LexVEN: A quick vocabulary test for proficiency in Venetan

Maria Ferin
University of Konstanz
maria.ferin@uni-konstanz.de

Henrik Gyllstad
University of Lund
henrik.gyllstad@englund.lu.se

Ilaria Venagli
University of Konstanz
ilaria.venagli@uni-konstanz.de

Angelica Zordan
University of Verona
angelica.zordan@univr.it

Tanja Kupisch
University of Konstanz / UiT The Arctic University of Norway
tanja.kupisch@uni-konstanz.de

Received: 30-09-2023
Accepted: 17-10-2023
Published: 25-10-2023

DOI: https://doi.org/10.5565/rev/isogloss.374
Abstract

While there is growing awareness that knowledge of regional minority languages must be protected, assessing the vitality of such varieties is challenging due to the lack of assessment materials. Our contribution documents the creation and validation of LexVEN, a bimodal yes/no vocabulary task, with the purpose of assessing the proficiency of users of Venetan with receptive and/or productive skills. The materials will be freely available. The vocabulary items were validated in a survey with 161 speakers of Italian, of which 136 came from the Veneto area. We have relied on Classical Test Theory (CTT) and Item Response Theory (IRT) in validating test items, and we demonstrate correlations with self-reported measures of proficiency and language use. Given the absence of any other assessment tools of Venetan for cross-validation, we further collected speech samples from a subset of the Venetan participants with active proficiency whose proficiency was rated by native speakers of Venetan. We discuss the procedures within a validation framework, concluding that LexVEN is reliable as a proficiency assessment tool of Venetan.

Keywords: Venetan, regional minorities, vocabulary, proficiency test, validation.

1. Introduction

Venetan is a regional language in the Veneto region in North Italy. The population of Veneto amounts to 4.8 million people. A decade ago, speaker estimates ranged around 4 million (Marcato 2005: 297, Berruto 2018: 503, Eberhard et al. 2022), but the number of users in the Veneto region today is probably lower with 35-40% of the population using only Italian or another language. Venetan is also spoken in other parts of Italy, especially in the neighboring regions of Trentino and Friuli-Venezia Giulia, as well as in countries to which Italians have emigrated, e.g., Croatia and Slovenia, Australia and Brazil. The International Commission of European Citizens (2022) provides an estimate of 7.8 million Venetan speakers worldwide. Despite decreasing speaker numbers, Venetan is one of the most vital and most widely spoken regional languages in Italy. Based on Istat (2012), Berruto (2018: 503) estimated that 70% of the Venetan population uses the dialect actively for their communication, mostly within families, among friends, with elderly speakers, but also younger people use Venetan. Unlike many other regional minority languages, Venetan does not seem to have suffered from a loss of prestige to the same extent as other regional languages (Cortelazzo 1999: 154, Santipolo & Tucciarone 2006). Nevertheless, there are some people for whom the use of Venetan is stigmatized compared to the standard language.¹

¹ There is no scientific literature to back this up, but a lot of anecdotal evidence. For example, a recent post in a FB group on the Venetan language started by "Speaking Venetan is not rude, […]", suggesting that there is a sufficient number of people who believe this would be the case.
In the Italian tradition, regional languages are called “dialects” to highlight their different socio-linguistic status from the standard language (Loporcaro 2013: 3). However, Venetan is not a dialect of Italian (Marcato 2005). Similar to other varieties and modern Romance languages, it has developed independently from Latin. It has a rich literary tradition, and speakers of Venetan tend to see it as a language rather than a dialect. Nevertheless, like most of the other Italo-Romance dialects, it is not mentioned in the law 482/99 that was established to protect and promote language minorities in Italy (van der Jeught 2016). The Ethnologue (Eberhard et al. 2022) classifies Venetan as a ‘stable’ language, which means—according to this source—that the language is not being sustained by formal institutions, but is still the norm in the home and community that all children learn and use the language. In our view, considering Venetan to be the “norm” is a too strong statement, although tendencies to use Venetan in the home are probably stronger than for other regional languages in Italy. The Atlas of the World’s Languages in Danger (WALS) (Moseley 2010) classifies Venetan as “vulnerable,” which seems more appropriate.

Venetan is (primarily) acquired in families in the Veneto region in which parents and grandparents are Venetan, thus in which there is some kind of “heritage”. Moreover, it is a minority language that is acquired in a naturalistic setting. For these reasons, it seems justified to include it under the umbrella term of “heritage languages”. One important difference to “migrant” heritage languages is that Venetan is autochthonous in the region where it is acquired and that nowadays most children do not seem to develop Venetan prior to Italian or simultaneously. Rather, studies with Venetan children at preschool and primary school ages suggest that while all children are fluent in Italian, their proficiency in Venetan is comparatively lower and increases with age (Kupisch & Klaschick 2017, Sanfelici & Roch 2021).

Italian and Venetan differ with respect to a number of lexical, grammatical and phonological properties, including subject use (e.g., Brandi & Cordin 1981, 1989; Poletto 1993) and many phonological properties, such as consonant degemination and (in many areas) [l]-vocalisation or deletion, lenition of the intervocalic stops /p, t, k/, and deletion of word-final vowels in specific phonological contexts (depending on variety) (e.g., Marcato 2004, Zamboni 1988, Loporcaro 2013). However, Venetan itself also varies. It represents a continuum of dialects, which divide into four subgroups (Zamboni 1988, Loporcaro 2013: 104): (i) Veneziano (the dialects of the islands and cities surrounding the city of Venice), (ii) central Venetan, (iii) western Venetan (e.g., Veronese) and (iv) alto Veneto, northern Venetan (e.g., Feltrino-Bellunese). These four areas have a rough correspondence to the seven administrative provinces. Central Venetan covers the largest area and includes Padovano, Rodigino and Vicentino, which are easily distinguishable for speakers with these origins.

While psycholinguistic research has focused on second language acquisition in schools and immigrant minority languages, bilectal acquisition and the acquisition of regional minority languages has only recently sparked interest in the psycholinguistic community (e.g., Leivada et al. 2017, Vangsnes et al. 2015, Garraffa et al. 2015, 2017).
Although Venetan is amongst the few regional languages that has relatively good visibility (Cardinaletti 2013, Bonifacio 2014, Sanfelici & Rocchi 2021), there are no standardized proficiency assessment tools, as the development of such materials is prioritized for languages that are actively taught in schools. This lack becomes a problem in research on bilectals and heritage bilingualism where speakers are often unfairly tested in the standard language although they have additional (often undefined and unattested) proficiency in a regional variety (Leivada et al. 2017, Kupisch, Arona, Besler, Ferin, Gyllstad & Venagli 2023). Self-ratings have long been argued to be unprecise (Ferré & Brysbaert 2017, Lemhöfer & Broersma 2012, Brysbaert 2013, MacIntyre et al. 1997, Izura et al. 2014). Thus, there is a great need for performance measures, but, ideally, they should be quick and easy (not requiring special technical equipment) and for free. Such tests have been increasingly developed during the past ten years for national languages, starting with Lemhöfer and Broersma (2012) for English, Brysbaert (2013) for French, Izura et al. (2014) for Spanish, and Amenta et al. (2021) for Italian. Our goal in this study is to create a diagnostic tool called LexVEN for Venetan proficiency as a basis for (psycho-)linguistic research on speakers of Venetan.

