Symbolic Artificial Intelligence for Preserving Linguistic and Cultural Diversity

Radomira Lazarova[1,*,†], Veneta Tabakova-Komsalova[1,2,*,†], Stanimir Stoyanov[1,2,†] and Ivan Stoyanov[3,†]

[1]*Paisii Hilendarski' University of Plovdiv, Plovdiv, Bulgaria, 24 Tsar Assen Str.*

[2]*Centre of Excellence in Informatics and Information and Communication Technologies, Sofia, Bulgaria, Acad. G. Bonchev St., Block 2, 1113 - Sofia, Bulgaria*

[3]*Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Acad. G. Bonchev St., Block 2, 1113 - Sofia, Bulgaria*

Abstract

This article examines the role of symbolic artificial intelligence in preserving and operationalising linguistic and cultural diversity under low-resource conditions. It compares a Prolog-based intelligent system for the Rhodope dialects of Bulgarian with two broader symbolic models, Womb Grammar and Doughnut Computing. The contribution of the paper is threefold: (1) it positions the Rhodope Prolog system as an educational and research infrastructure rather than a simple dialectal lexicon; (2) it explains why symbolic and constraint-based methods remain preferable under low-resource, highly variable, and community-verifiable conditions; and (3) it proposes a staged connection between local dialect modelling, grammar induction, and socially responsible deployment. The paper argues that these approaches are complementary rather than competing: the Rhodope system captures fine-grained lexical, morphological and phonetic particularities; Womb Grammar supports translingual grammar induction; and Doughnut Computing extends constraint-based reasoning toward socially and ecologically orientated planning. By combining related work, a theoretical framework and a comparative analysis, the study shows that symbolic AI remains especially valuable where interpretability, modularity, community participation and data scarcity are decisive constraints.

Keywords

symbolic artificial intelligence, logic programming, Prolog, Rhodope dialects, Womb Grammar, Doughnut Computing, low-resource languages, digital humanities.

1. Introduction

The digital disappearance of languages and dialects today is taking place in parallel with their demographic weakening. When a speaking community loses active speakers, and at the same time lacks machine-readable corpora, lexical resources, educational materials, and robust analysis tools, the risk is not only that the language will narrow down in use, but also that it will remain outside of future search technologies, training, and cultural memory. That is why the problem of maintaining low-resource and endangered languages has gradually become a common topic for computational linguistics, documentary studies, education and language policy [1], [2], [3].

In many of these contexts, large neural models do not provide an automatic solution. They are powerful in large-scale corpora and widely spoken languages, but with local dialects, unstable spelling, limited data, and the need for clear verifiability, problems often arise with traceability of the inference, sustainability of assessment, and institutional trust. Therefore, symbolic artificial intelligence should not be thought of as an "obsolete" alternative but as an engineering and humanitarianly relevant paradigm

Fourth Prolog Education Workshop 2026, Lisbon, Portugal

*Corresponding author.

†These authors contributed equally.

radomira.lazarova@uni-plovdiv.bg (R. Lazarova); v.komsalova@uni-plovdiv.bg (V. Tabakova-Komsalova); stani@uni-plovdiv.bg (S. Stoyanov); ivanstojanov22@gmail.com (I. Stoyanov)

0009-0003-8231-634X (R. Lazarova); 0000-0002-0617-9844 (V. Tabakova-Komsalova); 0000-0002-3854-4260 (S. Stoyanov); 000-0002-5943-399X (I. Stoyanov)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-03

CEUR-WS.org/Vol-4311/PEG_regular_05.pdf

for tasks where transparency, modularity, and human control are essential requirements, not secondary improvements.

This paper does not claim that statistical or neural methods are irrelevant for dialect preservation. Rather, it argues that in the specific case of small dialectal datasets, unstable orthography, scarce aligned speech resources, and the need for local expert verification, a symbolic layer should precede or constrain statistical components. LLM-based tools may later support search, hypothesis generation or annotation assistance, but they should not replace explicit rules for morphology, phonetic variation, and microdialectal provenance when the goal is preservation, teaching, and community trust.

Bulgarian dialectology has a long tradition, but the transition from archiving to operational digital use remains partially complete. Projects for a digital corpus of Bulgarian dialects show how laborious it is to transform audio recordings, transcriptions, and metadata into research infrastructure. The Rhodope dialects are particularly important in this picture because they preserve archaic and regionally distinguishable phonetic and morphological features [4], but access to systematised digital resources for them is still limited. This makes each working prototype [5], [6] not just a technical experiment but part of a broader program for the preservation of cultural memory.

A further motivation is educational. The Rhodope system is not only a preservation tool, but also a Prolog-based learning environment in which students and young researchers practise declarative modelling, rule design, linguistic annotation, and culturally responsible knowledge engineering. This explicitly connects the project with the objectives of PEG: teaching logic programming through meaningful, socially grounded, and linguistically informed tasks.

