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

Pietro Cerrone
New York University Abu Dhabi
Program in Psychology
Abu Dhabi, UAE
pietro.cerrone@nyu.edu

Jon Sprouse
New York University Abu Dhabi
Program in Psychology
Abu Dhabi, UAE
jon.sprouse@nyu.edu

1. Introduction

A critical empirical claim in generative syntactic theory, first introduced by Perlmutter (1978), is that intransitive verbs can be divided into some number of distinct categories based on syntactic and/or semantic properties – a phenomenon known as *split-intransitivity*. The justification for this claim crucially hinges upon identifying robust diagnostics of these categories; 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 split-intransitivity 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 split intransitivity and references therein). The traditional claim regarding *ne*-cliticization is that it divides intransitive verbs into two classes: *unaccusative* verbs, which allow *ne*, and *unergative* verbs, which do not as in (1-2) below (Perlmutter 1978).

(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 primary goal in this study is to experimentally test to what extent *ne*-cliticization is a robust diagnostic of split-intransitivity.

We focus on *ne*-cliticization in this study because it is one of the most frequently cited split-intransitivity diagnostics, appearing across a wide range of work in generative grammar: e.g., it occupies a privileged space in the description of the evidence for split-intransitivity in Haegeman's 1994 textbook on government and binding; it appears in many of the most cited works on split-intransitivity (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 a set of studies that have challenged the judgments reported in (2), instead reporting that *ne* can appear with both unaccusative and unergative verbs under certain circumstances. (Lonzi 1986, Saccon 1992, Calabrese and Maling 2009, Glushan and Calabrese 2014). If true, this could suggest that *ne* is not a diagnostic of split-intransitivity. To explore this possibility, we test 20 verbs in Italian across a range of 5 lexical-semantic classes that instantiate both the binary unaccusative/unergative distinction and a gradient lexico-semantic distinction in two acceptability judgment experiments (with 41 and 45 participants, respectively). Anticipating our results slightly, we find no evidence that *ne* is sensitive to subclasses of intransitive verbs. We describe the logic of our experimental designs and results in more detail below; but our ultimate conclusion is that *ne*-cliticization appears not to be a split-intransitivity diagnostic, at least for the speakers of Italian who participated in our experiments.[1]

2. The Logic of the Present Study

There is an active debate in the split-intransitivity 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; Levin and Rappaport-Hovav 1995), while the LSH proposes several categories (up to 7) based solely 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 split-intransitivity, regardless of the form that it 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 categories in Sorace 2000: change of location, change of state, state (a category that combines continuation of a pre-existing state and existence of a state category from Sorace 2000), controlled motional process, and controlled non-motional process. From the perspective of the UH, these 20 verbs should 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 frequently 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:

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

Verb Class	**Verbs**

[1] We recognize that this marks a departure from the original reported observations. It is possible that there has been a change in the idiolects of Italian over the past 45 years. It is also possible that future work could identify intonational or semantic contexts for *ne* cliticization that might cause the split to re-emerge. For this study, our goal is to explore *ne* cliticization in the default, standalone presentation that has been used in the existing literature.

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

We use these categories a priori to design our experiments in order to ensure a representative selection of verbs, but we will be conducting verb-level cluster analyses on our experimental results such that we will be able to detect any categorical distinctions that arise, regardless of which putative lexical-semantic category the verbs are theoretically assigned to. In this way, we will avoid losing information due to unknowingly averaging different verb types together.

In the first experiment, we test only the *ne* cliticization diagnostic. We built the items with a sentence-initial prepositional phrase with the two conditions as in (3a-b) in order 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 split-intransitivity diagnostics. The first is the *ne* cliticization diagnostic again, but this time without the sentence-initial PP as in (4a-b).

(4) a. Ne arrivano molte, di amiche.
NE arrive.3.PL many.F.PL of friend.F.PL
"There come many friends."

b. Arrivano molte amiche
arrive.3.PL many.F.PL friend.F.PL
"Many friends come."

We re-tested *ne* without sentence-initial PPs because Saccon 1992 claims that sentence-initial PPs license *ne*. If that were the case, then the failure to find split-intransitivity in the first experiment

could be because of the presence of the sentence-initial PPs. If we find the same lack of split-intransitivity without sentence-initial PPs, we can be more confident that *ne* is not a diagnostic of split-intransitivity. The second diagnostic that we tested is the absolute small clause (ASC) as in (5a-b) (Perlmutter 1989, Belletti 1981, 1990, 1992, 1999, Egerland, 1996, Cinque 1990, Dini 1994; see Loporcaro 2003 for a critical overview). We included the ASC as a baseline comparison for what a successful diagnostic would look like under our cluster analyses.

(5) 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."