2. Vocabulary tests as proxies for general proficiency measures

It stands to reason that the best way to measure general proficiency in a given foreign language is to administer a suite of tests tapping into the various subskills, i.e., speaking, listening, reading and writing, that together make up the concerted ability to perform in that language. However, as is well-known, from a practical perspective, this is often not doable, as time and other resources are limited. Sitting tests of several subskills may take several hours at best. For this reason, it is common to rely on a proxy test, which is shorter but at the same time a good indicator of overall language skills. Potential proxy tests of general proficiency have been suggested, for example C-tests, gapped texts that require lexical and morphosyntactic skills. An alternative is to use a vocabulary test.

By convention, vocabulary knowledge is seen to consist of three dimensions: size (a.k.a. breadth), depth and fluency (Gyllstad 2013). Of these, the dimension of size is the most relevant for our purposes (though see caveat below). Vocabulary size refers to the number of words for which a language user has at least a basic meaning knowledge. Meara (1996) has argued that the dimension of size is the most basic one. Importantly, size has empirically been shown to be a robust indicator of general language proficiency. For a sample of 600-700 test-takers, Alderson (2005) observed correlations between the DIALANG yes/no vocabulary placement test and tests of writing ($r = .70$), reading ($r = .64$), listening ($r = .61$), and grammar ($r = .64$). Gyllstad (2007) reported links between vocabulary and reading ($r = .69$), whereas Staehr (2008) found correlations between vocabulary size and reading ($r = .83$), writing ($r = .73$) and
listening \( (r = .69) \). Uchihara and Clenton (2020) found a relation between the scores in a yes/no vocabulary test and L2 speaking abilities \( (r = .55) \). In light of Plonsky and Oswald’s (2014) recommendations for correlations as effect sizes (.25 = small, .40 = medium, and .60 = large), there is no doubt that receptive vocabulary knowledge can reliably predict general language proficiency.

As to the test format for LexVEN, we needed one that is practical, conventionally seen as effective, and reliable. One format commonly used in vocabulary tests is the yes/no format (Meara & Buxton 1987). Since the early 1990s, a large number of studies have investigated this format (e.g., Amenta et al. 2021, Beeckmans et al. 2001, Brysbaert 2013, Eyckmans 2004, Harrington & Carey 2009, Huibregtse et al. 2002, Meara & Miralpeix 2017, Mochida & Harrington 2006, Pellicer-Sánchez & Schmitt 2012, Salmela et al. 2021, Stubbe 2012, Zhang et al. 2020). The format entails presenting test-takers with a list of words and typically asking whether they know the meaning of each word; alternatively, whether the word exists or not (a difference we return to below). The test-taker then indicates whether their answer is yes or no for each word, though without having to supply any evidence of their stated knowledge. This can in theory lead to participants overstating the number of words they know. For this reason, yes/no tests typically include ‘pseudowords’, non-existing words that follow the phonotactics of the target language, and whose function it is to counter overestimation of test-taker scores. This is done by penalizing test-takers who extensively answer “yes” to a pseudoword (referred to as ‘false alarms’) during the scoring procedure.

There are many advantages of the yes/no test format; it is accompanied by cognitively undemanding instructions, which in turn means that a large number of items can be administered in one sitting (Hashimoto 2021). The yes/no test format is prominently featured in the DIALANG project, whose goal was to provide an online language assessment tool in 14 European languages, including Italian (Alderson 2005, Alderson & Huhta 2005). The DIALANG test suite begins with a placement test, which determines which difficulty level is appropriate for the continued testing. This placement test consists of 75 items, with 50 real verbs and 25 pseudoverbs in the infinitive form, sampled from dictionaries. Alderson (2005) argued that the use of verbs was an effective means of accessing a relatively large range of vocabulary in a given language with only few items.

A central question is what type of word knowledge is required when engaging with the items in a yes/no test. Schmitt (2010), based on work by Laufer and Goldstein (2004), has proposed a 2 x 2 matrix (Figure 1), which indicates what word knowledge is needed as a response to the information given in a test item. The context for this matrix was foreign language (L2) acquisition, but it can be mapped onto the situation of regional languages.
**Figure 1.** Matrix for classifying item formats for required word knowledge (adapted from Schmitt, 2010: 86)

<table>
<thead>
<tr>
<th>Word information provided</th>
<th>Form</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recognition</strong></td>
<td><strong>meaning recognition</strong> (select definition/L1 translation, etc.)</td>
<td><strong>meaning recall</strong> (supply definition/L1 translation, etc.)</td>
</tr>
<tr>
<td><strong>Form recognition</strong></td>
<td>(select L2 item)</td>
<td><strong>form recall</strong> (supply L2 item)</td>
</tr>
</tbody>
</table>

On the left, the kind of word information provided in an item prompt, ‘form’ or ‘meaning’, is indicated, and at the top, the two knowledge modalities, ‘recognition’ or ‘recall’, are shown. A standard multiple-choice test with a word form prompt and four alternatives to choose from would be a case of a meaning recognition format. In the case of the yes/no test, the word information given is a word form, and since no alternatives are given, it is a meaning recall format. The difference from a typical meaning recall format, however, is that no meaning knowledge is verified, as the test-taker simply states a yes or no. Technically, then, it is a form recognition format, but the test-taker is assumed to recall and activate a mapped meaning representation, which arguably blurs the boundary between recognition and recall. Irrespectively, the yes/no format is a receptive task, which is suitable for dialect speakers, because it provides the possibility to target lower proficiency level users of a language, which would be impossible with a productive task requiring a spoken or written recall answer.

In what follows, we describe in detail the purpose, development and initial validation of a vocabulary test called LexVEN. We will use an argument-based validation framework and rely both on Classical Test Theory (CTT) and Item Response Theory (IRT) approaches.

3. **Using a validation framework for test development**

3.1 **General considerations**

In the wake of a substantial increase in the number of vocabulary tests being developed over the last two to three decades, there has been a recent call for more rigorous test development and validation in the field of vocabulary assessment (Schmitt et al. 2020). The main critique has been directed towards the proliferation of tests without sufficient validation. More specifically, there is an identified need for more precise specification of the purpose of a test, what particular vocabulary knowledge facet is targeted, how
test scores should be interpreted, and the intended test-takers and potential educational context. Furthermore, better support for the test’s users by way of a users’ manual has been highlighted (What skills does it test? For which test takers? How do we interpret the scores?). The use of a validation framework has been suggested as a mainstay of the desired improvements.

A number of approaches exist. Kane (2012: 3) has summed up his argument-based approach to validation contribution, called a “interpretation/use argument” (IUA), as follows: “First, specify the proposed interpretations and uses of the scores in some detail (i.e., developing an interpretive argument). Second, evaluate the overall plausibility of the proposed interpretations and uses (i.e., compiling a validity argument)”. Specifically, the approach rests on the use of a flexible number of inferences that are combined with appropriate warrants. Kane (2013) emphasizes that it is the proposed score interpretations and uses that are validated, not the actual test or the test scores, and that validity is a matter of degree, which may change over time as the interpretations/uses develop and as new evidence accumulates.

