Representational Adequacy in Bantu Language Resources: Tone as a Case for Property-Graph Language Databases

Mpho Primus
SARChI Chair in Big Data and the Social Sciences
University of Johannesburg
South Africa
mraborife@uj.ac.za

Sabine Zerbian
Institute of Linguistics: English
University of Stuttgart
Germany
sabine.zerbian@ifla.uni-stuttgart.de

Abstract

Language models typically treat orthographic text as sufficient, but for Bantu languages like Sesotho, this text omits linguistic features such as tone and vowel quality – leading to ambiguities. We argue this is a representational, not data, problem. We propose shifting from static corpora to property-graph, multimodal, versioned databases that link orthography with speech, phonology, morphology, syntax, semantics, variation, and provenance, enabling task-specific dataset generation. Developing such databases requires linguistic expertise; we outline a curriculum to build capacity for creating phonological and phonetic representations, including tone, for computational use.

1 Introduction and Background

1.1 Motivation

The development of large language models has intensified efforts to increase the amount of digital data available for African languages. This work is necessary as African languages remain substantially underrepresented in the digital resources used to train and evaluate contemporary language technologies (Primus, 2025). However, resource scarcity is not only a problem of quantity. It is also a problem of representation.

Computational language resources commonly begin with a written representation of language. Text is collected, normalised, segmented, labelled, and converted into datasets for tasks such as machine translation, sentiment analysis, named-entity recognition, language identification, or language modelling. Even speech resources frequently use orthographic transcription as the primary symbolic representation aligned to the acoustic signal. This approach implicitly assumes that the written string contains, or provides sufficient evidence for recovering, the linguistic distinctions relevant to computational modelling. For many languages, however, orthography is a partial representation of linguistic structure. Features that are contrastive in speech may be absent from ordinary writing. Morphological or phonological distinctions may consequently be collapsed into identical orthographic forms.

1.2 The Representational Problem

Bantu[1] languages provide an important case for examining the limits of orthography-centred computational resources. Most Bantu languages are tonal[2], yet tone is frequently not represented in orthography. Other linguistic distinctions, including vowel quality and aspects of prosodic structure, may similarly be incompletely represented.

Greenberg's foundational work (Greenberg, 1948) established tone as a reconstructed phonological feature of Proto-Bantu, while Odden (1988) showed the considerable diversity of tonal organisation across Bantu languages. Some systems exhibit lexically contrastive tone; in others, tonal patterns are strongly conditioned by morphological or prosodic structure. In either case, linguistically active tonal distinctions need not be visible in the orthography.

Sesotho[3] illustrates the computational consequences particularly clearly. Orthography leaves tone unmarked and collapses phonemically distinct mid-vowel qualities under the graphemes *e* and *o*. Sesotho orthography uses five vowel letters (*a*, *e*, *i*, *o*, *u*), while the Sotho-Tswana vowel system distinguishes more vowel qualities than the orthography marks; the graphemes *e* and *o* each correspond to more than one vowel phoneme (see Cole, 2000 and

[1]The Bantu language family comprises about 600 languages spoken across Central, Southern, Eastern, and Southeast Africa.

[2]Tone is a historically inherited feature of Bantu languages, as reconstructed for the proto-language (Greenberg, 1948), although its functional load and phonological realisation vary considerably across individual languages.

[3]A Bantu language of the Sotho-Tswana group spoken primarily in Lesotho and South Africa.

This work is licensed under CC BY SA 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/deed.en
The copyright remains with the authors.

Khoali, 1991 for the Sotho-Tswana vowel system and its orthographic representation). Consider the question phrase *o batlang?* (*what do you want?*, second-person singular) and *ó batlang?* (*what does he/she want?*, class-1 third-person singular), discussed by Roux and Wissing (2017); the acute accent here is a phonological transcription convention used to indicate the tonal contrast, not standard Sesotho orthography. The same orthographic sequence *o* corresponds to different grammatical interpretations depending on the tonal realisation of the subject marker: the second-person singular and the class-1 third-person marker are associated with different tonal representations. Because orthography does not mark tone, the two forms are written identically.

