparison (sections 4.2-4.3).

*4.2 Hierarchical cluster analysis for the three diagnostics*

We employed a hierarchical cluster algorithm to divide the twenty verbs into clusters based on the size of the difference scores for each of the three diagnostics (*ne*+PP, *ne*, and ASC). Given that the choice of clustering procedure can affect the number of clusters, we decided to choose a procedure that is biased toward smaller clusters, in line with the LSH, and contrary to the UH and no-split hypotheses. We chose this to bias against our prior personal beliefs in the UH, even though distinguishing these two theories is not a primary goal of this study. To that end, we performed agglomerative hierarchical clustering with "complete" linkage using the hclust() function in R. The dendrograms at the bottom of Figures 5-7 report the full result of the clustering (showing 2 through 20 clusters). The bar plot of Figures 5-7 re-plot the difference scores (from Figures 2-4), organized in ascending order, and split into the two theoretically relevant cluster options: 2 clusters (indicated by a solid black line) or 5 clusters (indicated by 4 dashed lines). We have retained the bar colors to indicate the by-hypothesis lexical-semantic classification according to the LSH (substituting gray for white for visibility reasons) so that interested readers can also qualitatively evaluate the mapping between lexical semantic class and cluster (though that is not a direct question of our study).

**Figure 5:** Clustering results for *ne*+PP




Turning first to the two *ne*-cliticization tests, we see that the pattern in Figures 5 and 6 descriptively corroborate the general impression of Figure 2 and 3 that there is no step-like division between verbs based on the *ne* effect size that could be used to argue that *ne*-cliticization is a SI diagnostic according to either the UH or LSH. The same conclusion emerges from exploring the two-class and five-class options in more detail: the classes are each a mix of lexical semantic verb types, which runs contrary to both the UH and the LSH.

Turning next to the ASC test, Figure 7 descriptively corroborates the impression from Figure 4 that there is a split between subclasses of verbs, particularly for the division into two classes. The division into lexical-semantic categories is also relatively uniform in the two-class split (with just *volare* 'fly' and *correre* 'run' as mismatches).

This is what we might expect from a SI diagnostic under the UH. The division into five classes shows a gradient that is similar to what is predicted by the LSH, but some of the details

might require further exploration under the LSH. *Volare* is in a class by itself because its effect goes in the opposite direction to the others (the control condition is less acceptable than the ASC condition). And, while the other 4 classes are roughly organized into two mostly-blue and two mostly-red empirical classes, there are some potential lexical-semantic misalignments: class 2 contains three types of verbs; class 3 contains four types, class 4 contains three types, and class 5 contains two types. This is not necessarily problematic for the LSH; it just means that there may be more work to do exploring the lexical-semantic properties of each of the verbs.