For the present purposes, the IUA for scores on the LexVEN includes five inferences, presented in Table 1 together with the proposed warrants: domain definition, scoring, generalization, extrapolation, and decision. In the remainder of this paper, we will use this framework as a scaffolding tool. In a nutshell, argument-based approaches to validation start with a clear and explicitly stated purpose and provide structured and comprehensive evidence for justifiable interpretations.
Table 1. Inferences, their scope and proposed warrants for the validation of the LexVEN test

<table>
<thead>
<tr>
<th>Inference</th>
<th>Inference scope</th>
<th>Warrant</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Domain definition inference</td>
<td>The evidence revolves around item selection, whether it is representative and relevant of the domain in question. In addition, it deals with the effectiveness of the test’s task or item format.</td>
<td>A1. The LexVEN real word items are easily identified by proficient users of Venetan. A2. The yes/no task and item format is practical and efficient.</td>
</tr>
<tr>
<td>B. Scoring inference</td>
<td>Going from test-takers’ observed performance on a task (or response to some stimuli) to an observed score. It typically makes assumptions about the appropriateness and accuracy of the scoring criteria and the rules for combining scores, based on judgment and/or on statistical criteria.</td>
<td>B1. The yes/no test scoring procedure effectively takes false alarms into consideration and compensates for test-taker overestimation.</td>
</tr>
<tr>
<td>C. Generalization inference</td>
<td>It investigates how well test-takers can perform in some larger domain of tasks over some range of occasions and conditions of observation. It goes from the observed sample of performances to claims about expected performance in a universe of possible observations or to an estimated trait value that can be used to draw conclusions about the future performances.</td>
<td>C1. Scores on the LexVEN are reliable in terms of Rasch estimates of internal consistency as well as Cronbach’s alpha coefficients C2. Scores on the LexVEN have a good fit to an IRT model (e.g., Rasch) C3. Scores on the LexVEN differentiates between groups of test-takers with assumed different ability levels</td>
</tr>
<tr>
<td>D. Extrapolation inference</td>
<td>The extrapolation inference is the link between the universe score to a level of real-life ability or skill. It investigates whether test scores are indicative of test-takers’ real-world abilities in the target domain.</td>
<td>D1. Scores on the LexVEN correlate with rated spoken samples of Venetan D2. Scores on the LexVEN have a relation with self-reported Venetan proficiency and use scores</td>
</tr>
<tr>
<td>E. Decision inference</td>
<td>The decision inference takes us from score interpretation to score use. Test scores are used to make decisions which entail actions and consequences.</td>
<td>E1. LexVEN scores can be used for low-stakes situations where an estimate is needed of proficiency in Venetan.</td>
</tr>
</tbody>
</table>
4 Developing the LexVEN

4.1 The purpose of the test and its domain

The purpose of the LexVEN is to develop a proficiency test which can be used to locate users of Venetan on a continuum of proficiencies; the purpose is not to develop a bona fide vocabulary size test. A proper size test would need to be developed on the basis of systematic item selection based on sampling from frequency lists (see e.g., Gyllstad et al., 2021) and drawing on dictionary information, which is not available for Venetan.

The test consisted exclusively of verbs (real verbs and pseudoverbs). We focused on verbs for two reasons. First, for compatibility with other yes/no tasks that are freely available (section 2.1). Second, there is high lexical variability in the dialectal continuum within the Venetan-speaking area, and this was expected to be mostly found within nouns, while verbs were expected to be more homogeneous across provinces. The final list of verbs was created through three successive steps: (i) creation of a first list of items and selection based on pre-judgements (section 4.2), (ii) administration of a first test version and performing IRT analysis (section 4.3), and (iii) further test validation (section 6).

4.2 Item selection procedure

In the first phase, we compiled a list of 102 Venetan verbs and created 59 pseudoverbs. The real verbs were selected intuitively by creating a list of verbs with characteristic features of the Venetan dialect, e.g., intervocalic <j> or word-initial <s-c> (pronounced [stʃ]), and of which most had a low degree of similarity to existing Italian verbs. The pseudoverbs were either freely invented, such that they included the same characteristic features of Venetan, or were derived from existing noun or verb stems, e.g., tola (‘table’) > tolare. Similarity with existing Italian verbs was avoided, though not entirely eliminated.

4.2.1 Pre-judgements on the initial lists

One major goal for item selection was to achieve a balanced spread over a continuum of difficulty. Since no frequency dictionaries are available for Venetan, we needed an alternative strategy for a preliminary assessment of the items. Therefore, we had the verbs judged by ten Venetan speakers with at least one representative from each of the seven provinces: one judge from Belluno, Rovigo, Treviso, and Venezia; two judges from Padova, Verona and Vicenza.

2 The grapheme <j> conventionally refers to the palatal glide [j], which etymologically derives from the palatalization of the Latin <LJ> cluster. It is realized as the glide [j] in many varieties, as in svejarse ‘wake up’, tajare ‘cut’, but in some varieties of Venetan (e.g., Venetian) also as [ʤ].
For the real verbs, the judges were asked whether they knew the meaning of the verb (yes/no) and its perceived level of difficulty on a 4-point scale (1 = molto facile ‘very easy’, 2 = facile ‘easy’, 3 = difficile ‘difficult’, 4 = molto difficile ‘very difficult’). Some judges asked whether perceived difficulty may coincide with frequency, which we confirmed. Real verbs that were unknown to more than one judge were excluded, as they could potentially compromise the validity of the test as covering the entire Veneto region. This left us with a final sample of 71 real verbs. For the pseudoverbs, the judges were asked to rate the level of plausibility (1 = altamente implausibile ‘highly implausible’, 2 = poco plausible ‘not very plausible’, 3 = plausible ‘plausible’, 4 = altamente plausible ‘highly plausible’). Of the remaining pseudoverbs a subset of 35 verbs was selected that was equally spread over all plausibility ranges. These initial judgements from native speakers (henceforth pre-judgements) were used in later analyses as an indirect measure of frequency or difficulty, to remedy for the lack of documented frequencies for Venetan.

4.3 Administration of a first test version of LexVEN and IRT analysis

In the second phase, we created an initial yes/no vocabulary test with the 71 real verbs and 35 pseudoverbs that were selected in the first phase. This version was administered to a large group of participants (see below). We then conducted an IRT analysis of the collected data in order to select the list of verbs for the final version of the test.

