

VerboWeb 3.0: Decoding Verb Behavior in Brazilian Portuguese – From Lexical Semantics to Web-Verified Syntactic Patterns

Márcia Cançado

Universidade Federal de Minas Gerais-UFMG

`mcancado@ufmg.br`

Abstract. *This paper introduces VerboWeb 3.0, a lexical-semantic and syntactic database of Brazilian Portuguese verbs. Building on previous releases, it catalogs over 2,000 lemmas and about 20,000 examples (in progress), organized into seven semantic categories and more than 65 classes. Each class is validated against web-attested sentences to ensure accuracy. A key innovation is flexible granularity, allowing verbs to belong to multiple overlapping classes. Classes have been reorganized to better reflect the structure of the BP verbal system. The user interface has been redesigned for intuitiveness and accessibility. VerboWeb 3.0 supports theoretical linguistics through argument structure analysis, provides training data for NLP tasks, and aids education by facilitating the creation of language exercises.*

1. Introduction

The study of verb behavior is at the core of both theoretical and descriptive linguistics. [VerboWeb 3.0](#) provides a comprehensive resource for the Brazilian Portuguese (henceforth BP) verbal lexicon, integrating lexical-semantic and syntactic insights with empirically attested usage patterns. The database currently catalogs over 2,000 verbs and approximately 20,000 constructed examples (still in development), systematically organized into seven semantic categories, over 65 classes, and their respective syntactic properties, many of which validated against real-world sentences retrieved from the web.

Building on the framework established in VerboWeb 1.0 (Cançado et al., 2017), version 3.0 introduces three key innovations. First, it relaxes rigid notions of granularity by allowing verbs to participate in multiple overlapping classes, each defined by a core semantic feature. Second, although many of these classes have their origins in earlier versions, they have been reorganized to better reflect the underlying structure of the BP verbal system. Third, it incorporates authentic, web-attested examples (via Google and other corpora) to ensure that its syntactic profiles capture real-world BP usage rather than idealized sentences. In addition, the user interface has been significantly improved, making the database more intuitive and accessible for a broader range of users.

These enhancements make VerboWeb 3.0 not only a powerful descriptive tool for linguists but also a practical resource for language teachers and NLP practitioners. By combining formal theoretical frameworks with corpus-based validation, VerboWeb 3.0 provides an empirically grounded platform for analyzing argument structure, alternation patterns, and lexical-aspectual phenomena in BP.

In this paper, we first survey the theoretical foundations of lexical-semantic verb classification (Section 2). In Section 3, we describe our data collection, and system architecture. Section 4 presents the core contents and interface of VerboWeb 3.0. Section 5 illustrates possible applications in theoretical linguistics, elementary education, and natural language processing. In Section 6, we conclude with a discussion of limitations and future directions.

2. Theoretical Background

This section reviews key lexical-semantic approaches to verb classification, laying the groundwork for the comprehensive syntactic–semantic framework of VerboWeb 3.0. The first step is to define what constitutes a verb class within lexical semantic approaches. Authors such as Fillmore (1970), Pinker (1989), Dowty (1991), Levin (1993), Levin and Rappaport Hovav (1995, 2005), Van Valin (2005), and Wunderlich (2012), among many others, propose that similarities in meaning components alone are not sufficient to classify verbs in a generalized and systematic way.

For instance, Pesetsky (1995) shows that there is no kind of syntactic generalization contrasting verbs that denote emission of loud speech (*holler*, *shout*) with verbs that denote emission of quiet speech (*whisper*, *murmur*). Conversely, the distinction between English verbs that denote a manner of speaking (*holler*, *whisper*) and verbs that denote a content of speaking (*say*, *propose*) seems to be relevant for their selection properties, since only the latter accept sentential complements: *Mary said that she is hungry is well-formed*, but *Mary whispered that she is hungry* is not.

Therefore, in the view adopted by these authors, classifying verbs implies grouping them in clusters that share an array of semantic properties that impact their syntactic behavior, such as possible argument realizations, passivization, reflexivization, etc. Thus, the semantic information carried by a verb is not just a list of idiosyncratic meanings, but contains types of meanings that are grammatically relevant. These meanings are encoded in the semantics of verbs and can be represented in different ways.