**Figure 6:** Clustering results for *ne* (without PP)

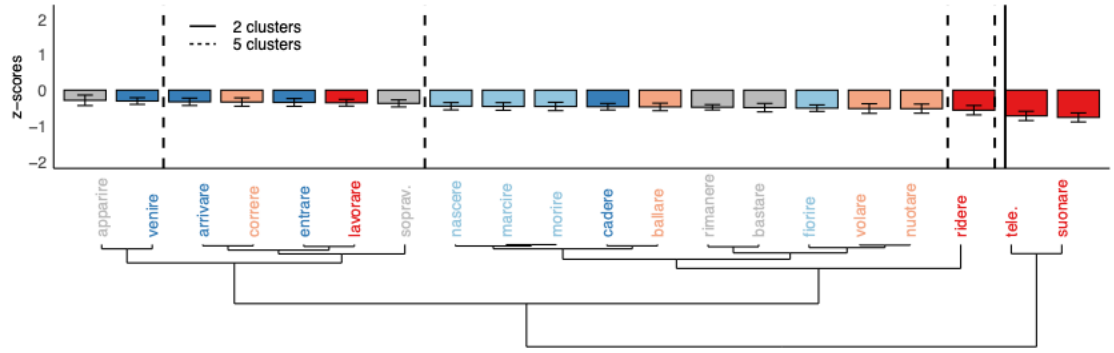
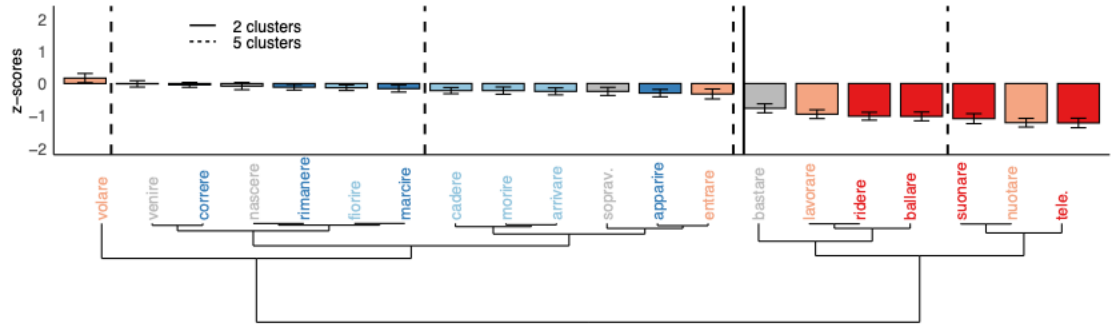



**Figure 7:** Clustering results for ASC




*4.3 Linear mixed-effects model selection for the number of clusters*

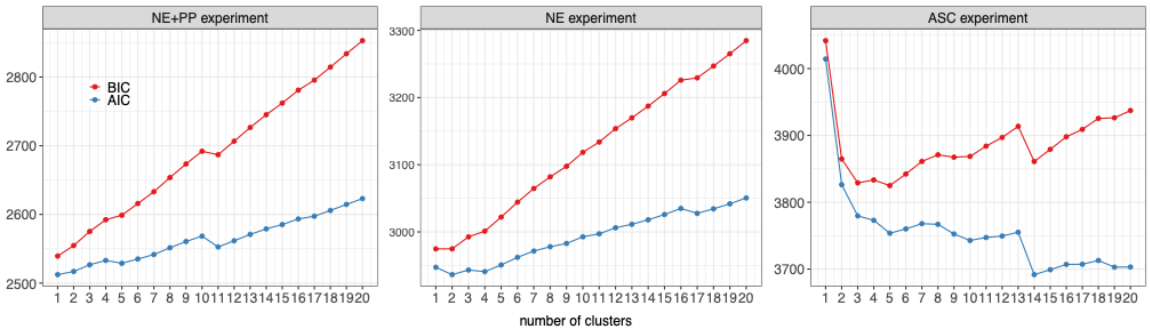
The final step of our analysis attempts to select the empirically optimal number of clusters for each diagnostic. To that end, we constructed linear mixed effects models to predict acceptability based on the interaction of target/control conditions and the number of CLUSTERS (derived from the hierarchical cluster analysis), with subject and item as random effects (intercepts only) using the lme4 package in R (Bates et al. 2015). We constructed a distinct model for each possible number of clusters (1 through 20), so that we could then compare the models using two popular model comparison metrics: the Akaike Information Criterion (AIC) (Akaike 1973) and the Bayesian Information Criterion (BIC) (Schwarz 1978). Both the AIC and BIC evaluate how well

each model predicts the observed data, and both penalize more complex models (in our case, models assuming more clusters) to navigate the trade-off between empirical coverage and theory complexity. For both metrics, the absolute value of the metric is not typically interpreted. Instead, the relative value is interpreted: a lower score is preferred to a higher score. (Though there is no categorical interpretation of either AIC or BIC, common rules of thumb are that a difference less than 2 means that two models fit the data equally well, a difference between 2 and 10 begins to favor the model with the lower value, and a difference greater than 10 is strong evidence in favor of the model with the lower value.)