4.3.1 Participants

Data from a total of 163 participants from three groups were collected: Veneto, North (Lombardy) and South. These three groups were selected because they were expected to vary in their degree of lexical knowledge of Venetan, with speakers from the Veneto having the best knowledge of the given verbs, while speakers from the region of Lombardy (geographically and linguistically closer to Veneto) would recognize some verbs but not all, and speakers from the South would recognize the fewest verbs. An overview of the speakers is provided in Table 2. The Venetan speakers (age range: 18–78, M = 37.3, SD = 15.7) had different provinces of origin, including Vicenza (n = 48), Padova (n=32), Treviso (n=17), Venezia (n=15), Verona (n = 14), Rovigo (n = 6) and Belluno (n = 2); one participant did not answer. The education level of the Venetan speakers included 67 speakers (49%) with a university degree, 45 (33%) with a high school diploma and 23 (17%) a lower degree.
Table 2. Overview of Participants

<table>
<thead>
<tr>
<th></th>
<th>Veneto</th>
<th>North</th>
<th>South</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>136</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>Age</td>
<td>18–78</td>
<td>20-27</td>
<td>20-55</td>
</tr>
<tr>
<td>M</td>
<td>37.3</td>
<td>34.8</td>
<td>29.6</td>
</tr>
<tr>
<td>SD</td>
<td>15.7</td>
<td>13.6</td>
<td>7.8</td>
</tr>
<tr>
<td>Gender</td>
<td>98 F, 37 M</td>
<td>5 F, 5 M, 1 unspecified</td>
<td>9 F, 6 M</td>
</tr>
</tbody>
</table>

4.3.2 Materials and procedures

The task was administered through the online software SoSci survey (Leiner 2014). In addition to providing the written form of each item, a spoken form was also recorded by a native speaker of Venetan; this was done since Venetan is primarily spoken and because various orthographic conventions exist. The task instruction to the participants was that they would read and hear existing and non-existing verbs and that they would be asked to judge whether they knew the meaning of each verb by pressing a “yes” or a “no” key. The instruction emphasized not to focus on the orthography, and reminded participants that Venetan is subject to regional variation and that an item was to be indicated as “known” if they thought they knew the meaning of a verb, even if it sounded different in their own variety. This created a small risk of overaccepting nonce words (‘false alarms’), which however, as pointed out below, did not turn out to be an issue. The order of the test items was randomised. The online instrument also contained a questionnaire with biographical data including self-rated proficiency, and amount of language use within and outside the home.

Finally, for the assessment of the vocabulary data, we also collected short picture-based narratives to be subsequently judged by native Venetan speakers. This part of the survey was optional. The participants were asked to describe a series of pictures taken from the MAIN narrative task (Gagarina & Bohnacker 2022). The sequences were between 18 and 78 seconds long, and we extracted snippets of 9-12 seconds duration. Fifty-six out of 136 Venetan participants completed this part of the task.

4.3.3 Initial item analysis

The goal of the initial item analysis was to reduce the number of real verbs from 71 to 50 and the pseudoverbs from 35 to 25. We first excluded participants with a high number (above 50%) of false alarms. This resulted in the exclusion of 2 participants, leaving us with 161 participants whose responses were analyzed. The response data were then assessed considering the following criteria: i) point biserial correlation values, ii) difficulty values (beta values) of items given by the IRT analysis, iii) person-item parameter distribution, iii) item discrimination power (given by infit- and outfit-
statistics of each item), iv) variety-specificity of items, as well as v) their orthographic Levenshtein distance from the equivalent Italian verb.

We computed point biserial correlations between participants’ score on a single item (a dichotomous variable, either 1 or 0) and their total score on the remaining items.

Table 3. Point biserial correlation values for removed words and nonwords

<table>
<thead>
<tr>
<th>Item</th>
<th>Translation</th>
<th>Point biserial correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imbriagarse</td>
<td>It. <em>ubriacarsi</em> ‘get drunk’</td>
<td>.17</td>
</tr>
<tr>
<td>Imbrojare</td>
<td>It. <em>imbrogliare</em> ‘to cheat’</td>
<td>.14</td>
</tr>
<tr>
<td>Lagnarse</td>
<td>It. <em>lamentarsi</em> ‘complain’</td>
<td>.08</td>
</tr>
<tr>
<td>Rimambirse</td>
<td>It. <em>rimambirsi</em> ‘to goof off’</td>
<td>.03</td>
</tr>
<tr>
<td>Stufarse</td>
<td>It. <em>stufarsi</em> ‘get fed up with something’</td>
<td>.14</td>
</tr>
<tr>
<td>Pironare</td>
<td>(pseudoword)</td>
<td>.16</td>
</tr>
</tbody>
</table>

Following Salmela et al. (2021: 810), we used a cut-off point of .20. This led to the removal of 5 real words and 1 nonword (Table 3). We attributed the poor correlation of these real words with the participants’ overall performance to their high similarity with authentic Italian verbs, leading individuals to answer correctly regardless of their background (here, Venetan or non-Venetan background) (see also Salmela et al. 2021: 810).

Next, we performed IRT analyses using the RM function from the eRm package in R (Mair et al. 2021), complemented by analyses using Winsteps (Linacre 2021). For each item, we obtained beta values indicating the probability of getting the item correct as a function of participants’ proficiency on the task. We started by reducing the number of items with the same beta values, i.e., items with similar item difficulty across proficiency levels. For example, there were five real words with a beta value of 0.55: *ciacolare* (It. *chiacchierare* ‘chat’), *copare* (It. *uccidere* ‘kill’), *pareciare* (It. *preparare* ‘prepare’), *rampegarse* (It. *arrampicarsi* ‘climb’), and *scorlare* (It. *scuotere* ‘shake’). We assessed accuracy scores across the provinces of Veneto, in order to identify potential variety-specific items. The accuracy scores for the verb *scorlare* differed across provinces, with participants from Venezia and Treviso displaying poor recognition of this item. Since this finding may indicate variety-specificity, the item was removed. *Rampegarse* was removed because participants from the South easily recognized it, indicating that the item is too similar to its translation equivalent in southern dialects; *maridarse* (It. *sposarsi* or *maritarsi* ‘get married’) and *menare* (It. *menare* or *picchiare* ‘hit someone’), were removed for their high similarity to Italian.

We followed the same steps for pseudoverbs. When removing items with similar beta values, we considered the nonword orthographic similarity to existing Italian verbs, as well as participants’ over-acceptance of the item. For example, the pseudoverb *stramasare* was over-accepted, probably because of its high similarity with the Italian verb *stramazzare* ‘collapse’.
In total, we removed 21 real words, and 10 pseudoverbs (see below for the complete list of words and pseudoverbs, including Italian and English translations).