This paper has three goals. First build a meaningful theoretical framework for the role of symbolic AI in preserving linguistic diversity. Second, systematically present the prototype intelligent system for the Rhodope dialects based on the project manuscript provided and related publications. Third, compare it with Womb Grammar [7] and Doughnut Computing [8] two universalising symbolic models – one aimed at grammatical induction and transfer and the other towards socio-ecological planning. The main thesis is that there is no contradiction between local language specificity and universal symbolic architectures: they operate at different levels and can be combined productively [8].A further motivation is educational. The Rhodope system is not only a preservation tool but also a Prolog-based learning environment in which students and young researchers practise declarative modelling, rule design, linguistic annotation, and culturally responsible knowledge engineering.

The paper makes three specific contributions. First, it reframes the Rhodope Prolog prototype as an executable model of dialectal knowledge with educational value, rather than as a dictionary-like resource. Second, it compares this local system with Womb Grammar and Doughnut Computing in order to distinguish local specificity, formal universality, and normative universality. Third, it identifies a staged development path in which expert-built symbolic resources can later support controlled grammar induction and carefully bounded hybrid extensions.

2. Related works

2.1. Endangered languages, low-resource NLP and documentary infrastructure

The current literature [1], [9], [10] on endangered languages increasingly insists that language documentation and NLP should evolve collaboratively, rather than sequentially. Early efforts in computational linguistics for endangered languages often concentrated on extracting small resources, while newer approaches think of the archive, annotation tool, corpus, and training use as a connected ecosystem. In this sense, the digital support of small languages is not simply a matter of “catching up” with larger languages but a different design problem in which the fundamental units are the community, variability, and the long-term maintenance of resources.

A significant part of this literature [2], [10], [11] emphasises that low resources are not just a quantitative data deficit; they include a lack of standardised spelling, low institutional visibility, a lack of robust annotation practices, and limited participation of local communities in technological modelling. Precisely for this reason, publications related to indigenous language processing increasingly emphasise

collaborative resource development, ethical responsibility, training of local participants, and preserving the connection between language, territory, and cultural practices. In this way, what might seem like a purely engineering problem becomes a matter of infrastructural justice.

2.2. The Bulgarian Context: From Archiving Speeches to Teaching and Research Use

For the Bulgarian language, the idea of digital dialect corpora is not new. Publications such as [4] on the creation of a digital corpus of Bulgarian dialects clearly show that the deep value of such a resource is not only in the preservation of sound recordings but also in the combination of audio, transcription, annotation, and metadata in a way that allows for subsequent phonological, lexical, and social research. Because it shows that any new system of dialect analysis should be thought of not as a closed dictionary but as a growing infrastructure for documentation and interpretation.

In more recent Bulgarian publications[6], [12], related specifically to the Rhodope speech, there is an interesting change: digitisation is beginning to be associated with project-based learning, logical programming and participation of pupils and students in the structured introduction of lexical material. This is an important transition from an archive to an active pedagogical environment. Instead of the language resource remaining available only to a narrow circle of specialists, it becomes an object of training in grammar, formalisation, and cultural identity. It is in this line that the system for the Rhodope dialects discussed in this article is positioned.

2.3. Constraint-based grammars, Womb Grammar, and logic models with a public focus

In terms of formal models, the constraint-based tradition offers a particularly productive bridge between linguistic theory and executable programming. Foundational work in CHR (Constraint Handling Rules) and CHRG (CHR grammars) [13], [14], [15], [16], [17] show that grammatical dependencies can be formulated as constraints rather than necessarily tree-like rewriting rules. In Property Grammars, this idea is further developed through properties such as precedence, obligation, uniqueness, exclusion, and dependency. It is this modularity that is especially valuable in language systems with high variability, because changing a property can extend to an entire set of constructs without completely rewriting the grammar.

The Womb Grammar Model [18] builds on this line by turning constraint-based description into a grammatical induction mechanism. Instead of assuming that the grammar of the target language is already available, the model derives constraints for it through a combination of a lexicon, a representative corpus, and a known syntactic system of a closely or universally constructed source language. From its initial formulation in grammatical induction, through applications in second language tutoring and language acquisition to work on Yorùbá, Ch'ol, and Mapuche, Womb Grammar is emerging as one of the few consistent symbolic programs for under-resourced grammar enginering [7], [19], [20],[21], [22].

At the same time, in the broader debate on AI for Social Good [7], [8], [23] symbolic models are being developed, the purpose of which is not linguistic but normative and socially oriented. Doughnut Computing is an illustrative case: it uses inferential programming and restrictive dependencies but directs them to public service planning and sustainable urban policies. Although it goes beyond linguistic analysis, this model is relevant to this article because it shows how the same symbolic epistemology can be transferred from grammar and lexicography to socio-ecological decision-making.

3. Theoretical framework

3.1. Symbolic, sub-symbolic and neurosymbolic artificial intelligence

Within this article, symbolic AI is understood as an approach where knowledge is explicitly represented through rules, facts, types of dependencies, and verifiable constraints. Subsymbolic AI, on the other hand, works primarily through statistical parameterisations and distributed representations, in which

the rules are not immediately observable, but are reconstructed from the behaviour of the model. Neurosymbolic AI tries to combine the two logics, but it does not negate the need for a clear delineation of tasks: there are problems where the advantage of symbolic representation is not in accuracy "at any cost" but in controllability, verifiability, and the ability to model rare, regional, and culturally specific phenomena.