We note that there is debate in the statistics literature about the relative pros and cons of AIC vs BIC. They differ in terms of their complexity penalties (e.g., the BIC tends to have a more severe complexity penalty than the AIC, making the AIC more likely to favor complex models), and they differ in terms of the philosophical approach that they instantiate (e.g., the AIC focuses on the likelihood function of the model, while the BIC focuses on the posterior probabilities). It appears that there is a slight preference for BIC in the model comparison literature, primarily because of its larger complexity penalty, and perhaps because of the rise in popularity of Bayesian methods in general. We agree with these arguments, but in the interest of providing the maximal amount of information, we will report both metrics, and to the extent possible, look for agreement between them. (We also note that for the AIC, we report values corrected for small sample sizes out an abundance of caution, but for our sample sizes, there is no difference between the corrected and uncorrected AIC.)

Figure 8 reports the BIC in red and AIC in blue for each of the 20 models for each of the three diagnostics. Turning first to *ne*+PP (left panel), we see an overall monotonically increasing pattern for both BIC and AIC: as the number of clusters increases, the BIC and AIC. For *ne*+PP, there is an increase for each model from 1 to 10, then a small decrease to 11, and then an increase from 12 to 20. This suggests that the optimal number of clusters is 1 for *ne*+PP, and that any increase in explanatory value gained with each additional cluster is outweighed by the penalty for the increase in model complexity.

**Figure 8**: AIC and BIC for the 20 models for each diagnostic




For *ne* without PP (center panel), the overall pattern is similar, except there is no change in BIC from 1 cluster to 2, and perhaps a small decrease in AIC from 1 to 2. We are not inclined to interpret this small decrease in AIC from 1 to 2 clusters as indicative of split intransitivity for

three reasons. First, this decrease is at least an order of magnitude smaller than the one that we see for ASC in the third panel, suggesting it is a different kind of effect. Second, this decrease is not seen with the BIC, suggesting that it is not a robust effect. The AIC penalizes model complexity less than the BIC, making it more sensitive to small fluctuations in predictive power. This decrease could be one such fluctuation. Finally, as we saw in the previous subsection, the second cluster in this case would be just two verbs (*telefonare* and *suonare*), leaving 18 in the first cluster, contrary to the typical prediction of the UH that the two classes would be roughly equal in size based on the verbs that we selected for our experiments. Thus, the overall pattern for the two *ne* tests is that 1 cluster is optimal. This is what we would expect from a generally small gradient in effect sizes with no clear step-like breaks indicative of cluster boundaries. We take this as quantitative corroboration of the patterns we observed in the acceptability judgments and in the cluster analyses: *ne* does not pattern like a SI diagnostic.