4.3.4 Rasch analysis of selected real word items
As the use of Rasch comes with the assumption of unidimensionality, i.e., the state where only one meaningful latent attribute exists in the data, and the fact that previous studies (e.g., Anonymous et al 2023a) involving both real and pseudoverbs found two distinct dimensions, the subsequent analyses were carried out on the real words separately. A Rasch analysis of the final 50 real word items selected through the previous analyses resulted in the values summarized in Tables 4 and 5 below. An item measure order is provided in Appendix A and a Wright map is provided in Appendix B. The summarized person and item values in Table 4 indicate that the item and person infit and outfit statistics were generally within the expected range (MNSQ 0.5-1.5) (Green 2013). Three items had infit values higher than 1.5 (R27: *misiare*, It. *mescolare* ‘mix’, R32: *ramenarse*, It. *rotolarsi*, ‘roll (oneself) around’, and R51: *smorsare*, It. *spegnere* ‘turn/switch (something) off’ and one item lower (R12: *fracare*, It. *premere* ‘push/press’). One person had an infit value above 1.5 (919 North).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Measured</th>
<th>Total</th>
<th>Logit</th>
<th>SE</th>
<th>Infit MNSQ</th>
<th>Infit ZSTD</th>
<th>Outfit MNSQ</th>
<th>Outfit ZSTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons</td>
<td>M</td>
<td>44.60</td>
<td>4.04</td>
<td>1.19</td>
<td>0.99</td>
<td>0.15</td>
<td>1.08</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>10.90</td>
<td>2.24</td>
<td>0.61</td>
<td>0.16</td>
<td>0.67</td>
<td>1.18</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>4.00</td>
<td>-3.02</td>
<td>0.55</td>
<td>0.64</td>
<td>-2.48</td>
<td>0.11</td>
<td>-1.93</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>50.00</td>
<td>5.81</td>
<td>1.84</td>
<td>1.60*</td>
<td>3.78</td>
<td>5.97</td>
<td>2.75</td>
</tr>
<tr>
<td>Items</td>
<td>M</td>
<td>143.50</td>
<td>0.00</td>
<td>0.40</td>
<td>1.00</td>
<td>-0.09</td>
<td>1.01</td>
<td>-0.19</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>8.20</td>
<td>1.23</td>
<td>0.06</td>
<td>0.36</td>
<td>1.48</td>
<td>1.46</td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>116.00</td>
<td>-2.64</td>
<td>0.59</td>
<td>0.43</td>
<td>-2.97</td>
<td>0.15</td>
<td>-2.75</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>157.00</td>
<td>3.03</td>
<td>0.27</td>
<td>2.12</td>
<td>3.56</td>
<td>9.90</td>
<td>3.84</td>
</tr>
</tbody>
</table>

Note. *There were 71 “maximum measure” values.

In Table 5, the person separation index reveals how well the items separate persons measured, whereas the item separation index shows how well a tested sample of people is able to separate the items. Item separation indices of 3 or greater are desirable (Linacre 2002). For the person separation index, 1.5 is conventionally acceptable, 2.0 is good, and 3.0 is excellent.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Separation</th>
<th>Rasch reliability</th>
<th>Cronbach’s alpha reliability</th>
<th>Mean Item-Total correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>1.34</td>
<td>.64</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Item</td>
<td>2.63</td>
<td>.87</td>
<td>.98</td>
<td>.70</td>
</tr>
</tbody>
</table>
In the current study the item separation of 2.63 falls somewhat short of the desired threshold of 3, and the person separation of 1.34 is somewhat low. The Cronbach’s alpha coefficient was observed at .98, which together with the item reliability of .87 suggests that the replicability of the item performance on a similar test population is sufficiently high. The mean item-total correlation was .70 (all items but one reached > .30, range .28 - .89), and in this regard the 50 items included in the final version of LexVEN have a good discriminative power. The Wright map in Appendix B indicates that the final version set of items matches the abilities well for the mid and lower segment, but that there are few items that can match the persons with the very highest abilities (Logits of 4 to 5). As many as 71 participants scored 50 out of 50.

5 Proficiency rating task as an objective proficiency measure of LexVEN takers

The final step in our validation process for LexVEN was based on an analysis of participants’ performance in the task based on biographical measures and linguistic proficiency (section 6). Given the lack of independent measures to assess Venetan proficiency, we used samples of recorded speech (section 5.1) in a rating task as an independent measure of proficiency in the dialect, which was then used for the final validation analysis.

5.1 Materials and procedure

The speech samples existed for a subset of 56 LexVEN takers. They were presented to 36 raters, all speakers of Venetan who had spoken Venetan since childhood (13 from Padova, 4 from Treviso, 6 from Verona, 6 from Venezia, 7 from Vicenza). The rating task also included control samples in standard Italian from 8 speakers of Italian of a northern (Lombard) or southern (Sicilian) variety as well as 8 speakers of Italian from the Veneto region. The samples were randomized and presented in two separate lists, each containing 28 samples plus the 16 controls. The raters were asked to judge each sample, imagining they were teachers of Venetan who had to assess how well someone already speaks Venetan. They were given a scale from 1 to 6, where 1 meant “someone who speaks only Italian”, and 6 “someone ‘native’, who has spoken the language for a long time and can easily express everything”. The task took about 15 minutes to complete.

Data from one rater were excluded because of very low self-reported use and proficiency. The origin of the raters had no effect on the ratings, and will not be discussed any further.
5.2 Intra-rater reliability and mean ratings

Inter-rater reliability, measured as Intraclass Correlation Coefficients (ICC), showed values above .80 (list 1: ICC = .87 [CI .82-.92]; list 2: ICC = .84 [CI .78-.89]), which indicates excellent agreement beyond chance. Figure 2 shows the mean ratings for each participant. All control groups (who spoke standard Italian) received very low scores, indicating that raters consistently gave a rating of 1 or 2 when they heard a sentence in Italian, even if the speaker had a Venetan accent (IT-Veneto: $M = 1.2$, SD = 0.1; IT-North: $M = 1.1$, SD = 0.1; IT-South: $M = 1.0$, SD = 0.03). The target group received on average high ratings ($M = 4.8$, SD = 1.1), indicating that the raters perceived a clearcut difference between the two linguistic codes (Venetan and Italian). The mean rating was used as a measure of objective proficiency for subsequent analysis.

Figure 2. Mean ratings in the proficiency judgment task per participant group

6. Analysing the performance of assumed ability groups

Based on the final selection of 50 real words and 25 pseudoverbs, we next investigated the scores from the three assumed ability groups (Table 6 below). This approach is what Bachman (2004: 290) refers to as a non-equivalent group design, i.e., running “the same test to several groups of individuals who are known, or who are believed, on the basis of some prior criterion, to differ in the ability to be assessed”. First, we analysed the effect of item and participant variables on performance in the LexVEN
task, considering participants’ individual responses (section 6.1). Subsequently, we analysed participants’ performance based on calculated test scores (section 6.2).

### 6.1 The effect of item and participant-related variables on LexVEN performance

The analyses on the effect of item and participant variables were based on data for both the real words and the pseudoverbs. As an independent item parameter, we considered the **pre-judgements** on the words as a facility (or familiarity) parameter for the verbs\(^4\). Lower scores indicated easier items both for words and nonwords (see section 4.2.1).

For participants, we considered three additional proficiency and biographical measures. These were collected for Venetan speakers only, since the other groups have no proficiency in Venetan. In particular, we collected **self-assessed proficiency** in Venetan. Proficiency was self-assessed on a scale from 1 (min) to 6 (max) for comprehension and production, from which we calculated a mean proficiency score. The mean proficiency was 4.6 (SD = 1.49). We also calculated two use and exposure scores: a **UseHome** score for use of Venetan in the family, and a **UseOutside** score for exposure outside of the family. The scores were calculated from questions, such as “which language(s) do you speak with your mother/father/friends…?” on a scale of 1 (only Italian) to 5 (only Venetan) (Besler et al. 2023b). The two Use scores were calculated as an average of the answered questions in both categories (see Anonymous et al. 2023b). Finally, we used mean ratings in the proficiency judgment task (Figure 2) as a measure of **objective proficiency** for Venetan speakers. All numerical predictors were scaled.