Low-resource languages and dialects are just such a case. When data is scarce, spelling is variable, and the goal is not only classification but also the restoration of correct lemmas, grammatical features and regional features, symbol systems have several immediate advantages: they allow manual authentication of knowledge, support local exceptions without "blurring" them into common vector representations, and can be executed in an environment with minimal computing resources. This does not mean abandoning hybrid solutions but prioritizing the right base layer for the specific type of task.

3.2. Constraint-based representation and executable grammar

Constraint-based models [13] are particularly suitable for linguistic heritage because they describe the structure in terms of obligations and compatibilities, and not solely through permissible trees. In Property Grammars [16], for example, phrasal acceptability is determined by which properties are satisfied and which are violated. This is crucial in cases of variation, where both “error” and regional markedness carry analytical value. If the model can indicate exactly which constraint has been violated, it becomes useful not only as a parser but also as a tool for diagnosis, comparison, and training.

The technical side of this paradigm is provided by CHR and CHRG, which turn logical constraints into directly enforceable grammar. Instead of rules being seen only as a descriptive layer, they become an operational mechanism for recognition, induction, and transformation. It is at this point that dialect analysis and Womb Grammar meet: both approaches need knowledge that is formalized enough to be executed but also modular enough to tolerate local peculiarities, exceptions, and iterative additions.

3.3. Prediction, unification and robustness in variable inputs

The new work of Philippe Blache [24] provides further theoretical support for the symbolic approach, showing how prediction and unification can be thought of as a single cognitively motivated architecture for linguistic understanding. In this perspective, base units are not just linear words but richer characters containing formal and semantic information that are activated, compared, and unified in context because it shows that symbolic representation is not necessarily "rigid" or mechanistic: it can work with partial specification different degrees of activation, and controlled loosening of constraints.

Rhodope speech, like many under-resourced languages, is characterized by variations in spelling, articulation, morphological endings, and local preferences. A model that operates only with idealized standardised input misses much of the real linguistic dynamics. Constraint-based and unification-oriented approaches [18] allow for a finer description: to fix which properties are mandatory, which are variant, and under what conditions several surface forms can be interpreted as implementations of the same grammatical unit.

4. From Local Dialectology to Universal Symbolic Models: The Intelligent System for the Rhodope Dialects, Womb Grammar and Doughnut Computing

4.1. The Intelligent System for the Rhodope dialects: objectives, architecture and linguistic scope

The main objective of the system is twofold: firstly, to preserve as precisely as possible the lexical, morphological and phonetic features of the Rhodope dialects [6], and secondly, on this basis to support the distinction between individual Rhodope microdialects. This goal is researchally significant because it changes the focus from the simple preservation of dialect units to the formal modelling of the differences

between them. In this sense, the system [6] goes beyond a lexical archive: it represents an attempt to provide an executable description of regional variation in Bulgarian, implemented through declarative means.

Architecturally, the prototype [6] is conceived as a fully Prolog-based environment in which the distributed knowledge base, management layer, and user interface are implemented in a single logical ecosystem. The manuscript describes a modular division into parts of speech – nouns, adjectives, verbs, invariable words, idioms, later numerals, and full nominative dialectisms – as well as local web access through the SWI-Prolog HTTP stack. The web interface is implemented within the same SWI-Prolog ecosystem as the knowledge base, which is methodologically important because the storage of lexical facts, the derivation of grammatical features, and the presentation of results remain traceable within a single declarative environment. Low-level implementation details of the HTTP stack are secondary to this architectural point. The knowledge base is divided into modules, according to the belonging of the words to the parts of speech. This choice is important not only technically, but also methodologically: the same environment supports the storage of facts, the derivation of grammatical characteristics, and the interface presentation of the results.

One of the system's greatest strengths is its focus on detailed grammatical metainformation. In the described knowledge base, each lexical unit is accompanied by characteristics depending on the category: gender, number, lemma, adjective type, conjugation, reflexivity, impersonality, stress, and regional affiliation. This allows the processing not to be limited to a simple substitution of "standard form – dialect form", but to model systemic dependencies between phonetics, morphology and microdialect. This is where symbolic representation shows its power: it can fix rules and exceptions at the same time, without sacrificing interpretability.

Especially indicative is the phonetic layer. In the manuscript, the Rhodope dialects are divided into six microdialects, and the phonetic phenomena are grouped into three large classes: transitions and transformations in word formation, regional phonetic differences, and variations in spelling and articulation. Formal predicates for consonants, vowels, palatalisation, accent, sonority, diacritically modified vowels and specific regional transitions are also described. For dialectology, this is an essential move, because it translates traditional descriptive observations into a machine-executable scheme that can be used for both analysis and training.

The morphological module for nouns and adjectives is another key component. [6] describes the dynamic construction of article morphemes and the processing of movable stress, as well as separate procedures for different types of adjectives. It is not only technical ingenuity that is important but the fact that the system refuses to reduce the Rhodope data to "noise" in relation to the standard language. On the contrary, it treats dialect inflection as its own system of dependencies, which requires separate predicates, separate categories, and separate rules for two-way processing – from standard to dialect language and vice versa.