Furthermore, Levin (2010) emphasizes that verb classes can vary in granularity, reflecting different levels of specificity in their semantic characterization. For instance, verbs can be grouped into coarse-grained, medium-grained, or fine-grained classes, each capturing a distinct level of semantic detail.

Medium-grained classes are the most canonical, typically defined by thematic role structures or decomposition patterns (e.g., change of state verbs), which capture the core event structure and argument relationships. Fine-grained classes, on the other hand, represent highly specific semantic groupings, such as verbs of bodily expulsion, which are distinguished by precise, context-dependent meanings. Finally, coarse-grained classes capture broader distinctions, such as the well-established contrast between internally caused verbs (e.g., *grow*) and externally caused verbs (e.g., *break*), as proposed by Levin and Rappaport Hovav (1995), which reflect more general aspects of event causation.

In all cases, these differences in verb class granularity are ultimately grounded in the lexical-semantic representation of verbal meanings, reflecting the varying degrees to which semantic properties influence syntactic behavior. In any case, we expect that these differences in verb class grain-size ultimately find their source in the lexical semantic representation of the verbal meanings.

Building on the foundational work cited above, VerboWeb 3.0's framework operates under the following key assumptions:

Semantic–syntactic coupling: the semantic properties of a verb—or more broadly, a verb class—determine some syntactic organization of sentences.

Non-granular classification: verbs can be classified through multiple semantic lenses, some intersecting with morphological or pragmatic factors. These classifications coexist without a fixed hierarchy of *grain size*.

Multi-class membership: a single verb may belong to multiple classes, provided it shares the semantic feature(s) central to a given classification. It is this feature—specific to a class—that drives syntactic patterning.

The framework proposes semantic categories for verb classification: all verbs are categorized according to their event structure, thematic-role structure, and lexical aspect—properties inherent to verbs and directly linked to their syntactic behavior. Other categories such as semantic content, semantic-morphological content, and semantic-pragmatic content are context-dependent. These classifications apply selectively, depending on the grammatical relevance of a verb's semantic content, whether purely lexical (e.g., meaning), morphologically encoded, or pragmatically constrained. Each syntactic property a verb displays is rooted in specific semantic information tied to its respective category.

These categories are further organized into semantic classes that capture distinct verb behaviors and are systematically linked to corresponding syntactic properties, following the progression: Categories → Classes → Syntactic Properties.

3. Methodology: data and analysis

The methodological goal is to identify, list, describe, analyze, and classify Brazilian Portuguese verbs as exhaustively as possible. To that end, we can list three main stages.

The data collection is based on previous studies and working hypotheses. We choose one target class (e.g., change-of-state verbs) and isolate a defining diagnostic property (e.g., causative-inchoative alternation). Guided by this property, we inspect each entry in a specific verb dictionary (Borba, 1990), a comprehensive BP verb resource, since it provides semantic classifications, corpus examples, and definitions. We also use Houaiss electronic dictionary. We verify each candidate verb against the diagnostic property until we obtain as complete a list as possible.

The individual analysis and classification operate on the hypothesis that semantics drives syntax. We subject each verb to a battery of tests—aspectual diagnostics, passivization, argument-deletion, acceptable adjunction patterns, and so on—to determine its precise syntactic and semantic profile. Verbs are grouped according to shared syntactic behaviors that reflect a common semantic trait. Polysemous verbs may split across classes if their distinct senses correspond to different argument structures. For instance, the verb *uniformizar* can mean either *make uniform*, a change-of-state sense, alongside *quebrar* 'break' (e.g., *O pintor uniformizou a parede* 'the painter made the wall uniform'), or *provide with a uniform*, a change-of-possession sense, alongside *amanteigar* 'butter' (e.g., *O gerente uniformizou os funcionários* 'the manager provided the workers with uniforms'). If multiple senses share the same core event structure, they remain in a single class.

Empirical attestation ensures that the grammatical structures analyzed genuinely reflect BP grammar: crafted sentences are systematically cross-checked with real usage, mainly through Google-indexed examples and supplemented by data from Linguateca and the Corpus do Português. Since some structures lack attestation due to their oral or infrequent nature, we also create illustrative sentences based on dictionary entries to exemplify usage and test specific properties. These constructed examples, validated against attested data, confirm which structures are grammatical or impossible in BP, thereby incorporating negative evidence in line with Chomskyan methodology.