For the statistical analysis, we ran generalized linear mixed effect models in R (version 4.2.1). The **glmer** function (**lme4** package, Bates et al. 2015) was used to run the models, while the **emmeans** and **emtrends** functions (**emmeans** package, Lenth 2023) were used for the post-hoc analysis of interaction terms. The model selection was performed with the functions **anova** and **drop1** from the **stats** package (R Core Team 2022). In the models, **Accuracy** was used as a response variable assuming a binomial distribution: correct answers were scored 1, and incorrect answers were scored 0. For real words, 1 represented a correctly accepted item, while for nonwords 1 represented a correctly rejected item.

---

\(^4\) For real verbs, 1 was ‘easiest’ and 4 was ‘hardest’. For pseudoverbs, 1 was ‘least plausible’ and 4 was ‘most plausible’. We reasoned that a highly implausible pseudoverb would be the easiest to reject, thus equating the two scales. Pre-judgments were significantly correlated with the items’ beta values (indicators of item difficulty) outputted by the RM Rasch model. The Spearman rank correlation was -0.52 (p < .001) for real verbs and -0.66 (p < .001) for nonverbs.
6.1.1 Model 1: Predictive ability of pre-judgments on LexVEN (raw) accuracy scores

The first model investigated the predictive ability of the pre-judgments on accuracy in the three groups. The model formula included a three-way interaction between Pre-Judgments, Condition (Word vs. Nonword) and Group (North, South and Venetan), as well as random intercepts for Participants and Items. To keep the random effect structure maximal (Barr et al. 2013), the model included a random slope for Condition over Participants. The model yielded a significant effect of Condition ($\chi^2 = 5.87$, df = 1, $p = .015$) and Group ($\chi^2 = 7.22$, df = 2, $p = .027$). Pre-judgment scores alone did not significantly predict accuracy ($\chi^2 = 0.06$, df = 1, $p = 0.807$). The three-way interaction was also significant ($\chi^2 = 8.64$, df = 2, $p = .013$). The model output is illustrated in Figure 3.

Figure 3. The predictive ability of the pre-judgments on accuracy in the three groups

The Venetan group was most accurate on real verbs, with an estimated marginal mean of accuracy of 98%. The Northern group had a lower accuracy (M = 65%) and it was even lower for the Southern controls (M = 31%). The difference between the groups was significant (South vs. Venetan: $\beta = -5.53$, SE = 0.45, $z = -12.41$, $p < .001$; North vs. Venetan: $\beta = -3.98$, SE = 0.5, $z = -7.99$, $p < .001$; North vs. South: $\beta = 1.55$, SE = 0.6, $z = 2.57$, $p = .031$). As for nonwords, predicted accuracy was almost at ceiling for all groups: North = 90%, South = 96% and Venetan = 97%. Significant differences in the nonword condition emerged between the North and the Venetan group ($\beta = -1.36$, SE = 0.54, $z = -2.51$, $p = .036$), and between the North and South group ($\beta = -1.88$, SE
= 0.78, \( z = -2.43, p = .045 \), but not between the Venetan and South group (\( \beta = 0.52, SE = 0.6, z = 0.86, p = 1 \)).

Significant effects of Pre-Judgments on real word accuracy were observed for all groups, such that easier verbs were accepted more frequently (North: \( \beta = -1.18, SE = 0.26, z = -4.55, p < .001 \); Venetan: \( \beta = -1.19, SE = 0.24, z = -4.99, p < .001 \); South: \( \beta = -0.49, SE = 0.24, z = -2.07, p = .038 \)). Interestingly, the effect of Pre-Judgments on nonword accuracy was significant for the Venetan group only (\( \beta = -0.62, SE = 0.17, z = -3.59, p < .001 \)), showing that with increasing plausibility (i.e., increasing difficulty) of the nonword, Venetan speakers are more inclined to accept them.5

6.1.2 Model 2: Predictive ability of (self-assessed) proficiency, use and pre-judgment scores on LexVEN (raw) accuracy scores

To further investigate whether performance in the lexical task by Venetan participants was predicted by their proficiency and use, we ran a second model. The model was run on the subset of Venetan participants (\( n = 56 \)) for whom we had a measure of objective proficiency (the rating task). In the model, we looked at the predictive ability of (i) Pre-Judgments, (ii) dialect Use in the home environment, (iii) Self-Assessed Proficiency and (iv) Rated Proficiency. We performed a backward stepwise model selection to determine the best-fit model, which included two interaction terms: one between pre-judgments and condition, and one between use and condition. Rated Proficiency was included as a single main effect, while Self-Assessed Proficiency was dropped. Further, the model included random intercepts for Participants and Items, and allowed for a random slope of Condition on Participants.

The model yielded a significant effect of Pre-Judgments (\( \chi^2 = 8.76, df = 1, p = .003 \)), and a significant positive effect of Rated Proficiency (\( \chi^2 = 4.59, df = 1, p = .032 \)), as shown in Figure 4a. The two-way interaction between Condition and Use was also significant (\( \chi^2 = 5.36, df = 1, p = .021 \)). Condition (\( \chi^2 = 1.99, df = 1, p = .158 \)) and Use (\( \chi^2 = 0.53, df = 1, p = .467 \)) alone were not significant, nor was the two-way interaction between Pre-Judgments and Condition (\( \chi^2 = 2.37, df = 1, p = .123 \)).

---

5 A model with the full dataset of Venetan participants only was run to explore whether this effect was driven by participants’ (self-assessed) proficiency. The model output and the post-hoc analysis of the interaction between Condition and participants’ (self-assessed) Proficiency indicated that this effect is not driven by their Proficiency. For reasons of space, the model is not reported in the paper, but is available in the supplementary materials (see link to OSF files at the end of the paper).
**Figure 4.** The predictive ability of rated proficiency, pre-judgements and use at home on accuracy in the Venetan group.

![Graphs showing predictive ability](image)

The post-hoc analysis on the two-way interaction between Pre-Judgment and Condition confirmed previous results: The effect of Pre-Judgments was significant and negative for both real words ($\beta = -1.24, SE = 0.32, z = -3.93, p < .001$) and nonwords ($\beta = -0.65, SE = 0.22, z = -2.96, p = .003$), as illustrated in Figure 4b (Recall that low scores means “easy items”).

Finally, we tested the two-way interaction between Use and Condition. Use significantly predicts accuracy in the real word condition ($\beta = 0.83, SE = 0.3, z = 2.79, p = .005$), but not in the nonword condition ($\beta = -0.17, SE = 0.24, z = -0.73, p = .467$), as illustrated in Figure 4c.

### 6.2 LexVEN score: descriptive and inferential statistics

We calculated the final LexVEN score with a correction for false alarm responses, including the final 50 real verbs and 25 pseudoverbs. Following Brysbaert (2013) and Amenta et al. (2021), we calculated the test score as in formula (1), which penalized guessing behavior (Izura et al. 2014). The maximum score of 50 can only be obtained by saying ‘yes’ to all the words and to none of the non-words.