A special case is the transformation of standard-language nouns with mobile stress. Here, the system not only outputs the corresponding dialectal lemma, but also dynamically constructs the definite form according to the Rhodope three-way definite article system, taking into account the mobile stress and activating a phonetic transformation of the vowel. In this process, the stressed vowel becomes unstressed, while the vowel to which the stress shifts is converted into a stressed one. This transformation is not symmetrical because it cannot be said that stressed and unstressed vowels simply swap places; rather, it follows the described rules governing the transition from stressed to unstressed vowels, as well as the transformation of Old Bulgarian nasal vowels across different microdialects. It is this type of example that shows that the Prolog-based architecture does not function as a simple dictionary converter but as an executable model of interdependencies between morphology and phonetics, which makes it especially valuable for low-resource dialect resources.

For verbs, numerals and full nominative dialectisms, the prototype adopts the same general philosophy: categorical specificity is not masked behind a universal but imprecise mechanism but is described by separate modules and rules [6]. This is an important limitation and an important dignity at the same time. Limitation – because maintenance becomes more labour-intensive. Dignity – because the system remains faithful to the linguistic reality of Bulgarian and its Rhodope varieties, where word

formation and inflection are rich, exceptions are many, and "unified" processing often leads to erroneous generalisations.

Prototype design results [6] – over 500 lexical entries, 94% accuracy in POS recognition, 96% in lemmatisation, and 92% recall with minimal resource consumption. These values are promising, but we interpret them cautiously until they are confirmed with a clearly defined corpus, protocol and error analysis. They should be considered as indicators of prototype maturity and not as a definitive standard for comparison. We believe that searching for similar results in a symbolic environment is already significant for under-resourced NLP.

4.2. Womb Grammar Model: from constraint properties to grammar induction

The original contribution of the Womb Grammar Model [7] was that it turned constraint-based grammar from descriptive formalism into an induction mechanism. In the original formulation of Dahl and Miralles, the source language provided a certain syntactic system, and the target language provided a lexicon and a representative corpus of correct phrases. The hybrid parser used the syntax of the former and the lexicon of the latter; the broken properties in the corpus of the target language then served as a signal for conversion of grammar so that it begins to accept the target constructions. This is conceptually elegant because the underlying operation is not massive statistical training, but analytical "weeding" and rearranging constraints [13], [14], [15].

Even more radical is the idea of Universal Grammar Parsing [20], in which the model does not necessarily start from a specific close language, but from a universal set of potential constraints between valid categories. If the corpus is sufficiently representative, the system gradually removes the unsatisfied properties, leaving only those that characterize the addressed subset of the target language. A operational: it is located in the space of possible formal constraints, from which a specific grammar is derived by eliminating the incompatible ones.

Further developments [18], [19], [20]; show that Womb Grammar (WGM) does not remain a one-off conceptual idea. It has been applied in second language tutoring models, in research on universal limitations in language acquisition, in automated transformations between constraint language models, and in descriptions of partial grammars. This line is important for this article because it shows that WGM functions as a generative framework for multiple cases and not just as a single workshop proposal.

The empirical applications to Yorùbá [18] and Ch'ol are particularly revealing. In Yorùbá, inductive procedures are developed for a subset of noun phrases, demonstrating that constraint-based grammatical induction can also work for a language with its own structural and tonal specificities. In Ch'ol, the model is explicitly presented as a solution for under-resourced language grammar induction. the specific result, and the repeatability of the methodology: lexicon, representative corpus, restrictive properties, and gradual reconfiguration of grammar.

Very recent work further develops the same logic in yet another critical context. Correa et al. [21] conclude that for Womb Grammar the corpus and lexicon are not secondary additions, but the actual inputs to future grammatical induction. Thus, the model acquires a clearly documented dimension: before "learning" grammar, it requires organized, well-motivated data collection. This is directly relevant to the Rhodope system because it shows how carefully organized lexical and corpus work can become a prerequisite for the next layer of automation.

At the same time, WGM also has distinct limitations. In classical publications, it is most convincingly demonstrated on subsets of phrasal grammar, especially noun phrases; it is dependent on the quality and representativeness of the housing; and in some applications, it still relies on significant expert control over categories and properties. However, this is not a weakness that invalidates it compared to the Rhodope project. On the contrary, this is where the possible complementarity comes into play: a locally built and expertly validated knowledge base can become "ripe ground" on which Womb Grammar can add semi-automatic or automatic induction at a higher level.

4.3. Doughnut Computing and Logical AIs for Societal Good

At first glance, Doughnut Computing stands far from dialectology. But if we consider not the theme, but the formal logic of the model, the proximity to symbolic language analysis becomes obvious. In its programming formulation, Doughnut Computing is an inferential-programming methodology that uses Doughnut Economics' benchmarks to model solutions that combine social well-being and environmental constraints. That is, again we have explicitly presented dependencies, eligibility criteria and verifiable inference logic – but this time applied to service planning, indicators, and public policy.

This model is important for this article for two reasons. First, it shows that symbolic AI is not limited to grammar, formal logic, or expert systems in the narrow sense, but can work as a normative compass in societal systems. Second, it brings the topic of cultural heritage closer to the field of AI for Social Good [8], [23]. If a language technology is truly socially responsible, it cannot be limited to technical accuracy; must take into account accessibility, community inclusion, institutional sustainability, and effects on education and local cultural policy.