4. VerboWeb 3.0: Architecture and Content

The framework proposes seven semantic categories for verb classification, though more categories may emerge as our understanding deepens. These categories are as follows.

Event structure is the temporal and causal organization of an event, defined by the predicate. It describes how the event unfolds over time—its sub-events (initiation, causation, process, result state), participants, and their relations—aligned with Vendler’s aspectual classes. To capture this, we use a predicate decomposition metalanguage

Thematic-role structure is the system that assigns semantic roles to the participants of an event described by a verb. These roles define how entities take part in the event (e.g., who performs, undergoes, or benefits from the action) and bridge semantics and syntax by determining how arguments map to grammatical positions such as subject and object.

Lexical aspect captures the inherent temporal properties of verbs, distinguishing states, activities, accomplishments, and achievements according to how events unfold in time and whether they reach a natural endpoint.

Selectional restrictions are semantic constraints verbs impose on their arguments, such as requiring an animate subject (*John speaks*) or an inanimate theme (*The glass broke*), reflecting how verb meaning interacts with argument properties. Basically, they govern what kinds of entities can participate in the eventualities denoted by the verb.

Semantic Content refers to the inherent, shared meaning that unites a group of verbs based on their core, context-independent semantic properties. For example, verbs of manner of motion (e.g., *roll*, *spin*, *wave* etc.) all share the semantic feature of describing specific types of movement.

Semantic-Morphological Content concerns the relation between verb meaning and morphological derivation from nouns denoting instruments, locations, or means. Such verbs retain a semantic link to the source noun, which can often reappear syntactically as an adjunct (e.g., *Ele martelou o prego com um martelo de pedra* ‘he hammered the nail with a stone hammer’).

Semantic-Pragmatic Content captures verb meanings that are context-dependent and shaped by cultural and grammatical conventions. Such verbs operate only within specific frameworks where syntax, semantics, and pragmatics intersect (e.g., *Eu cortei o cabelo com o famoso cabeleireiro* ‘I cut my hair with the famous hairdresser’). In this case, the preposition *com* (‘with’) introduces an adjunct indicating delegated agency, which is licensed because this type of action is pragmatically conventional—typically performed by someone else upon request.

First, we present the category organization in VerboWeb's interface, as shown in Figure 1.

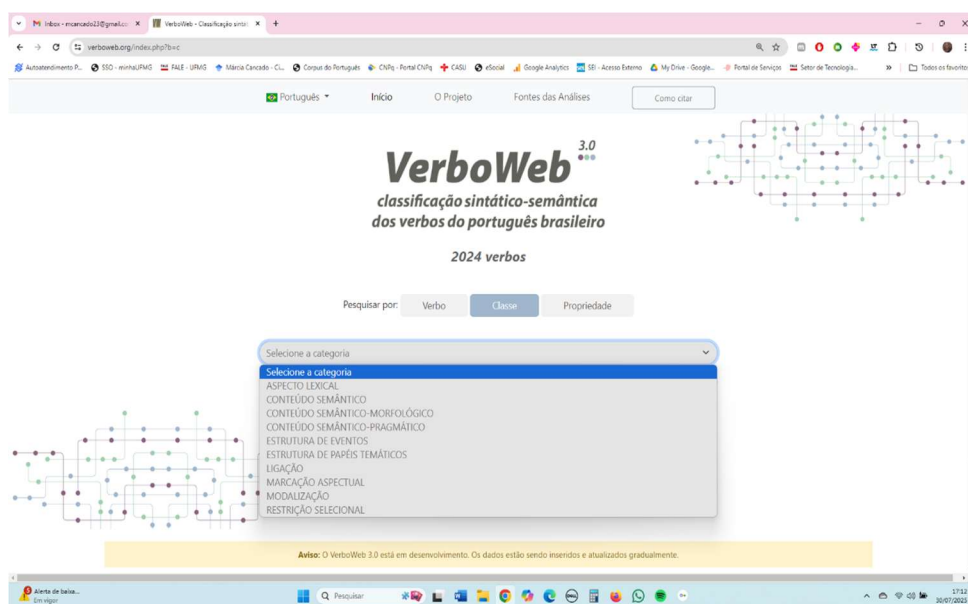


Figure 1. Semantic categories

Figure 2 illustrates how specific categories are linked to a particular verb.