(1) LexVEN Score = $N_{\text{yes to words}} - 2 \times N_{\text{yes to nonwords}}$

Descriptive statistics for the ability groups are shown in Table 6 and is visualized in Figure 5.
Table 6. Descriptive statistics for the participating ability groups

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>CI Lower</th>
<th>CI Upper</th>
<th>Mean HR</th>
<th>Mean FA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Veneto</td>
<td>135</td>
<td>44.86</td>
<td>4.02</td>
<td>44.18</td>
<td>48.20</td>
<td>0.97</td>
<td>0.08</td>
</tr>
<tr>
<td>North</td>
<td>11</td>
<td>27.09</td>
<td>12.84</td>
<td>19.50</td>
<td>32.34</td>
<td>0.67</td>
<td>0.13</td>
</tr>
<tr>
<td>South</td>
<td>15</td>
<td>14.73</td>
<td>8.13</td>
<td>10.75</td>
<td>18.88</td>
<td>0.32</td>
<td>0.04</td>
</tr>
<tr>
<td>Total</td>
<td>161</td>
<td>40.84</td>
<td>11.39</td>
<td>39.06</td>
<td>42.61</td>
<td>0.89</td>
<td>0.08</td>
</tr>
</tbody>
</table>

As expected, participants from the Southern group had the lowest scores, and speakers from Veneto scored the highest. Participants from the Northern group scored a mean just below twice as high as the South group. Results from a one-way ANOVA indicated significant differences, $F(2, 160) = 182.3, p < 0.001$, and a Games-Howell post hoc test showed that all three groups were significantly different from each other, at $p < 0.01$. Table 6 also shows the Mean Hit Rate (HR) and the Mean False Alarm Rate (FA) for the groups, indicating the very high HR and low FA for the Veneto group participants.

Figure 5. Box-and-whiskers graph for the participating ability groups’ performances

6.2.1 Correlations between LexVEN, proficiency and background measures

We ran correlation analyses between LexVEN scores and proficiency and background measures, described in section 6.1, for the group of Venetan speakers only. We compared the LexVEN score, corrected for false alarms (third column), with the raw accuracy in the task, for real verbs only (first column) and on real and pseudoverbs
together (second column). As can be seen in Table 7, the correlations pointed to a positive relation between the rated spoken skill in Venetan and the performance on the LexVEN test, confirming the positive effect found in the model. The correlations were relatively similar across the three score calculations. Self-assessed proficiency also correlated significantly with the three scores, although less strongly. On the whole, objective proficiency comes across as the strongest predictor.

**Table 7.** Correlations (Spearman’ Rho) between speech sample rating scores and various LexVEN scores for a subset of test-takers (n = 56).

<table>
<thead>
<tr>
<th></th>
<th>Real words correct (hits) in LexVEN</th>
<th>Real words and pseudoverbs correct in LexVEN</th>
<th>LexVEN score corrected for false alarms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective proficiency (rating task)</td>
<td>.443**</td>
<td>.484**</td>
<td>.465**</td>
</tr>
<tr>
<td>Self-assessed proficiency</td>
<td>.386**</td>
<td>.384**</td>
<td>.340*</td>
</tr>
<tr>
<td>Use in the family</td>
<td>.454**</td>
<td>.339*</td>
<td>.271*</td>
</tr>
<tr>
<td>Use outside the family</td>
<td>.290*</td>
<td>.307*</td>
<td>.256 (n.s.)</td>
</tr>
</tbody>
</table>

Note. *Correlation significant at < .05. **Correlation significant at < .01. (n.s.) Correlation not significant. Based on a sample of n = 56

7. Discussion

7.1 LexVEN as a proficiency test

We have developed and provided initial validation of LexVEN, an online vocabulary assessment tool for Venetan. The test was modeled after the DIALANG placement test, but we adapted the instructions and presented the stimuli in both oral and written mode because familiarity with dialect in the written mode cannot be taken for granted. Instead of comparing L1 and L2 speakers, as is commonly done in the development of proficiency tests, we compared bilectals in the target region (Veneto) with Italians from other regions in the North (who are potentially familiar with a related variety) and with Italians from the South (who are unlikely to be familiar with any variety closely related to Venetan). The LexVEN scores were highest amongst the Venetans, lower amongst speakers of non-Venetan Northeners, and lowest amongst speakers from the South (Figure 5).

For the final set of items in the LexVEN, we ran correlations (using the corrected LexVEN score) and inferential statistics models (using raw, binary scores) to assess the predictive ability of factors that were shown to relate to vocabulary scores in previous studies, specifically language use, self-rated proficiency and rated
(objective) proficiency. We have added rated proficiency because there are no other assessment tools we could have compared the results to.

The correlations with use and (rated/objective) proficiency were restricted to the subset of participants who used the language actively and could provide a speech sample to be rated. The results showed the strongest correlations between the LexVEN and proficiency, especially rated proficiency, confirming that the test taps into this dimension. Although language use also correlated with the corrected LexVEN scores, this correlation was weaker (Table 7). This is expected since being proficient in a language does not necessarily imply that one uses the language. The statistical analysis on the LexVEN raw scores confirmed the predictive power of rated proficiency and use (at home). Interestingly, while rated proficiency has a significant positive effect on LexVEN accuracy scores (Figure 4a), regardless of whether the item was a real or a pseudoverb, use at home significantly predicted performance only with real verbs, showing a non-significant yet opposite trend with pseudoverbs (Figure 4c), which highlights that real and pseudoverbs tap into two different dimensions.

We further validated the test by looking at the relation between pre-judgments and (raw) LexVEN scores. Recall that during pre-judgments, individual items were judged for difficulty, which for real words was assumed to coincide with frequency, i.e., high frequency verbs being less difficult. For pseudoverbs, higher difficulty meant increasing plausibility, i.e., participants were more tempted to think that they actually existed. The analysis to assess the predictive ability of pre-judgments on LexVEN accuracy scores could be run on all participant groups. The statistical model yielded a significant positive effect of pre-judgment on LexVEN real verbs, while the effect of pre-judgment on pseudoverbs was significant only for the Venetan group, showing a tendency to over-accept more plausible pseudoverbs (Figure 3c). This effect was not driven by participants’ (self-assessed) proficiency. Again, this shows that the pseudoverbs tap into a different dimension than real verbs, as has been previously argued by Gyllstad et al. (2023). The fact that non-Venetan speakers rejected pseudoverbs consistently (except for two speakers who were excluded due to an exceptionally high number of false alarms) shows their conservative behavior, rejecting words that they have never seen before (Figure 3a and 3b). In the group of Venetan speakers, pseudoverbs and real verbs pattern together (Figure 3c), confirming the predictive power of pre-judgments in both conditions. Interestingly, pre-judgments had a somewhat weaker effect on pseudoverbs compared to real verbs, indicating that despite the strong resemblance of pseudoverbs with existing verbs, Venetan participants were conservative and correctly rejected them.