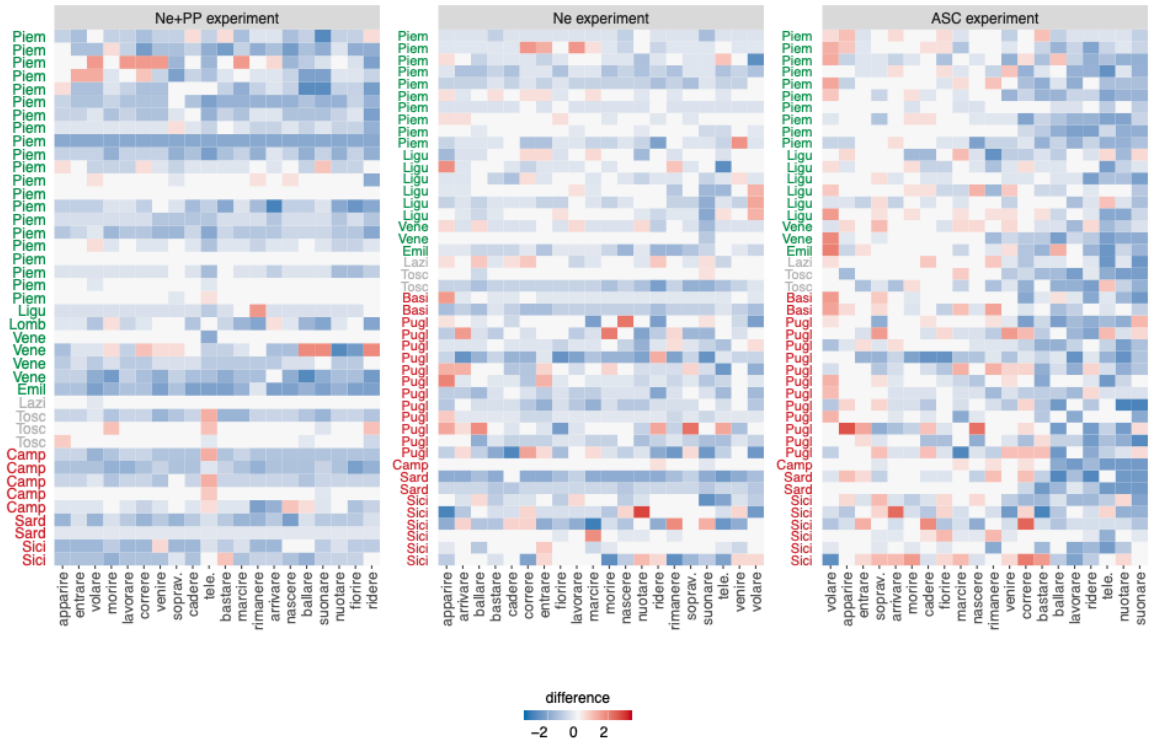
Turning next to the ASC in the right panel, we see a large decrease in both BIC and AIC from 1 to 2 clusters. We see a small additional decrease in both BIC and AIC from 2 to 3 clusters. However, from the dendrogram at the bottom of Figure 7, we can see that this third cluster only contains *volare*. This is because *volare* is an outlier - the direction of its effect is opposite to all of the others. Therefore, we are reluctant to interpret this as a statistical argument for a 3 class SI theory, and more inclined to interpret it as an argument to treat *volare* as an outlier (and perhaps explore why it behaves differently than the others in a follow-up study). For BIC, clusters 3 through 5 are roughly equivalent, followed by a general increase for each cluster up to 20 except for a decrease at 14. This suggests that the BIC identifies the optimal number of clusters as 2 (with *volare* as an outlier). The AIC result is a little more complicated. The AIC values are roughly equal for 3 clusters through 14 clusters, suggesting that 2 clusters might be optimal in this range; but then there is a decrease at 14 clusters that yields a new minimum AIC value. We suspect this is a consequence of the AIC's smaller complexity penalty (i.e., we do not believe the LSH would predict 14 distinct classes of verbs). Therefore, we are inclined to interpret the AIC as also identifying the optimal number of clusters as 2 (again, with *volare* as an outlier). Taken together, the BIC and AIC seem to converge on the ASC identifying two classes of verbs in Italian. Though this could be taken as evidence for the two-category split predicted by the UH, we note that this study was not explicitly designed to test differences between the UH and LSH (that would require a much broader range of diagnostics, including those that are central to the LSH). Therefore, we are inclined to interpret this conservatively as evidence that the ASC is a SI diagnostic, and note that future research may want to include the ASC in a broader investigation of the predictions of the UH and LSH.

### *4.4 Individual variation*

One possibility is that there are individuals who show an SI pattern for *ne*, but their results were obscured by a larger group for whom *ne* is not an SI diagnostic. To investigate this, we constructed a heatmap in Figure 9 that reports the difference between the target conditions (*ne*, ASC) and the control conditions for each verb for individual participant (i.e., target – control).