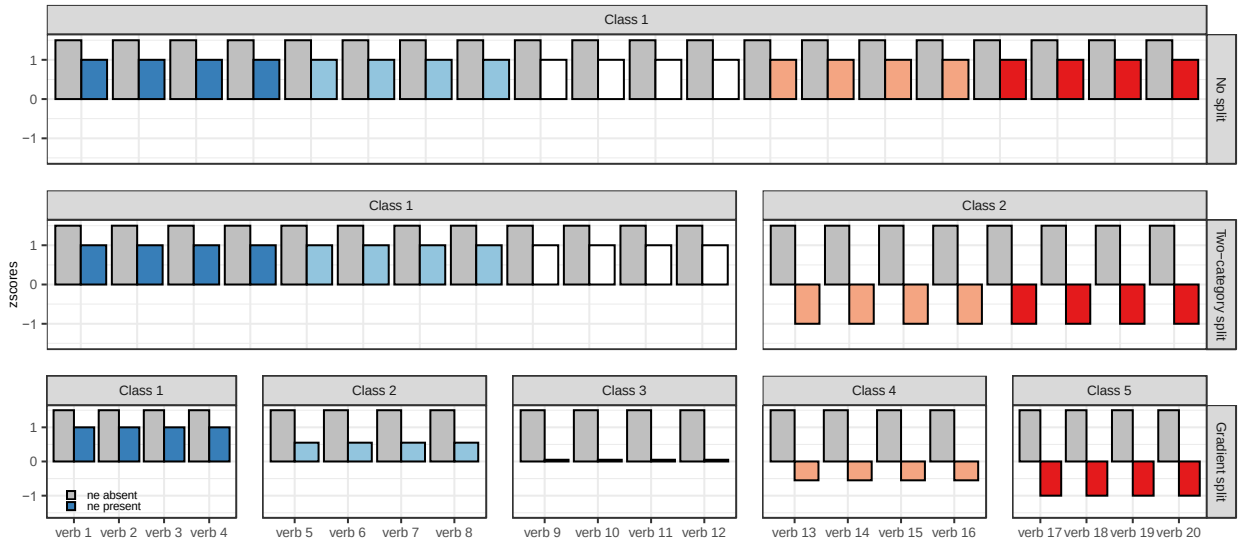
In (5a) we see an example of the absolute small clause, where absolute means "without auxiliary". In this construction, a "participial" absolute small clause precedes the matrix clause. Perlmutter first noticed that only unaccusative verbs can be used as participial in (5a), but not unergative verbs. We built control conditions as in (5b), where instead of the participial absolute form of the verb, we used the complex form auxiliary + past participle which is grammatical with all intransitives. The control conditions have complex prepositions such as *dopo che* "after that", that make explicit the circumstantial relation with the matrix clause that remains implicit in the meaning of the absolute small clause.

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 physical experiment reasonable for participants, and therefore to minimize satiation and/or fatigue effects. However, we also wanted to be able to look for individual or regional differences in the acceptability of *ne*. Therefore, we gave each participant (in both experiments) all 4 sub-experiments, with at least 1 week between each sub-experiment. The sub-experiments were given to each participant in a different order to partially counterbalance for ordering effects. This allows a completely within-participants analysis of the verbs, and crucially, allows us to look at both the group and individual level for evidence of split intransitivity. We didn't restrict participants to a specific area of Italy, and we recorded demographic information about age and area of Italy of each participant. We did not find any evidence of dialect variability based on area or age, therefore for space reasons, we do not present those analyses here. That said, all of the data for both experiments will be publicly available on the authors' websites for other researchers to analyze.

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 split intransitivity. In this case, we expect to find no significant difference in the acceptability of *ne* across the verbs (as if they are all one class). This

pattern does not make any specific prediction on the actual acceptability of *ne*-cliticization: whatever level of acceptability it shows among speakers of Italian, it would be stable across the lexico-semantic categories. This potential outcome is illustrated in the top row of Figure 1.

Figure 1: The three possible outcomes of the experiment. The top row represents no split-intransitivity, the middle row represents two categories as predicted by the UH. And the bottom row represents five categories as predicted by the LSH. We include color to indicate the putative lexical-semantic category of each verb regardless of the empirical category determined by our results. Here, we have indicated perfect alignment, but this won't necessary be the case for the actual results.



The second pattern is that *ne*-cliticization is a diagnostic of split intransitivity, and that split intransitivity entails two categories (e.g., as predicted by the UH). Under this scenario, we expect to find a clear distinction between two groups of verbs with one class showing acceptability and one class showing unacceptability of *ne*-cliticization. This is illustrated in the second row of Figure 1. The third pattern is that *ne*-cliticization is a diagnostic of split intransitivity, and that split intransitivity entails multiple categories (e.g., as predicted by the LSH). In this case, we expect to find a gradient in acceptability: the level of acceptability of *ne*-cliticization is predicted to gradually decline across some number of classes. This is illustrated in the third row of Figure 1 for five classes in line with the classes we used to construct our materials. It is important to note again 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 (this is the strongest prediction of the two existing theories in the literature), or the empirical classes could be misaligned with the lexical-semantic categories. This is why we have labeled the classes in the columns and the verbs along the facet columns generically. Our cluster analyses below will be agnostic about the composition of the classes to ensure that we can detect any empirical distinction across verbs in *ne*. We will use

color to track the theoretical lexical-semantic class of each verb (i.e., the verb will always have a given color) in order to make the alignment or misalignment visible.

3. Experiment 1: *Ne-cliticization* with Sentence-initial Prepositional Phrases

3.1 Division into 4 Sub-experiments

As mentioned above, to keep the length of the surveys reasonable, we split the 20 verbs into 4 sub-experiments, each containing 5 verbs, one from each lexical-semantic class. Table 2 lists the distribution of verbs in each sub-experiment.

Table 2: The set of verbs used in each sub-experiment for experiment 1.

Sub-experiment 1	**Sub-experiment 2**	**Sub-experiment 3**	**Sub-experiment 4**
Arrivare 'arrive'	Venire 'come'	Cadere 'fall'	Entrare 'come-in'
Fiorire 'bloom'	Morire 'die'	Nascere 'be born'	Marcire 'rot'
Rimanere 'stay'	Sopravvivere 'survive'	Bastare 'be enough'	Apparire 'appear'
Nuotare 'swim'	Ballare 'dance'	Volare 'fly'	Correre 'run'
Telefonare 'call'	Ridere 'laugh'	Suonare 'play'	Lavorare 'work'

3.2 Survey and Materials Construction

The surveys for experiment 1 consisted of 3 anchor items in the instructions, 6 items in the same order for each participant at the start of each survey to help participants acclimate to the task, 10 target items (5 verbs x 2 *ne* conditions), 8 filler items that are the target conditions for an unrelated experiment about island effects, and 9 independent filler items, for a total of 3 + 33 items. In the remainder of this subsection, we describe the construction of the items in the surveys in detail.