In this sense, Doughnut Computing adds to Womb Grammar and the Rhodope system not a new linguistic technique, but a new layer of assessment. WGM asks: how to extract or reformulate grammar through constraints? The Rhodope project asks: how to model a local dialect complex as accurately as possible? DCM asks, 'How can this formalisation serve human and environmental well-being when it becomes part of public decisions, educational programmes, and institutional infrastructures?' It is this shift in scale that makes the comparison between the three models productive.

The PEG 2.0 educational initiative adds to this picture because it formulates a more general programme for the widespread dissemination of reliable logical thinking skills and trustworthy coding. Here the connection with the Rhodope system is immediate: if students, teachers, and regional researchers are involved in structuring a dialect resource through Prolog, then the project is no longer just a storage tool but an environment for the formation of logical, linguistic, and civic competence. Thus, symbolic AI is not a substitute for humans but a technique for organising joint knowledge.

5. Benchmarking: similarities, differences and conceptual additions

The three cases – the Rhodope system, Womb Grammar, and Doughnut Computing – should not be seen as competing solutions to the same problem. They belong to different levels of abstraction and to different application domains. But precisely because they are different, the comparison between them brings to light a common symbolic epistemology: explicit representations, rules and constraints, human validation, modular extensibility, and the ability to trace the causes of a given result. This distinguishes them sharply from the black-box paradigms, especially in conditions of limited data and the need for culturally responsible interpretation. Table 1 summarises the comparison across six dimensions that are central to low-resource symbolic AI: object of modelling, basic units, source of rules, prerequisites, advantages, risks, and most appropriate use cases.

The comparison shows that the three models solve different parts of the same preservation pipeline. The Rhodope system contributes depth and local accountability; Womb Grammar contributes a method for moving from curated data toward constraint induction; Doughnut Computing contributes a framework for evaluating whether a technical system serves broader educational, cultural and social goals. The analytical value of the comparison is therefore not that one model outperforms the others, but that each one occupies a different level in a staged symbolic architecture.

The most significant similarity is in the idea that knowledge should remain visible. In the Rhodope project[6] visibility is in the form of lexical notations, morphological metadata, and phonetic predicates; in Womb Grammar [7] – as properties whose satisfaction or distortion can be observed directly; in Doughnut Computing [8], as criteria, indicators, and constraints by which a public plan is evaluated. This visibility has a scientific cost and benefit: it makes systems slower to initially construct but much more suitable for critique, verification, training, and collective revision.

However, the differences are just as important. The Rhodope system is locally and morphologically focused: its success depends on the ability not to lose detail. Womb Grammar [7], on the other hand,

Table 1
Comparison between the intelligent system for the Rhodope dialects, Womb Grammar and Doughnut Computing

Measurement	Rhodope system	Womb Grammar	Doughnut Computing
Object of modelling	Rhodope microdialects, vocabulary, morphology, phonetics	Phrasal grammar and induction of constraints for the target language	Urban and Community Systems, Services, Resources, Sustainability
Main units	Lemmas, grammatical metadata, phonetic predicates, modular rules	Property constraints, categories, representative corpus, lexicon	Indicators, social and environmental constraints, planning scenarios
Source of rules/constraints	Mostly hand-modelled by experts and project team	Partially known or universal grammar + elimination of impaired properties	Regulatory framework of Doughnut Economics + inferential programming
Prerequisites for work	Curated dialectal lexicon, expert validation, microdialect labels, morphological and phonetic rules	Representative corpus, lexicon, category inventory, initial/universal constraint set	Policy indicators, social and ecological thresholds, stakeholder-validated planning constraints.
Main advantage	Maximum preservation of local detail and high interpretability	Semi-automatic/automatic output of grammatical properties	Links the formal conclusion to societal and environmental criteria
Primary risk	Difficult maintenance and incomplete coverage with resource growth	Insufficiently representative housing or partial induction	Regulatory framework without real community participation
Best Use	Preservation, education, dialect analysis, local interfaces, translation	Grammatical induction for under-resourced languages	AI for Social Good, Public Planning, Sustainable Policies

strives for formal universalisability: it seeks a mechanism by which constraints can be transferred, transformed, and filtered out so as to obtain a grammatical description of a new language or linguistic subset. Doughnut Computing [8], on the other hand, universalises not the formal categories of language but the normative framework of decision-making. For him, the central question is not What is an admissible sentence, but What is a socially and environmentally acceptable policy?

This difference allows us to distinguish between two forms of universality. Womb Grammar [7] offers formal universality: a space of potential properties from which a particular grammar is deduced through constraints. Doughnut Computing [8] offers normative universality: a framework in which different domains – urban planning, services, resources – can be evaluated through a common core of social and environmental objectives [23]. The Rhodope system, by contrast, does not claim to be universal. Its value lies in the precise preservation of local specificity. Thus, the three models do not oppose but fill different fields between concreteness, generalization and public purpose.