Overall, our analysis shows that pseudoverbs and real verbs might yield slightly different patterns in some user groups, even if both are necessary in order for a yes/no task to yield reliable results. We acknowledge that amongst our participants, users with intermediate proficiency are underrepresented, because it is harder to recruit them and convince them to provide a speech sample. This could be amended in future studies. The inclusion of more speakers from this group might increase the predictive
power of pseudo-verbs. While low proficiency speakers are unfamiliar with both real and pseudoverbs and while high proficiency speakers confidently reject verbs they have never seen or heard before, speakers of intermediate proficiency, especially when being sensitive to dialectal variation, might be inclined to accept phonotactically and orthographically plausible pseudoverbs.

7.2 Support for inferences and warrants

When creating and evaluating LexVEN, we have followed Kane’s (2012, 2013) argument-based approach to validation contribution. Table 1 shows the inferences and warrants postulated for our validation. As for the Domain definition inference (A), the results linked to Warrant A1 showed that Venetan users could easily identify the LexVEN real words, with a mean hit rate (HR) of .97. In comparison, the North and the South group participants had mean HRs of .67 and .32, respectively. As to Warrant A2, the yes/no format is practical in its administration, with a large number of items that take little time to do. The low False Alarm rate (FA) speaks also to the functionality of the task instructions.

In terms of Scoring inference (B), the correction formula for overestimation, which took the number of correctly identified real words minus two times the number of incorrectly accepted pseudoverbs (FA), resulted in markedly lower scores for some test-takers. On the whole, though, the FA rates were low for all three groups. As to practicality, the correction formula should be straightforward to use by non-specialists as it entails summing up the correct answers for the real words (e.g., 43 out of 50), and then taking that score subtracted by the doubled number of pseudoverbs selected as known real words (e.g., \(2 \times 7 \text{ FAs} = 14\)), arriving at the LexVEN score of 29 points (43 - 14). There were no cases of participants receiving negative scores. The possibility of such scores have been highlighted as a flaw of the correction formula (see Meara & Miralpeix 2017 for a discussion on this). Further validation work should investigate alternative correction formulas. Notwithstanding, inference B was arguably supported.

For the Generalization inference (C), three Warrants were postulated. C1 dealt with the reliability of the LexVEN test. The Rasch item reliability was observed at .87 for the 50 real word items, and a Cronbach’s alpha coefficient of internal consistency of .98. These are wholly acceptable levels for a vocabulary test like LexVEN. The observed mean item-total correlation for the 50 real words was .70, indicating a high level of discriminative power. The Rasch item and person separation indexes, however, were a bit lower than what was hoped for. As discussed, this need not be a weakness of the task, but potentially result from the high number of highly proficient speakers in the sample (see also Figure 2). Further validation work on the test should target a higher number of speakers with intermediate proficiency, even if these are harder to find. Warrant C2 dealt with data fit to the Rasch model. The analysis reported in Table 4 yielded a very small number of misfitting items and persons, and it can be concluded that the scores for the 50 real word items had a good fit to the model.
Finally, the C3 warrant was supported as the different-ability groups assumed were significantly different in terms of an ANOVA with post-hoc tests for mean scores. In conclusion, the (C) inference was largely supported.

As to extrapolation inference (D), we have discussed above in Section 7.1 that the scores of the LexVen correlated with spoken samples of Venetan, as well as with self-rated proficiency and use. We can therefore conclude that the test is suitable for low-stakes situations where an estimate of lexical proficiency in Venetan is needed (Decision Inference E).

Finally, an always present challenge when developing tests for primarily spoken varieties with no agreed-on standard is the presence of substantial variation. We tried to eliminate the risk of including (sub-)region specific verb items during the prejudgment-phase by recruiting consultants that represented the entire Veneto region, and we eliminated verbs that were apparently unknown to users from specific geographical regions. The risk of disadvantaging users from specific regions extends to pseudoverbs. For example, a reviewer remarked that the problem with the pseudoverb pironare (derived from the noun pirón ‘fork’), which we had eliminated due to unsatisfying point-biserial correlations (Table 3), is similar to the verb inpironare. Following the reviewers’ comment, we consulted with some Venetan users, who either did not know the verb or agreed that it means ‘to skewer (with a fork)’. The example shows that also pseudoverbs may appear more or less plausible, depending on a user’s variety. Thus, work on assessment with regional languages requires particular caution when it comes to variation within the region. Relatedly, we acknowledge that during the validation of LexVEN, the Alpine dialects (spoken in the province of Belluno and in the Northern area of Treviso) were underrepresented and they differ comparatively more from the other Venetan dialects. Thus, we recommend using the test with caution in these areas, where, ideally, further validation should be carried out.6

As emphasized by Chapelle and Lee (2021), validation is an ongoing process of inquiry and no single study can result in a validated test instrument. We have here provided initial validation evidence related to the scores on LexVEN as indicative of lexical proficiency in Venetan.

7.3 Availability and conclusions

LexVEN is a reliable and quick tool to assess language proficiency in Venetan for anyone who would like to work with users of Venetan and determine their proficiency objectively. The completion of the test requires about five minutes, and it can be used as a pen and paper test or on a computer, tablet, or smartphone. We administered our items in a randomized order and we have presented them in a bimodal fashion (aurally

---

6 We also had fewer speakers from Rovigo, but this is less of a concern because the dialect is more similar to that spoken Padova and Vicenza, where we had many participants.
and written). Since we could not rely on previous assessment tools for Venetan, we have used recorded speech samples (besides self-assessed proficiency and language use) for the validation of LexVEN, which turned out to be a very good predictor.

The final list of the LexVEN items and the instruction we have used are included below (to facilitate reading we added translations, which should be removed when using the test). Since Venetan is primarily spoken, we recommend using a bimodal version. The recorded items and the data analysis can be downloaded from Iris [https://osf.io/f8vpu/].

Acknowledgments

We wish to thank our colleagues, friends and family and all speakers who supported the data collection and validation process. This research did not receive any specific grant from funding agencies. Special thanks to Alexandra Besler and Sebastiano Arona who have supported the experimental set-up.

References


Besler, Alexandra, Ferin, Maria, Venagli, Ilaria, & Tanja Kupisch. 2023. *On the relation between attitudes and dialect maintenance (Sicilian and Venetan) in Italy*. Ms, University of Konstanz.


Cardinaletti, Anna. 2013. On the bilingual acquisition of Italian and venetan dialects: a focus on subject and object clitic pronouns. *Paper Presented at the Generative Linguistics in the Old World (Glow) 36, April 2-6 2013*, Lund University.

Chapelle, Carol, & Hye-won Lee. 2021. Understanding argument-based validity in language testing. In C. Chapelle, & E. Voss (eds), *Validity argument in language testing*, 19–44. Cambridge University Press. [https://doi.org/10.1017/9781108669849.004](https://doi.org/10.1017/9781108669849.004)


Ferré, Pilar, & Marc Brysbaert. 2017. Can Lextales-Esp discriminate between groups of highly proficient Catalan–Spanish bilinguals with different language dominances?.


Linacre, John Michael 2021. ‘Winsteps® Rasch measurement computer program’ Version 5.1.6. Winsteps.com