For the target conditions, we created 8 lexically matched pairs for each verb for a total of 320 items (20 verbs x 2 *ne* conditions x 8 tokens). We divided the target items for each sub-experiment into 8 lists using a Latin Square design such that participants did not see the same lexicalization either within or across verbs

For the filler items, 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 the 8 items from a separate experiment, we expect 2 items to be judged unacceptable, and 6 to be judged acceptable. The 9 novel fillers were constructed to span the lower rating of the judgment scale. We therefore expect roughly half of the items to be in the acceptable range of the scale and half of the items to be in the unacceptable range of the scale.

For the anchor items in the instructions, we created 3 items to demonstrate ratings of the two endpoints and midpoint of the 7-point scale (1, 4, 7). For the unannounced practice items, we created 6 items that span the range of acceptability to help participants figure out how to use the scale before rating target items. These appear as the first 6 items in the survey in the same order for all participants. All materials will be made publicly available along with the raw data.

3.3 Participants and Procedure

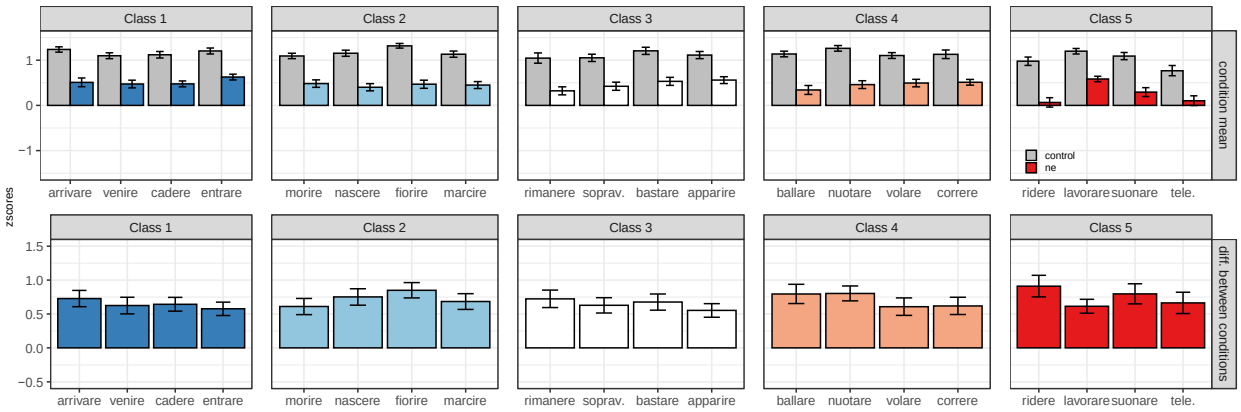
We recruited 41 participants. All are self-reported native speakers of Italian who reside in Italy. (Though we see no evidence of dialectal variation in our samples, we list each anonymous participant's age and geographic region in the publicly available data file for researchers interested in potential dialectal variation.) Each participant was asked to complete all 4 sub-experiments, with each sub-experiment separated by at least one week's time. Participants were paid 2 Euros for completing each sub-experiment, and a 2 Euro bonus for completing all 4 sub-experiments. For a 33 item survey, this equates to a rate of roughly 15 Euros per hour.

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". We used IBEX (Drummond 2013) to present the items one at a time, 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.

3.4 Results for the Ne + PP diagnostic

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. The top row of Figure 2 plots the means for the control (*ne* absent) and *ne* conditions for each individual verb, organized by lexico-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 also indicates the lexico-semantic class (redundantly in this plot, but it will be useful for the cluster analyses). The bottom row of Figure 2 plots the difference between the control condition and *ne* for each verb to highlight the effect size for each verb.

Figure 2: The top row reports means (z-scores) for each condition (control and *ne*) for each verb, organized by category. The bottom row reports the difference between *ne* and the control condition. Error bars represent estimated standard error.

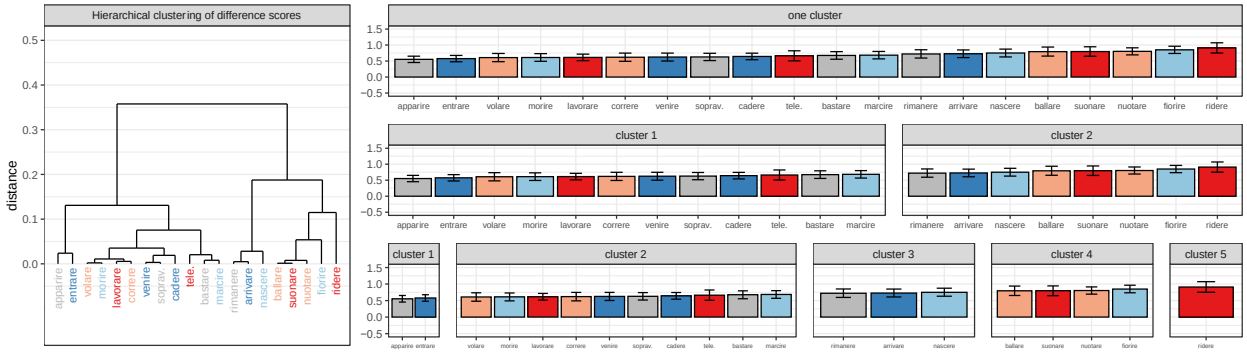


We find that all of the *ne* conditions are on the acceptable side of the scale (zero is the midpoint of the z-score scale), except for perhaps *ridere* and *telefonare*, which, though numerically positive, have error bars that overlap the midpoint (zero). This suggests that *ne* + PP is not a split-intransitivity diagnostic in the classic sense of creating a clearly ungrammatical sentence with unergative verbs. Though the classic conception of *ne* as a split-intransitivity diagnostic is likely incorrect, we could potentially reconceptualize the diagnostic to be one that focuses on effect size – perhaps the *ne* effect size varies by some number of verb classes. In the bottom row of Figure 2, we see that there is some minor variation in size across verbs, roughly between – 0.5 and – 0.9 in the z-score scale. It is relatively small, but nonetheless, we can ask whether it matches the predictions of the UH and LSH. To evaluate that question, we will use a combination of hierarchical cluster analysis and linear mixed effects modeling.