6. Importance for digital humanities, education and language policy

Placed in the field of digital humanities, the symbolic AI program under consideration is particularly important. It shows that digital humanities is not obliged to work only with large corpora and statistical summaries but can also extract value from small, densely annotated, culturally sensitive resources. Moreover, such a resource is not just a collection of texts, but structured knowledge of phonetics, morphology, local meanings, regional variants and educational uses. It is an infrastructure in which interpretation and formalisation are both humanitarian and measurable.

Figure 1 visualises the staged relationship proposed in the paper. The Rhodope system anchors the process in local linguistic accuracy; Womb Grammar generalises selected regularities into formal

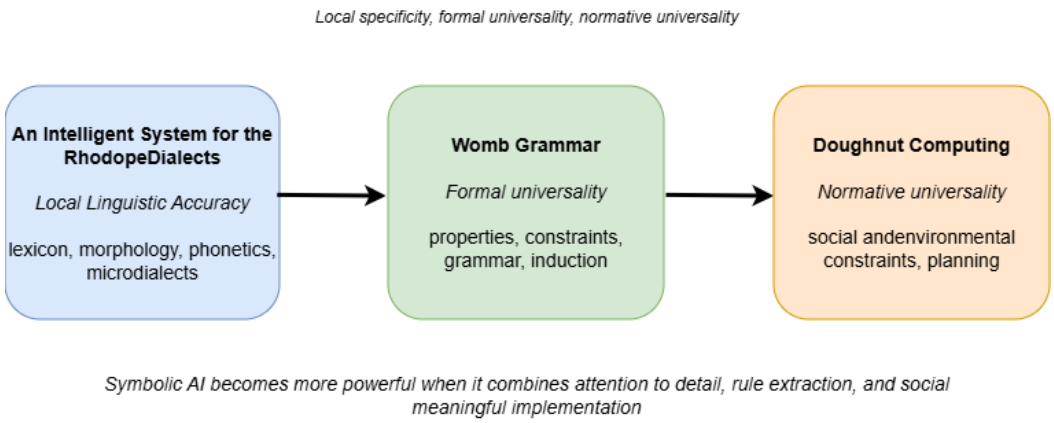


Figure 1: Three levels of symbolic generalisation: local specificity, formal universality, and normative universality.

constraints; and Doughnut Computing shifts the question toward responsible deployment in educational and social contexts. The arrows should therefore be read not as a linear replacement of one model by another, but as a possible path from documentation to induction and finally to socially meaningful use.

The published work of Lazarova et al. [6], as well as the submitted manuscript, suggests the participation of pupils and students in collecting, entering and formalising dialect material. Thus, the work on language heritage becomes integrated learning in grammar, logical programming, structured data, team coordination and cultural identity. In this context, symbolic AI has an important pedagogical advantage over black-box models: it can be explained, edited, and challenged by participants, and not just "used" as someone else's infrastructure.

In educational terms, the Rhodope system offers a concrete route for introducing Prolog through meaningful linguistic data. Learners do not only implement abstract predicates; they encode nouns, verbs, phonetic alternations, regional labels, and exceptions whose correctness can be discussed with teachers, dialectologists, and speakers. This makes the system a bridge between programming education and cultural participation: students learn declarative reasoning while contributing to a resource with local value.

Finally, from the point of view of language policy, such systems have implications far beyond academic publication. They can support local cultural visibility initiatives, serve as a base for museum and archival interfaces, support the creation of teaching aids, and facilitate dialogue between university, school, and local communities. If developed carefully, such symbolic resources can also become a bridge between local philology and broader public policies for multilingualism, digital inclusion, and the protection of intangible heritage.

7. Conclusion

The comparison between the intelligent system for the Rhodope dialects [6], Womb Grammar and Doughnut Computing [7], [8] shows that symbolic AI retains high scientific and societal relevance where tasks are low-resource, culturally specific, and require clear verifiability of the inference. The Rhodope system demonstrates how Prolog and declarative knowledge can describe the subtle dependencies of a local dialect complex. Womb Grammar shows how the constraint-based approach can be extended to a mechanism for grammatical induction and transfer between languages. Doughnut Computing, for its part, reveals that the same symbolic logic can be applied to socio-environmental planning and to the broader AI for Social Good agenda.

The most important general conclusion is that symbolic AI should not be thought of as just a set of techniques but as a form of responsible computability. Where linguistic material is rare, community-relevant and difficult to standardise, the engineering solution cannot be detached from the issues

of accessibility, human participation, and cultural validity. It is here that the power of symbolic description—its traceability, editorial visibility, and modular extensibility—turns out to be not just a technical advantage but a condition for scientific integrity.

In this sense, the article defends the thesis that local dialectology and universal symbolic models are most fruitful not when they compete but when they are connected in stages: first through reliable description and archiving, then through the induction of rules, and finally through socially meaningful implementation. Such a program does not promise an "automatic salvation" of the linguistic heritage. However, it offers a realistic, verifiable, and human-centred technological path for its long-term preservation.

For the PEG context, the central point is that logic programming is not presented only as an implementation technique, but as an educational method for making linguistic knowledge explicit, debatable, and revisable. The revised system therefore contributes simultaneously to dialect preservation, symbolic AI methodology, and the teaching of declarative programming through socially meaningful material.