Blue indicates a negative difference, meaning that the target condition is rated lower than the control condition (as we'd expect for unergative verbs), while white indicates no difference, and red indicates a positive difference, meaning that the target condition is rated higher than the control condition (the latter two we might expect for unaccusative verbs). Each horizontal line represents the results of a single participant for each of the 20 verbs. We have labeled the participants by their home region in Italy to help reveal any geographic dialectal variation. To make the geographic location of these regions in Italy immediately clear, we differentiated them by coloring the labels: green for northern regions, gray for central regions, and red for southern regions and the islands. The verbs are ordered by their empirical clusters (not by theoretical class), with positive differences on the left (unaccusative) and negative differences on the right (unergative), as this provides the best chance of revealing a visual pattern indicative of individual variation. What we would look for are participants that show a qualitative change in color along the row indicative of a cluster break, such as red or white on the left changing to blue on the right.

**Figure 9**: Heatmaps representing differences scores for each participant in the experiments (target – control). List of abbreviations: Piemonte (Piem), Liguria (Ligu), Lombardia (Lomb), Veneto (Vene), Emilia-Romagna (Emil), Lazio (Lazi), Toscana (Tosc), Basilicata (Basi), Campania (Camp), Puglia (Pugl), Sardegna (Sard), Sicilia (Sici).



We see the pattern indicative of split-intransitivity for the ASC diagnostic in the right panel – a visible shift for almost all participants from white or red on the left to blue on the right . But we do not see this pattern for many participants in the *ne* experiments in left and center panels. There may be a few participants with this pattern, but these are visually overwhelmed by most participants showing no pattern across the verbs, with red/white cells seemingly randomly interspersed among the row, or even red/white cells toward the right. Futhermore, we see no indication of regional variation (which would appear as a pattern only in top, central or bottom rows).

We note that this is only a descriptive (and therefore qualitative) analysis, therefore we invite readers interested in individual variation to download the data to perform any additional quantitative analyses that they may have in mind.

## 5 Discussion

In this study, we tested two SI diagnostics: *ne*-cliticization (with and without PPs) and ASC. We found that ASC shows the empirical hallmarks of SI according to a combination of hierarchical cluster and linear mixed effects model analysis. That analysis further suggests that a two-class division is more compatible with the data than a three-or-more class division, which aligns more closely with the Unaccusative Hypothesis (Burzio 1986, Perlmutter 1989) than the Lexico-Semantic Hypothesis (Sorace 2000). But we note that teasing apart the Unaccusative Hypothesis and Lexico-Semantic Hypothesis was not a primary goal of the experiment, and that testing these properly will require testing a wider range of SI diagnostics, so we note this finding only to motivate future research. In contrast, *ne*-cliticization (with or without PPs) does not show the hallmarks of SI. This suggests that *ne*-cliticization is not a diagnostic of SI for the participants recruited for our experiments, at least when it appears in basic declarative sentences with simple tense. This in turn suggests that researchers interested in SI should not consider *ne*-cliticization a robust diagnostic, at least when sentences are presented in isolation (with or without PPs). Instead, the results of the ASC experiment show that ASC is a potentially reliable diagnostic for split-intransitivity (and that it is compatible with the two-category Unaccusative Hypothesis).

Before concluding, it is worth noting that the results of our experiment differ from the judgments reported by several professional linguists. There are a number of possibilities for this difference, each of which would require a dedicated study. One possibility is that there is a difference in grammar. This could, in principle be tracked to geographic or generational differences. Another possibility is that *ne* serves as a diagnostic for SI under more restricted contexts than has previously been reported. This could, in principle, be tested with an experiment that manipulates whatever context researchers believe might be relevant. For our part, we consider this a first experimental study to determine the behavior of *ne* for a sample of Italian speakers that is not restricted to any particular dialect, and for *ne* presented in standalone declarative sentences (as it has typically been presented in the literature). Researchers interested in exploring whether the split might re-emerge with different experimental designs can use our

results to formulate and test new hypotheses. To that end, both the materials and results of these experiments [will be made] freely available for exploration on the authors' websites.