3.5 Hierarchical Cluster Analysis and Linear Mixed Effects Models for the Ne + PP Diagnostic

We analyze the results in two steps. The first step is to divide the twenty verbs into clusters based on the size of the *ne* effect. 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. To that end, we performed agglomerative hierarchical clustering with "complete" linkage using the hclust() function in R. The left panel of Figure 3 reports the full result of the clustering as a dendrogram (showing 2 through 20 clusters). The right panels of Figure 3 re-plot the effect sizes (from Figure 2), organized in ascending order, and split into 1, 2, or 5 clusters (the three theoretically relevant clusters). We have retained the bar colors to indicate the by-hypothesis classification according to the LSH (substituting gray for white for visibility reasons).

Figure 3: The results of the clustering algorithm on the *ne* + PP effect sizes. The left panel reports the full results of the clustering as a dendrogram (showing 2 through 20 clusters). The right panels plot the *ne* + PP effect sizes organized in ascending order, and split into 1, 2, or 5 clusters (the three most relevant numbers). The bar colors indicate the classification under the LSH.



The top-right panel of Figure 3 confirms the general impression of Figure 2 that there is no step-like division between verbs based on the *ne* effect size that could be used to argue that *ne*-

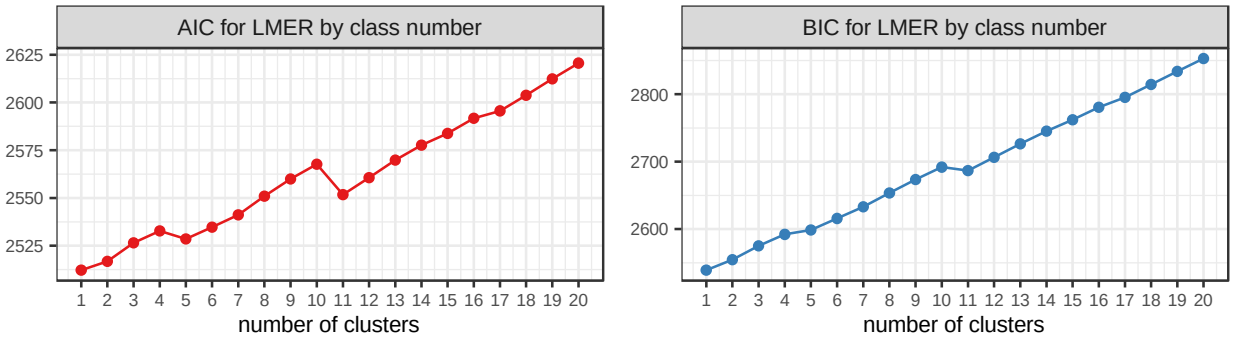
cliticization is a split-intransitivity diagnostic according to either of the two theories. The same conclusion emerges from exploring the two-class and five-class options in more detail: the classes are each a fairly extreme mix of lexical semantic verb types. It does not appear as though the classes straightforwardly map to either theory of split-intransitivity.

The second step of our analysis evaluates these descriptive impressions quantitatively. We constructed linear mixed effects models to predict acceptability based on the interaction of *ne* (ne/full-clause) 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 (particularly) relevant. Instead, the interpretation is a relative one: 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 volatile), 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). We will side-step this by reporting both metrics, and looking 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 with the uncorrected AIC.)

Figure 4 reports the AIC and BIC for each of the 20 models. Both the AIC and BIC yield lines that generally increase monotonically as the number of clusters increases. The overall monotonically increasing pattern suggests that the any increase in explanatory value gained with each additional cluster is outweighed by the penalty for the increase in model complexity. This overall pattern is what we would expect from a generally small gradient in effect sizes with no clear step-like breaks to indicate cluster boundaries. We take this as quantitative corroboration that the *ne* + PP effect size does not pattern like a split-intransitivity diagnostic.

Figure 4: The left panel reports the Akaike Information Criterion (corrected for small sample sizes) for linear mixed effects models using the full range of possible classes (1-20). The right panel reports Bayesian Information Criterion. For both metrics, models with lower scores are preferred over models with higher scores.



3.6 Conclusions for Experiment 1

In experiment 1, we tested *ne*-cliticization with an initial prepositional phrase. The results force us to conclude that *ne*-cliticization + PP does not appear to be a diagnostic of split-intransitivity in the classic sense (the *ne* conditions are in the acceptable range of the scale for all verbs) or based on *ne* effect sizes (the cluster analysis shows no benefit to dividing the verbs into two or more clusters).

4. Experiment 2: The ASC and *Ne*-cliticization diagnostics

In experiment 1 we found that *ne* cliticization is relatively acceptable with all the 20 verbs we tested. However, Saccon (1992) claim that a prepositional phrase can exceptionally license *ne*-cliticization with unergative verbs. Because our conditions in experiment 1 do include prepositional phrases, it is possible that our results are simply this exceptional *ne*-licensing. To rule out this possibility, and to potentially internally replicate our own results with a distinct sample of participants, we conducted a second experiment to test *ne*-cliticization, this time without a sentence initial prepositional phrase. For experiment 2 we also compare the pattern that arises with *ne* to the pattern that arises with the absolute small clause diagnostic as a baseline to demonstrate the behavior of a true split-intransitivity diagnostic.