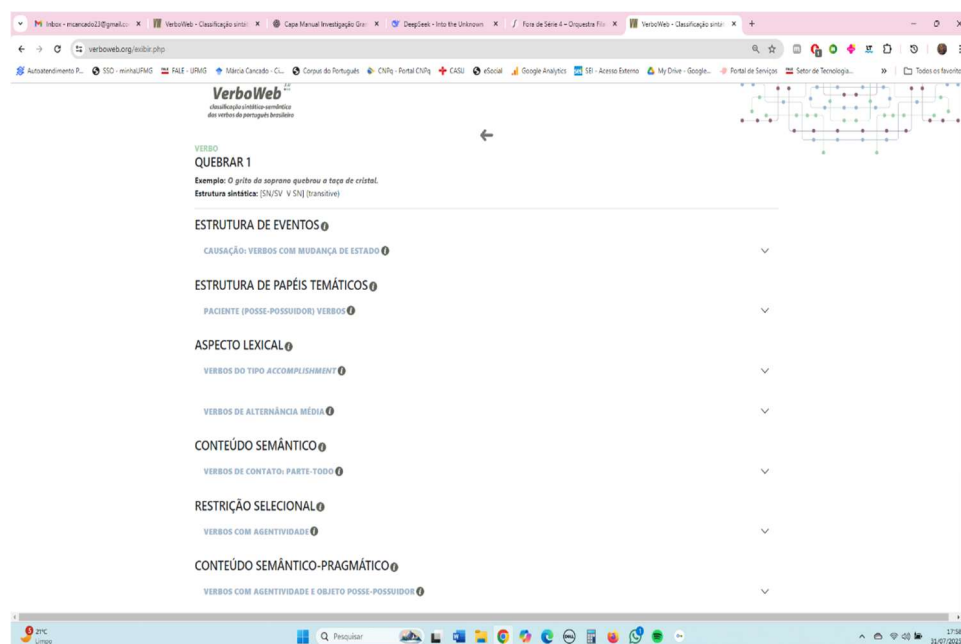


Figure 2. Specific categories linked to a particular verb

Each category comprises specific classes related to its semantic feature. For example, the event structure category includes causation types and their specifications (result events, state changes, etc.), culmination types and their specifications (results, locations, etc.), manner verb types and their specifications (affected, instrument, etc.), and state types and their specifications (locative, possessive, etc.), as shown in Figure 3.

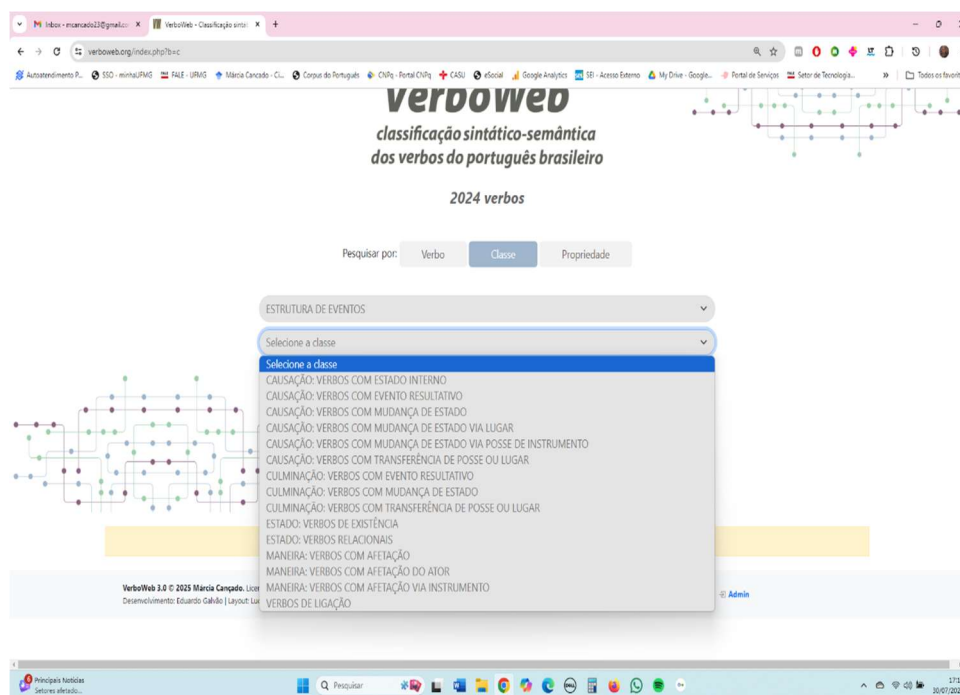


Figure 3. Semantic classes associated with the event structure category

The following illustrates how a specific class appears in the database:

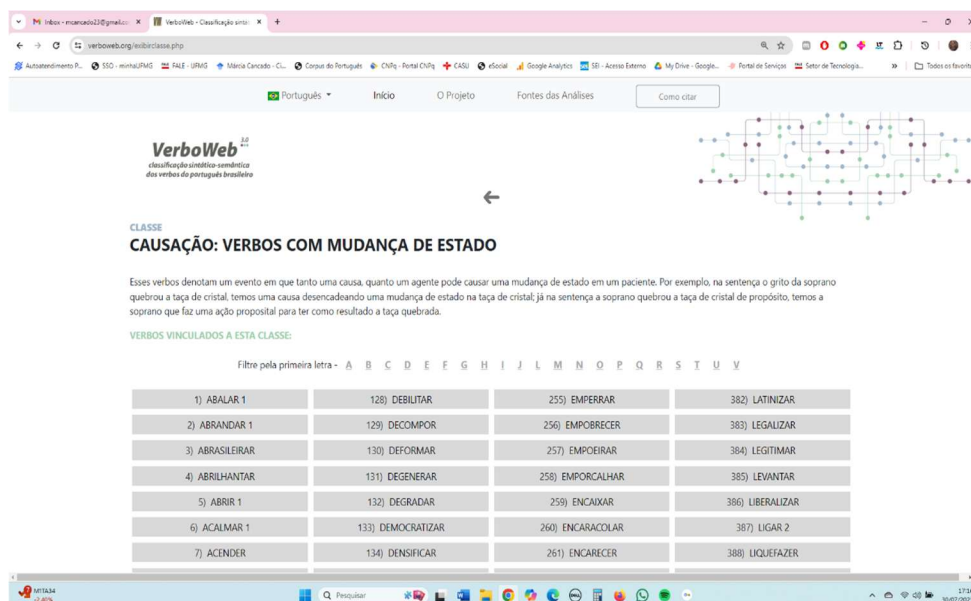


Figure 4. Detailed example of a specific Class

Furthermore, all of this information is consolidated in each verb's template, which records the class-defining properties (e.g., passive, causative-inchoative alternation, etc) derived from the verb's semantic structure. For instance, any verb whose event structure can be represented as $[[X \text{ ACT}] \text{ CAUSE } [\text{BECOME } [Y <RESULT-STATE>]]]$ will consistently exhibit: resultative passive, stative passive, causative-inchoative alternation (marked by the clitic *se*). Each property is illustrated with carefully constructed examples and corroborated by authentic Google attestations, as shown in Figure 5.

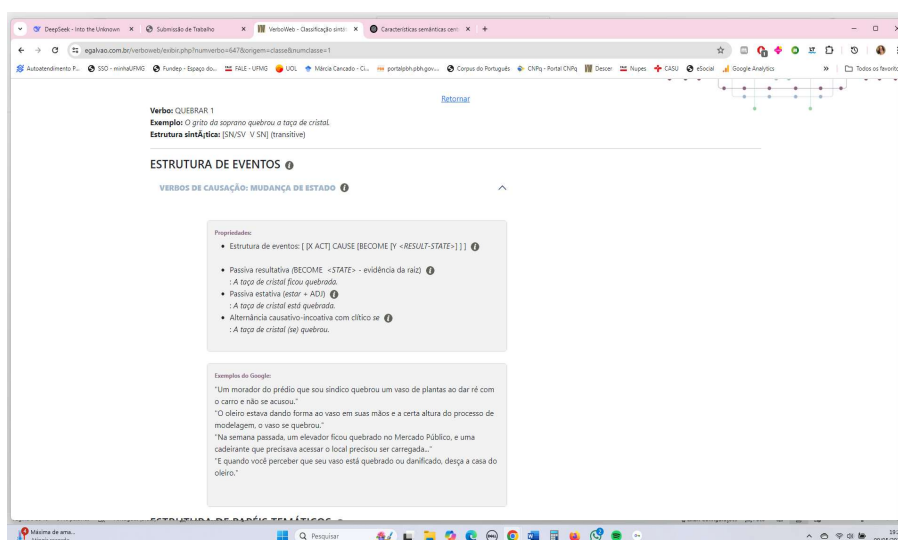


Figure 5. Properties and constructed examples, accompanied by attested occurrences