## 6 Conclusion

Our goal in this paper was to experimentally test to what extent *ne*-cliticization is a diagnostic of split-intransitivity when appears in plain declaratives (with and without a preposed PP). We also tested the absolute small clause diagnostic as a comparison. To that end, we tested a set of 20 verbs (4 each from 5 lexical-semantic) in two experiments, testing *ne*-cliticization in both experiments, and the ASC in one. Using a hierarchical clustering and model comparison we find no evidence in either experiment that *ne*-cliticization is a split-intransitivity diagnostic; but we do find evidence that ASC is a split-intransitivity diagnostic (and that is compatible with the two-category Unaccusative Hypothesis).

Though this is a negative result insofar as it suggests that *ne*-cliticization in simple declarative sentences cannot be used as a diagnostic for split-intransitivity, we see this study as a first step in a larger project to experimentally investigate split-intransitivity diagnostics in Italian. This study reveals both an experimental design and set of descriptive and inferential statistical analyses that can be used to investigate a wide range of split-intransitivity diagnostics, as evidenced by the success of the ASC diagnostic here. This study also points to the kinds of follow-up studies that could be developed for *ne*-cliticization if speakers of Italian find that it is still a diagnostic in their own judgments: *ne* could be tested in contexts that highlight telicity and agentivity; *ne* could be tested with complex tenses; or *ne* could even be tested in auditory experiments if speakers believe that intonation plays a role. Our hope is that this study generates new discussions about this classic phenomenon.

## References


Akaike, Hirotsugu. 1973. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, ed. by. Frigyes Csáki and B. N. Petrov, 267-281. Budapest: Akadémiai Kiadó.

Alexiadou, Artemis, Elena Anagnostopoulou, and Martin Everaert (eds) 2004. *The Unaccusativity Puzzle: Explorations of the Syntax-Lexicon Interface*, Oxford: OUP.

Baker, Mark. 2003. *Lexical categories: verbs, nouns, and adjectives*. Cambridge: Cambridge University Press

Bates, Douglas, Martin Mächler, Ben Bolker and Steve Walker. 2015. Fitting Linear Mixed Effects Models Using lme4. *Journal of Statistical Software* 67: 1-48.

Belletti, Adriana. 1981. Frasi ridotte assolute. *Rivista di Grammatica Generativa* 6: 3-32.

Belletti, Adriana. 1990. *Generalized Verb Movement*. Turing: Rosenberg & Sellier.

Belletti, Adriana. 1992. Agreement and Case in Past Participle Clauses in Italian. In *Syntax and the Lexicon. Syntax and Semantics 26*, ed. by Tim Stowell and Eric Wehrli, 21-44. San Diego: Academic Press.

Belletti, Adriana. 1999. Italian/Romance clitics: Structure and derivation. In *Clitics in the languages of Europe*, ed. by Henk van Riemsdijk, 543-579. Berlin: Mouton de Gruyter.

Belletti, Adriana, and Luigi Rizzi. 1981. The Syntax of *ne*: Some Theoretical Implications. *The Linguistic Review* 1: 117-154.

Bentley, Delia. 2004. Ne-Cliticisation and Split Intransitivity. *Journal of Linguistics* 40: 219-262.

Bentley, Delia. 2006. *Split Intransitivity in Italian*. Berlin, New York: De Gruyter Mouton.

Burzio, Luigi. 1981. Intransitive verbs and Italian Auxiliaries. Doctoral dissertation, MIT.

Burzio, Luigi. 1986. *Italian Syntax. A Government-Binding Approach*. Dordrecht: Reidel Publishing Company.