4.1 Division into 4 Sub-experiments

We once again divided the 20 verbs into 4 sub-experiments, each containing 5 verbs, one from each lexical-semantic class for the *ne* cliticization diagnostic, and a distinct set of 5 verbs, one from each class for the ASC diagnostic (to avoid any interaction between the *ne* and ASC conditions). Table 3 lists the distribution of verbs in each sub-experiment.

Table 3: The set of verbs for *ne*-cliticization and ASC experiments used in each sub-experiment.

Sub-experiment 1		**Sub-experiment 2**		**Sub-experiment 3**		**Sub-experiment 4**	
ne	**ASC**	***ne***	**ASC**	***ne***	**ASC**	***ne***	**ASC**
Venire	Arrivare	Arrivare	Venire	Entrare	Cadere	Cadere	Entrare
Morire	Fiorire	Fiorire	Morire	Marcire	Nascere	Nascere	Marcire
Sopravvivere	Rimanere	Rimanere	Sopravvivere	Apparire	Bastare	Bastare	Apparire

Ballare	Nuotare	Nuotare	Ballare	Correre	Volare	Volare	Correre
Ridere	Telefonare	Telefonare	Ridere	Lavorare	Suonare	Suonare	Lavorare

4.2 Survey and Materials Construction

The surveys for sub-experiments 1-3 consisted of 3 anchor items in the instructions to illustrate the task, 6 unannounced practice items in the same order for each participant, followed by 20 target items (5 verbs x 2 *ne* conditions + 5 verbs x 2 ASC conditions) plus 20 filler items in a pseudorandom order, for a total of 3 + 46 items. Sub-experiment 4 consisted of 3 anchor items, 6 practice items, 20 target items and 26 filler items in a pseudorandom order, for a total of 3 + 52 items. In the remainder of this section, we describe the construction of the items in the surveys.

For the target items, we created 8 lexically matched pairs per verb based on these *ne* and ASC conditions (as in 4a-b and 5a-b) for a total of 640 items (20 verbs x 2 *ne* conditions x 8 tokens + 20 verbs x 2 ASC conditions x 8 tokens). We divided the target items for each sub-experiment into 8 lists using a Latin Square design such that participants did not see the same lexicalization either within or across verbs.

For sub-experiments 1- 3, the 20 filler items consisted of 8 items from an unrelated experiment about island effects, and 12 entirely unrelated items. Among the 20 filler items, we expect 14 to be unacceptable and 6 to be acceptable, such that combining them with the target items should yield an equal balance between 20 unacceptable and 20 acceptable items in the experiment. For sub-experiment 4, the 26 filler items consisted of 12 items from an unrelated experiment about island effects, the 12 unrelated filler items from sub-experiments 1, 2 and 3, to which we added 2 more filler items. We expect 23 of the filler items to be unacceptable and 9 to be acceptable, such that combining them with the target items should yield an equal balance of 23 acceptable and 23 unacceptable items in the experiment. We used the same 3 anchor items and 6 practice items as experiment 1. All materials are publicly available on our websites.

4.3 Participants and Procedure

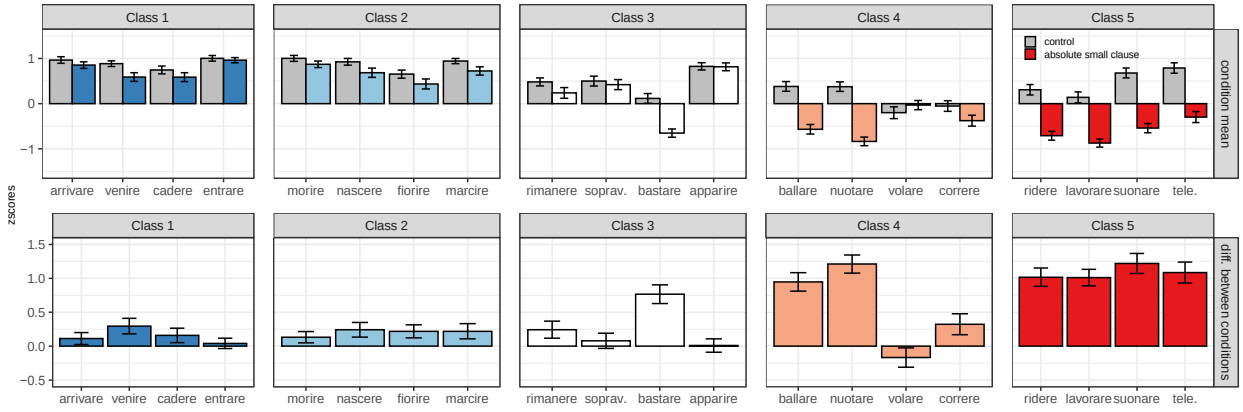
We recruited 45 self-reported native speakers of Italian who reside in Italy. Each participant was asked to complete all 4 sub-experiments (so that they rated all 20 verbs by the end), with each sub-experiment separated by at least one week. Participants were paid 2 Euros for each sub-experiment, and a 2 Euro bonus for completing all 4 sub-experiments (again for a rate of roughly 15 Euros per hour). All participants completed all 4 sub-experiments. We used the same task and presentation as experiment 1.

4.4 Results for the ASC diagnostic

We begin with an analysis of the ASC diagnostic because it serves as a baseline example for what the results of a split-intransitivity diagnostic will look like in our experiment. The top row of Figure 5 plots the means for the control (full adjunct clause) and ASC conditions for each individual verb, organized by lexico-semantic category for convenience, along with error bars that estimate the standard error for each mean. The bottom row of Figure 5 plots the difference between the control and the ASC conditions for each verb to highlight the effect size. The top row of Figure 5 shows

that there is variability in the absolute acceptability of the ASC conditions across verbs: for some verbs ASC is in the positive side of the scale, and for others it is in the negative side of the scale. This is in line with the logic of split-intransitivity 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. The bottom row of Figure 5 controls for this variability by subtracting the target condition from the control condition to reveal the size of the effect of ASC. There appears to be at least two types of verbs: those that show a relatively small difference between the control condition and ASC, and those that show a relatively large effect. To explore the number of classes quantitatively, we will again use a combination of hierarchical cluster analysis and linear mixed effects modeling.