Finally, Figure 6 illustrates how VerboWeb 3.0 delivers in-context explanations and references for every category, class, and property via pop-up windows (the trigger icons are highlighted in Figure 5).

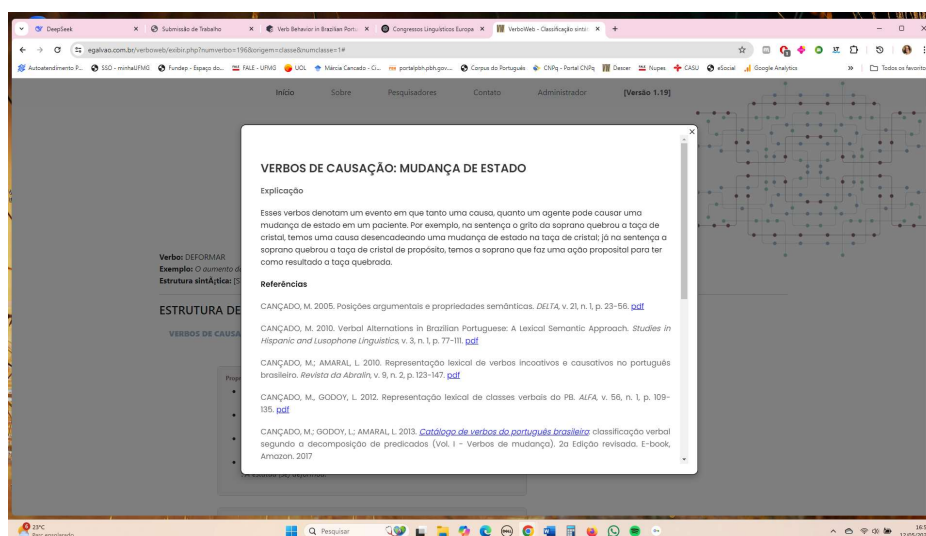


Figure 6. Pop-up activated

5. Applications and Relevance

There are many potential applications of the VerboWeb 3.0 that have yet to be explored. We highlight three key areas, as following.

Theoretical Linguistics Application: VerboWeb 3.0 currently documents over 2,000 BP verbs and provides approximately 20,000 illustrative sentences, with ongoing expansion. It systematically covers Vendler's four aspectual classes: activities, accomplishments, achievements, and states, offering a comprehensive mapping of the BP verb lexicon. This rich empirical resource supports detailed analyses of argument structure, alternation patterns, lexical-aspectual behavior, and other semantically relevant phenomena. Moreover, most of the 20,000 constructed examples are corroborated by web-attested sentences (e.g., via Google), ensuring empirical validity. Because the

database places no theoretical constraints on users, VerboWeb 3.0 is a highly versatile tool for a wide range of theoretical-linguistic investigations.

Potential in Natural Language Processing (NLP): VerboWeb 3.0 offers significant potential for processing NLP applications in BP. Its richly annotated thematic-role data—identifying Agents, Patients, Experiencers, and more—combined with many web-attested example sentences, can serve as high-quality, automatically generated training material for semantic role labeling systems. Additionally, the extracted syntactic profiles—which document properties such as transitivity, passivization, and alternation patterns—can be exported to construct comprehensive subcategorization lexicons for parsers and dependency analyzers, thereby constraining attachment decisions and reducing syntactic ambiguity errors.

Applications in Elementary Education and Portuguese as an Additional Language (PAL): VerboWeb 3.0’s syntactic profiles enable teachers to develop exercises in which students identify subjects, objects and verb complements within age-appropriate sentences, while its exportable examples make it easy to generate customized worksheets that target specific verb classes (such as action versus state verbs) or structures (transitive versus intransitive). By incorporating web-attested sentences, lessons become more engaging and grounded in real-world usage, allowing learners to see how verbs function in everyday communication. Furthermore, educators can leverage VerboWeb’s aspectual classifications to design activities with progressively increasing complexity—starting with simpler action verbs and advancing to more abstract states or achievements—thereby reinforcing both semantic understanding and syntactic proficiency over time.