8. Future development of the system for the Rhodope dialects

8.1. Consolidation of the linguistic and computational layers

The first necessary step is to consolidate the existing modular architecture into more clearly demarcated layers: a lexical layer, a morphological normalisation layer, a phonetic variability layer, and a syntactic/phrase processing layer. In the provided manuscript, lemmatisation is implemented with separate predicates by category, which is justified by language specificity, but in the long run it would be useful to build a central coordination mechanism for tracking decisions, typing errors, and reusing common subtasks.

It is equally important to build a separate layer for spelling and graphic variability. The manuscript itself indicates that various introduced forms and articulation variants are to be processed in an additional architectural layer. This solution is correct and should be developed through carefully defined levels of equivalence between graphic form, phonetic realisation, and lemma. Such a layer would increase the resilience of the system at real user input and would reduce the risk of excessive manual correction of eligible variant forms.

8.2. Corpus Expansion, Multimodality and Womb Grammar as the Next Stage

If the system is to move from a prototype knowledge base to a robust research infrastructure, it needs a wider range of contextualised sentences, audio recordings, speaker metadata, geo-labelling, and a clear description of microdialect affiliation. Here, the experience of Bulgarian dialect corpora and the newer workflows for endangered language documentation [5], [9], [10] is particularly useful: without a lasting connection between sound, transcription, and annotation, any subsequent automation remains limited.

It is on this basis that one can think of carefully introducing womb grammar as a next layer and not as a replacement for the existing system. The most realistic scenario is not a direct derivation of a "complete grammar" of the Rhodope dialects but the induction of restrictive properties for specific phrasal subsystems – for example, name groups, coordination or local construction patterns. Such an approach would allow the manually modelled lexical-morphological basis to be combined with semi-automatic derivation of syntactic dependencies, without sacrificing expert verification of the results.

Multimodal extension – audio, video, eventual limited speech recognition – is also promising but should be developed cautiously. In dialect data, automatic speech recognition often suffers from insufficient representativeness and dominance of literary standards. Therefore, a step-by-step strategy is more appropriate: first, sound and aligned transcription; then tools for semi-automatic annotation assistance; and only at a later stage, a more ambitious ASR (Automatic Speech Recognition) component. Thus, the symbolic core will not be replaced but helped by carefully selected statistical aids.

8.3. Evaluation, public deployment and careful hybridisation

A third direction is the methodological strengthening of assessment. If the system is to be used in education, cultural institutions and future scientific publications, it must have a transparent evaluation protocol: separate test sets by parts of speech, by microdialects, by types of variability, and by directions of transformation (standard to dialect and vice versa). Qualitative error analysis is particularly important because it shows whether the system is wrong due to an insufficient rule, due to a lexical gap or due to structural mixing of regional variants.

Evaluation should include not only standard NLP metrics but also pedagogical and documentary criteria. A revised protocol should specify the size and provenance of the test material, the distribution across microdialects, the parts of speech covered, and the types of errors observed. In addition, classroom use should be evaluated through the quality of student-created rules, the interpretability of explanations, and the ability of learners to revise predicates after linguistic feedback.

From a public perspective, the system would benefit from a clearer open access strategy, versioning, source traceability, and local community involvement. This includes not only a web interface, but also mechanisms for marking the origin of forms, the degree of confidence, regional validity, and editorial verification status. In cultural heritage infrastructures, trust is not built solely through a good interface but through visibility of the origin and the right of the community to participate in the redaction of knowledge.

Finally, hybrid and neurosymbolic extensions are promising but should remain subject to symbolic control. It is possible that embeddings or language-model-based auxiliary components help search for similar forms, prioritise hypotheses, and navigate the corpus but do not supersede the explicit rules for gender, number, phonetic transformation, and regional affiliation. It is especially important to emphasise that expansion to other dialects of Bulgaria should not be taken as an automatic next move. The project material itself marks this possibility cautiously. Therefore, it should be realised only after separate empirical validation and new corpus work for each specific dialect area.

Building interoperability with broader documentary standards and research formats is also a useful next step. Even when the system remains Prolog-based at its core, it could export structured metadata, annotations, and results in formats that allow reuse by archives, corpus environments, and educational platforms. This would reduce the risk that the project will remain an isolated application and facilitate comparative research, long-term archiving, and onboarding of new entrants without losing the traceability of the rules.

In addition, future development should also include a robust mechanism for community and expert management of the resource. For dialects, the question of who has the right to edit, confirm or dispute a form is not a technical detail but part of the preservation of cultural heritage itself. The combination of native speakers, teachers, dialectologists, and developers could provide an editorial model in which the system does not just describe speech but develops as shared knowledge. This is also the best way for the cautious expansion to other dialect areas to happen with local legitimacy, and not just with technical ambition.

Acknowledgments

The work was partially supported by Center for Excellence in Informatics and ICT under grant No. BG16RFPR002-1.014-0018-C01, funded by the Research, Innovation and Digitalization for Smart Transformation Program 2021–2027 and co-funded by the European Union.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] A. Anastasopoulos, et al., Endangered languages meet modern nlp, in: Proceedings of COLING 2020 Tutorials, 2020. doi:10.18653/v1/2020.coling-tutorials.7.

[2] B. Srivastava, et al., Editorial for the special issue on computational linguistics processing in low-resource indigenous languages, ACM Transactions on Asian and Low-Resource Language Information Processing (2023). doi:10.1145/3591208.