Figure 5: The top row reports means (z-scores) for each condition (control and ASC) for each verb, organized by category. The bottom row reports the difference between the ASC condition and the control. Error bars represent estimated standard error.



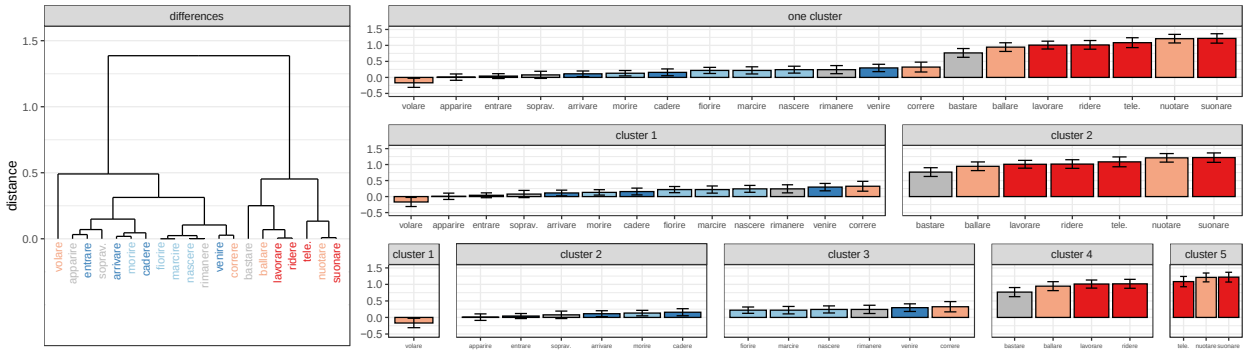
4.5 Hierarchical Cluster Analysis and Linear Mixed Effects Models for the ASC Diagnostic

We follow the same analysis procedure as experiment 1. The left panel of Figure 6 reports the full results of the clustering as a dendrogram (showing clusters 2 through 20 clusters). The right panels of Figure 6 re-plot the effect sizes (from Figure 5), organized in ascending order, and split into 1, 2, or 5 clusters (the three theoretically relevant clusters). We have retained the bar colors to indicate the by-hypothesis classification according to the LSH.

The right panels of Figure 6 confirm the impression from Figure 5 that there is a potential step-like split between verbs, particularly for the split into two classes (middle row). The division into lexical semantic categories is also relatively uniform in the two-class split (with just *volare* and *correre* as mismatches). This is what we might expect from a split-intransitivity diagnostic. The division into five classes is a bit less clear. *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, the by-verb details do not match the theoretical lexical semantic

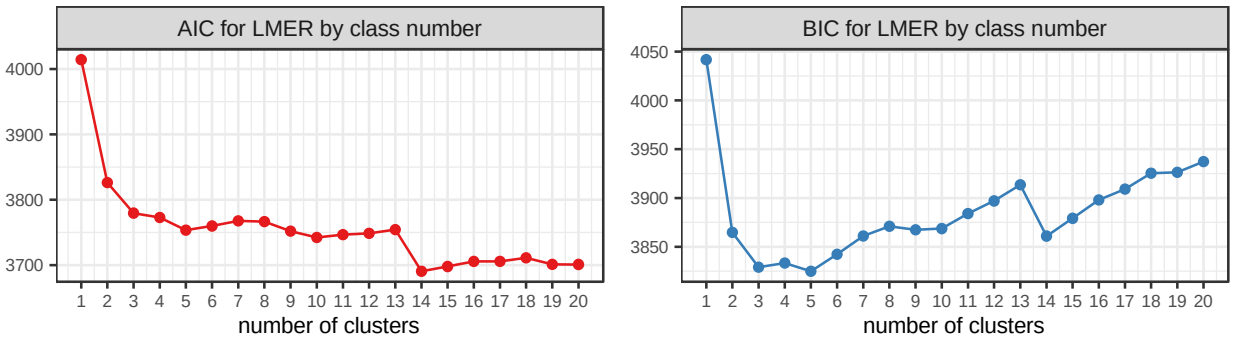
classes of the LSH. Class 2 contains 3 types of verbs; class 3 contains 4 types of verbs, class 4 contains 3 types of verbs, and class 5 contains 2 types of verbs.

Figure 6: The results of the clustering algorithm on the ASC effect sizes. The left panel reports the full results of the clustering as a dendrogram (showing clusters from 2 through 20). The right panels plot the *ne* effect sizes organized in ascending order, and split into 1, 2, or 5 clusters (the three most relevant clusters). The bar colors indicate the classification under the LSH.



To evaluate the cluster analysis quantitatively, we constructed linear mixed effects models and calculated AIC and BIC following the same procedure as experiment 1. The results are shown n Figure 7.

Figure 7: The left panel reports the Akaike Information Criterion (corrected for small sample sizes) for linear mixed effects models using the full range of possible classes (1-20). The right panel reports Bayesian Information Criterion. For both metrics, models with lower scores are preferred over models with higher scores.