Calabrese, Andrea, and Joan Maling. 2009. *Ne* Cliticization and Auxiliary Selection: Agentivity Effects in Italian. MS, University of Connecticut/Brandeis University. Cinque, Guglielmo. 1990. Ergative Adjectives and the Lexicalist Hypothesis. *Natural Language and Linguistic Theory* 8: 295-331.

Dini, Luca. 1994. Aspectual Constraints on Italian Absolute Phrases. *Quaderni del Laboratorio di Linguistica della Scuola Normale Superiore di Pisa* 8: 52-87.

Drummond, Alex. 2013. Ibex Farm. (Available until September 30th 2021).

Egerland, Verner. 1996. *The Syntax of Past Participles. A Generative Study of Nonfinite Constructions in Ancient and Modern Italian*. Lund: Lund University Press.

Glushan, Zhanna, and Andrea Calabrese. Context Sensitive Unaccusativity in Russian and Italian. In *Proceedings of the 31st West Coast Conference on Formal Linguistics*, ed. by. Robert E. Santana-LaBarge, 207-217. Somerville, MA: Cascadilla Proceedings Project.

Haegeman, Liliane. 1994. *Introduction to Government and Binding Theory. Second Edition*. Oxford: Blackwell Publishers Ltd.

Harves, Stephanie. 2009. Unaccusativity. In *Handbooks of Linguistics and Communication Sciences: Slavic Languages, Vol. 1, 32*, ed. by Jeroen Darquennes and Patience Epps, 415-430. Berlin: Mouton de Gruyter.

Levin, Beth, and Malka Rappaport Hovav. 1995. *Unaccusativity: At the Syntax-Lexical Semantics Interface*. Cambridge: MIT Press.

Lonzi, Lidia. 1986. Pertinenza della struttura tema-rema per l'analisi sintattica. In *Tema Rema in italiano*, ed. by Harro Stammerjohann, 99-120. Tübingen: Narr.

Loporcaro, Michele. 2003. The Unaccusative Hypothesis and participial absolutes in Italian: Perlmutter's generalization revised. *Rivista di Linguistica* 15: 199-263.

Parry, Mair. 2005. *Sociolinguistica e grammatica del dialetto di Cairo Montenotte. Parluma 'd còiri.* Società savonese di storia patria.

Perlmutter, David M. 1978. Impersonal Passives and the Unaccusative Hypothesis. In *Proceedings of the 4th Annual Meeting of the Berkeley Linguistic Society*, 157-189.

Perlmutter, David M. 1989. Multiattachment and the Unaccusative Hypothesis: The Perfect Auxiliary in Italian. *Probus* 1: 63-119.

R Core Team. 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. Available online at: https://www.R-project.org/

Rosen, Carol. 1984. The Interface between Semantic Roles and Initial Grammatical Relations. In *Studies in Relational Grammar 2*, ed. by David M. Perlmutter and Carol Rosen, 38-77. Chicago / London, The University of Chicago Press.

Saccon, Graziella. 1992. VP-internal arguments and locative subjects. In *Proceedings of the 22nd Annual Meeting of the North East Linguistic Society*, ed. by Kimberley Broderick, 383-397. Amherst, MA: GLSA.

Schütze, Carson, and Jon Sprouse. 2014. Judgment Data. In *Research methods in linguistics*, ed. by Robert Podesva and Devyani Sharma, 27-51. Cambridge: Cambridge University Press.

Schwarz, Gideon E. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461-464.

Sorace, Antonella. 2000. Gradients in auxiliary selection with intransitive verbs. *Language* 76: 859-890.

Suñer, Margarita. 1992. Clitics in the Northern Italian Vernacular and the Matching Hypothesis. *Natural Language and Linguistic Theory* 10: 641-672.

Zaenen, Annie. 1993. Unaccusativity in Dutch: Integrating Syntax and Lexical Semantics. In *Semantics and the Lexicon*, ed. by James Pustejovsky, 129-161. Dordrecht: Kluwer.