[3] UNESCO, Multilingual education, the bet to preserve indigenous languages and justice, 2023. URL: https://www.unesco.org/en/articles/multilingual-education-bet-preserve-indigenous-languages-and-justice.

[4] N. Ikonomov, M. Dobreva, Creation of a digital corpus of bulgarian dialects, in: Dagstuhl Seminar Proceedings, volume 06491, 2006. doi:10.4230/DagSemProc.06491.5.

[5] V. Dahl, Peg 2.0: Future-gazing through a socio-linguistic and historical lens, in: Joint Proceedings of the Workshops and Doctoral Consortium of ICLP 2025, volume 4117 of *CEUR Workshop Proceedings*, 2025. URL: https://ceur-ws.org/Vol-4117/PEG-Regular-1.pdf.

[6] R. Lazarova, S. Stoyanov, A. Dukovski, V. Tabakova-Komsalova, Development of an intelligent system for linguistic analysis and storage of dialects through logic programming (prolog), in: 2025 International Conference on Big Data, Knowledge and Control Systems Engineering (BdKCSE), 2025, pp. 1–8. doi:10.1109/BdKCSE67969.2025.11300513.

[7] V. Dahl, J. E. Miralles, Womb grammars: Constraint solving for grammar induction, in: Proceedings of the 9th Workshop on Constraint Handling Rules, 2012, pp. 32–40. URL: https://exia.informatik.uni-ulm.de/fruehwirth/chr-workshop-CW624-2012.pdf.

[8] V. Dahl, J. J. Moreno-Navarro, Doughnut computing in city planning for achieving human and planetary rights, in: Bio-inspired Systems and Applications: from Robotics to Ambient Intelligence, Springer, 2022. doi:10.1007/978-3-031-06527-9_56.

[9] P. T. Spencer, et al., Documenting endangered languages with langdoc: Nlp enhances archival documentation workflows, in: Proceedings of Field Matters 2024, 2024. doi:10.18653/v1/2024.fieldmatters-1.4.

[10] R. Zariquiey, et al., Cld^2: Language documentation meets natural language processing for revitalising endangered languages, in: Proceedings of ComputEL 2022, 2022. doi:10.18653/v1/2022.computel-1.4.

[11] S. Chiblow, P. J. Meighan, Language is land, land is language: The importance of indigenous languages, Wicazo Sa Review 36 (2022). doi:10.1177/19427786211022899.

[12] V. Dahl, L. A. Cecchi, Introducing prolog in language-informed ways, in: Prolog: The Next 50 Years, Springer, 2023. doi:10.1007/978-3-031-35254-6_13.

[13] P. Blache, Property grammars: A fully constraint-based theory, in: CSLP 2004, Springer, 2005. doi:10.1007/11424574_1.

[14] H. Christiansen, Logical grammars based on constraint handling rules, in: CICLOPS 2002, 2002. doi:10.1007/3-540-45619-8_43.

[15] H. Christiansen, Chr grammars, Theory and Practice of Logic Programming 5 (2005) 467–501. URL: https://arxiv.org/abs/cs/0408027.

[16] D. Duchier, T.-B.-H. Dao, Y. Parmentier, Model-theory and implementation of property grammars with features, Journal of Logic and Computation 24 (2014) 1201–1237. doi:10.1093/logcom/exs080.

[17] T. Frühwirth, Theory and practice of constraint handling rules, Journal of Logic Programming 37 (1998) 95–138. doi:10.1016/S0743-1066(98)10005-5.

[18] I. Adebara, Using womb grammars for inducing the grammar of a subset of yorùbá noun phrases, Triangle 14 (2020) 1–36. doi:10.17345/triangle14.1-36.

[19] L. Becerra Bonache, V. Dahl, J. E. Miralles, On second language tutoring through womb grammars, in: Advances in Computational Intelligence, Springer, 2013. doi:10.1007/978-3-642-38679-4_18.

[20] L. Becerra-Bonache, V. Dahl, J. E. Miralles, The role of universal constraints in language

acquisition, in: New Challenges in Computational Collective Intelligence, Springer, 2013. doi:10.1007/978-3-642-41578-4_1.

[21] L. A. Correa, et al., Hacia una womb grammar para el mapuche: construcción de un corpus de frases nominales, in: XXXI Congreso Argentino de Ciencias de la Computación, 2025. URL: https://sedici.unlp.edu.ar/handle/10915/191144.

[22] V. Dahl, G. Bel-Enguix, V. Tirado, E. Miralles, Grammar induction for under-resourced languages: The case of ch'ol, in: Analysis, Verification and Transformation for Declarative Programming and Intelligent Systems, Springer, 2023. doi:10.1007/978-3-031-31476-6_6.

[23] J. Cowls, et al., A definition, benchmark and database of ai for social good initiatives, Nature Machine Intelligence 3 (2021) 111–115. doi:10.1038/s42256-021-00296-0.

[24] P. Blache, A neuro-cognitive model of comprehension based on prediction and unification, Frontiers in Human Neuroscience 18 (2024) 1356541. doi:10.3389/fnhum.2024.1356541.