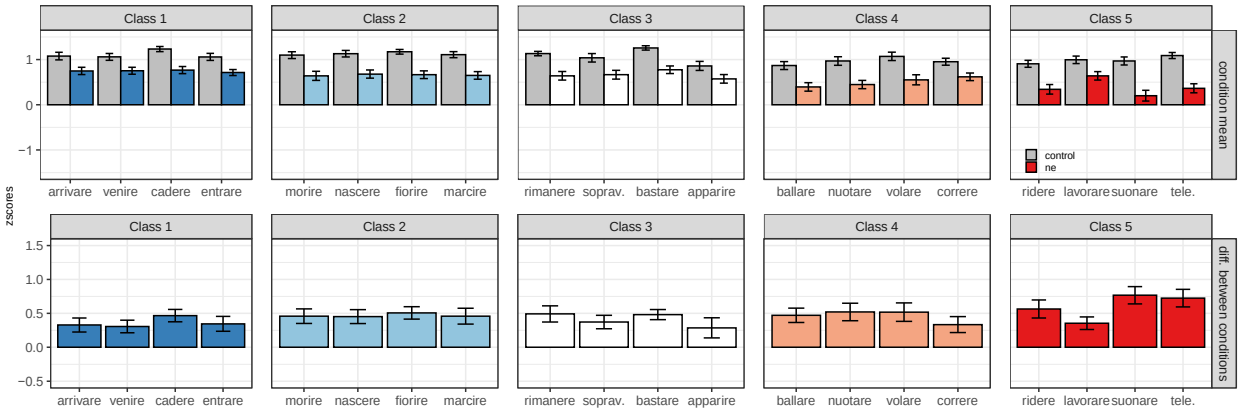
First, we see a large decrease in both AIC and BIC between the 1 cluster model and the 2 cluster model. This shows that splitting the verbs into at least 2 clusters dramatically increases the fit of the models (far beyond the complexity penalties). We take this to be a clear indicator of a split-intransitivity diagnostic – because there is a benefit to *splitting* the verbs into (two) classes. Next,

we see a smaller but still meaningful decrease between 2 clusters and 3 clusters. Recall that the third cluster (from the left panel dendrogram of Figure 3) only contains *volare*. This is because *volare* is an outlier - the direction of its effect is opposite to all of the others. Therefore we cannot interpret this as a statistical argument for a 3 class split-intransitivity theory, but rather as an argument to treat *volare* as an outlier (and perhaps explore why it behaves differently than the others in a follow-up study). Finally, we can ask whether adding additional clusters beyond *volare* increases the fit of the models as perhaps predicted by the LSH. For both AIC and BIC, we see roughly equivalent values after 3 clusters: for AIC, it extends up to 13 clusters; for BIC it only extends to 5 clusters before beginning to increase (suggesting worse fits). This range of equivalence suggests that there is no empirical benefit to further sub-dividing the classes after removing *volare*. We take this as evidence that the optimal division for the ASC is into two classes. That said, because 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), we consider this only suggestive evidence that the ASC is a binary diagnostic. But it is a finding that could merit further research.

4.6 Results for the Ne-cliticization diagnostic

With the ASC diagnostic as a baseline, we can now turn to the *ne*-cliticization diagnostic. The top row of Figure 8 plots the means for the control (*ne* absent) and *ne* conditions for each individual verb, organized by lexico-semantic category for convenience, along with error bars that estimate the standard error for each mean. The order of the verb classes reflects the order predicted by the lexical-semantic approach. The bottom row of Figure 8 plots the difference between the control and the *ne* conditions for each verb to highlight the effect size for each verb.

Figure 8: The top row reports means (z-scores) for each condition (control and *ne*) for each verb, organized by category. The bottom row reports the difference between *ne* and the control condition. Error bars represent estimated standard error.



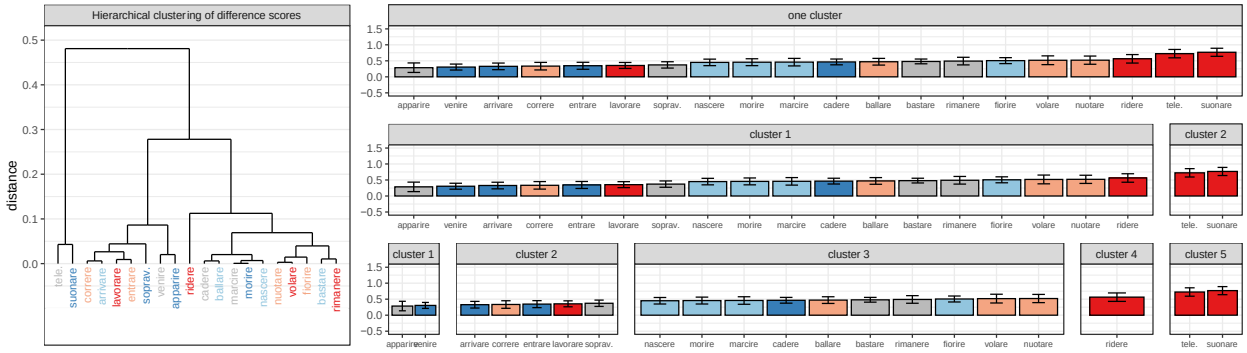
In the top row of figure 8 we see that, while there is some variation by verb in the ratings of both the control and *ne* conditions, both the control and *ne* conditions are consistently in the

acceptability range (positive side of the scale). This is evidence against the idea that *ne*-cliticization is categorically unavailable with unergative verbs, as proposed by both the UH and LSH. This suggests that the classic conception of *ne* as a split-intransitivity diagnostic that leads to unacceptability in one or more classes appears to be incorrect. In the bottom row of Figure 8, we again see some minor variation in effect size across verbs, roughly between 0.28 and 0.76 in the z-score scale. To evaluate whether this variability could be for a split-intransitivity pattern, we will again use a combination of hierarchical cluster analysis and linear mixed effects modeling.

4.7 Hierarchical Cluster Analysis and Linear Mixed Effects Models for the Ne-cliticization diagnostic

We follow the same analysis procedure discussed in section 3.5. The left panel of Figure 9 reports the full results of the cluster analysis as a dendrogram (showing 2 through 20 clusters). The right panels of Figure 9 re-plot the effect sizes (from Figure 8), organized in ascending order, and split into 1, 2, or 5 clusters (the three theoretically relevant clusters). We have retained the bar colors to indicate the by-hypothesis classification according to the LSH.