6. Final Considerations

VerboWeb 3.0 offers extensive coverage of the Brazilian Portuguese verbal lexicon—even in its current final development phase—and provides freely accessible data. In a comparison with existing resources such as VerbNet.Br and Verbo-Brasil, Rodrigues et al. (2022) conclude that VerboWeb delivers more precise linguistic analyses, making it a valuable tool for researchers in theoretical linguistics, natural language processing, and language education.

Nonetheless, several limitations remain. The dataset relies on manual data collection, which demands specialized expertise that is scarce in Brazil. This constraint has limited both the quantity of usage examples and the representativeness of oral or low-frequency constructions. Future work will prioritize the automatic extraction of examples from diverse corpora and the expansion of polysemous and complex verbal expressions.

We invite researchers, developers, and educators to explore VerboWeb 3.0, integrate it into their research and teaching practices, and contribute new examples or corrections via our collaborative platform. By pooling community expertise and feedback, we can continue to refine VerboWeb into an even more comprehensive, accurate, and versatile resource for the study and teaching of Brazilian Portuguese.

Acknowledgements

The author acknowledges the support of CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil) through a PQ Grant, and the use of ChatGPT (OpenAI) to assist in the revision and organization of the manuscript. All content was subsequently reviewed and edited by the author to ensure accuracy and originality.

References

- Borba, F. S. (1990). Dicionário de usos do português do Brasil. Ática.
- Cançado, M. (2025). *VerboWeb 3.0: classificação sintático-semântica dos verbos do português brasileiro*. Available at: <https://verboweb.org>.
- Cançado, M; Amaral, L.; Meirelles, L. (2017). *VerboWeb 1.0: classificação sintático-semântica dos verbos do PB*. Available at: [VerboWeb 1.0](#).
- Davies, M., & Ferreira, M. (2006). Corpus do Português: 1300s–1900s. Brigham Young University & Universidade de Lisboa. Available at <https://www.corpusdoportugues.org>
- Dowty, D. (1991). “Thematic Proto-Roles and Argument Selection”. *Language*, 67(3), 547-619.
- Fillmore, C. J. (1970). “The case for case”. In E. Bach & R. T. Harms (Eds.), *Universals in Linguistic Theory* (pp. 1-88). Holt, Rinehart and Winston.
- Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.
- Levin, B. (2010). “Verb Classes and Alternations Revisited”. In M. Rappaport Hovav, E. Doron, & I. Sichel (Eds.), *Syntax, Lexical Semantics, and Event Structure* (pp. 35-68). Oxford University Press.
- Levin, B., & Rappaport Hovav, M. (1995). *Unaccusativity: At the Syntax-Lexical Semantics Interface*. MIT Press.
- Levin, B., & Rappaport Hovav, M. (2005). *Argument Realization*. Cambridge University Press.
- Pesetsky, D. (1995). *Zero Syntax: Experiencers and Cascades*. MIT Press.
- Rodrigues, R., Lemos-Couto, M., Coelho, F. L., & Vale, O. A. (2022). “Bases lexicais verbais do português brasileiro”. *Domínios de Linguagem*, 16(4).
- Santos, D., & Bick, E. (2000). Linguateca: Resource Center for Portuguese Language Technology. Available at <https://www.linguateca.pt>
- Van Valin, R. D. (2005). *Exploring the Syntax-Semantics Interface*. Cambridge University Press.
- VerbNet.Br. (2022). Brazilian Portuguese Verb Classification Database. Available at <http://143.107.183.175:21380/verbnetbr/index.html>.
- Verbo-Brasil. (2022). Brazilian Portuguese Verb Database. Available at <http://143.107.183.175:21380/verbobrasil/index.php?lang=pt-br>.
- Wunderlich, D. (2012). “Lexical decomposition in grammar. In M. Everaert, M. Marelj, & T. Sioni (Eds.), *The Theta System: Argument Structure at the Interface* (pp. 139-166). Oxford University Press.