Figure 9: The results of the clustering algorithm on the *ne* effect sizes. The left panel reports the full results of the clustering as a dendrogram (showing clusters from 2 through 20). The right panels plot the *ne* effect sizes organized in ascending order, and split into 1, 2, or 5 clusters (the three most relevant clusters). The bar colors indicate the by-hypothesis classification under the LSH.

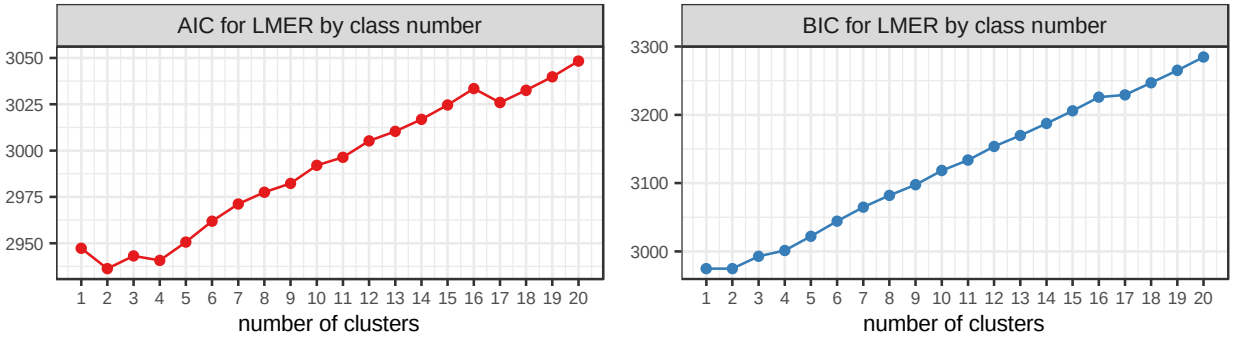


The top-right panel of Figure 9 is similar to what we saw in experiment 1. There is no step-like difference in effect sizes. The same conclusion emerges from exploring the two-class and five-class options in more detail. In the two-class option, we can see that the verbs are divided into two unbalanced sets: one set has 18 members and the other has two members. Though this is possible in principle according to the UH, this is not the by-hypothesis distribution of verbs to classes that we would expect if the *ne* effect size were a split-intransitivity diagnostic. Similarly, the five-class option groups verbs with different lexical-semantic properties together. Classes 1 through 3 each contain verbs that are by-hypothesis from distinct lexical semantic classes. Taken together, the *ne*

effect size does not appear to yield classes that straightforwardly map to either theory of split-intransitivity.

To evaluate the cluster analysis quantitatively, we constructed linear mixed effects models and calculated AIC and BIC following the same procedure as experiment 1. The results are shown in Figure 10.

Figure 10: The left panel reports the Akaike Information Criterion (corrected for small sample sizes) for linear mixed effects models using the full range of possible classes (1-20). The right panel reports Bayesian Information Criterion. For both metrics, models with lower scores are preferred over models with higher scores.



The BIC shows no difference between 1 and 2 clusters, and then generally increases monotonically as the number of clusters increases. This is in contrast to the patterns we saw with the ASC diagnostic, which revealed a large decrease between one and two clusters. This is what we would expect if the *ne* effect size does not pattern like a split-intransitivity diagnostic.

The AIC also generally increases monotonically, except for a small decrease between 1 and 2 clusters. The question is whether this could be evidence for split-intransitivity. We do not believe this decrease can be interpreted as split-intransitivity for three reasons. First, this decrease is at least an order of magnitude smaller than the one seen for the clear split-intransitivity effect with the ASC, 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 Figure 9, the second cluster would be just two verbs, 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 selected for our experiments.

4.8 Conclusions for Experiment 2

Our experimental design and analysis method appears to be able to distinguish diagnostics that are sensitive to split-intransitivity from those that are not. Our results suggest that ASC is likely a diagnostic for split-intransitivity. We see unacceptability for ASC conditions for some verbs. We also observe a step-like difference in effect sizes that suggests that there are two classes of verbs

with respect to the ASC diagnostic. Though it is not a primary research question for our study, hierarchical clustering and linear mixed effects models suggest that the best fitting model given our data has two classes, in line with the UH. (We also learned that *volare* behaves differently than all of the other verbs for reasons that remain unclear.) Our results also suggest that *ne*-cliticization is likely not a diagnostic for split-intransitivity. We do not see the predicted unacceptability of the *ne* conditions for some verbs (*ne* is always acceptable), nor do we see differences in the effect size of *ne* that would indicate split-intransitivity (as corroborated by hierarchical clustering and linear mixed effects models).

5. Conclusion

In this study, we tested two split-intransitivity diagnostics: *ne*-cliticization (with and without PPs) and absolute small clauses. We found that absolute small clauses show the empirical hallmarks of split-intransitivity 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 stress that this was not a primary goal of the experiment, so we note this finding only to motivate future research. In contrast, *ne*-cliticization (with or without PPs) does not show the hallmarks of split intransitivity. This suggests that *ne*-cliticization is not a diagnostic of split-intransitivity for participants recruited for our experiments. This in turn suggests that researchers interested in split-intransitivity should not consider *ne*-cliticization a robust diagnostic, at least when sentences are presented in isolation (with or without PPs). Finally, we note that we consider this a first experimental study to determine the behavior of *ne* in standalone sentences (as it is presented in the existing literature). Researchers interested in exploring whether the split might re-emerge with specific intonation or specific semantic contexts can use our results to formulate and test new hypotheses. To that end, the results of these experiments are freely available for exploration on the authors' websites.

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 UnaccusativityPuzzle: 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. 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.