For an orthography-based computational model, the relevant input may therefore be identical,

$$O(x_1) = O(x_2), \tag{1}$$

while the corresponding linguistic analyses differ:

$$L(x_1) \neq L(x_2). \tag{2}$$

Here, $O(x)$ represents the orthographic form and $L(x)$ a richer linguistic analysis containing grammatical and phonological information. Context may provide evidence from which a model can assign probabilities to alternative interpretations, but probabilistic recovery is not equivalent to representing the linguistic distinction in the resource itself. Increasing the quantity of data expressed in the same orthographic representation does not solve this problem: a billion orthographic tokens cannot directly encode a contrast that the orthography systematically leaves unspecified.

The issue is not that Sesotho orthography is deficient as a writing system. Orthographies are selective representations and are not intended to encode every phonological, phonetic, morphological, or prosodic property of speech. The computational problem arises when this selective representation is treated as though it were a complete representation of the language.

We therefore distinguish three cases that are often conflated: distinctions that orthography does not directly encode; distinctions that may nevertheless be inferred probabilistically from context; and distinctions that require additional phonological, phonetic, or acoustic evidence because no amount of orthographic text can supply them. Tone and mid-vowel quality in Sesotho fall into the third category for the purposes of computational modelling, even where individual tokens are contextually disambiguable.

The problem becomes still more important when the language itself varies or changes. Roux and Wissing (2017) report substantial intergenerational speaker variation in Sesotho tone and describe a system potentially "in a state of flux." A static resource may therefore conflate orthographic omission, linguistic variation, annotation uncertainty, demographic differences, and diachronic change. Representational adequacy requires a resource capable of keeping these dimensions distinct.

1.3 Database-Centred Approach

We propose a shift from static datasets to property-graph, multimodal, versioned language databases from which task-specific datasets can be reproducibly generated. This approach builds on prior work in database-centred resource construction, including relational lexical databases for speech and language systems (Gibbon and Trippel, 2000), database models for diverse linguistic data (Magidow, 2015), and large-scale typological databases (Haynie et al., 2023; List et al., 2025). The proposed structure would link orthography to audio, phonetic measurements, phonological representations, morphology, syntax, semantics, speaker metadata, and annotation uncertainty.

1.4 Contribution

This paper makes three linked contributions. First, it distinguishes language-resource coverage from representational adequacy and shows why increasing the volume of orthographic data cannot recover linguistic distinctions that writing systematically leaves unspecified.

Second, using tone in Sesotho and the broader Sotho-Tswana group as a case study, it proposes a property-graph, multimodal database architecture that links orthography to phonological, phonetic, morphological, grammatical, speaker, and provenance information.

Third, it connects this resource architecture to an interdisciplinary teaching initiative in which students acquire the linguistic and computational skills required to create reliable tone annotations for such resources. We use Sotho-Tswana as a case study, which is S30 in Guthrie's classification, spoken by 16 million people in South Africa alone as well as in Botswana and Lesotho.

This is a position and architecture paper: we present a schema and a research programme, not an implemented system. The contribution lies in the representational argument, the schema specification, and the teaching pipeline, rather than in a completed database or empirical evaluation.

1.5 Structure of the Paper

In this paper, we review the existing resources in Section 2 and present the representational problem in Section 3. In Section 4 we present the proposed property-graph database architecture. In Section 5, we present the outline for a joint teaching initiative which has two aims: First, to transmit the linguistic foundational knowledge on (South) African tone systems as an admittedly specialized yet still classical part of a curriculum of an MA Linguistics. Second, to apply this fundamental linguistic knowledge in the interdisciplinary field of Digital Humanities and provide labeled tone data to be used in the development of the multi-layered databases referred to above. Section 6 discusses limitations, and Section 7 concludes.

2 Coverage and Representational Adequacy

2.1 Existing African-Language Resources

Substantial progress has been made in developing computational resources for African languages. The National Centre for Human Language Technology (NCHLT) corpus supports automatic speech recognition across South Africa's official languages (Barnard et al., 2014); MasakhaNER supports named-entity recognition (Adelani and others, 2022); AfriSenti and SAfriSenti provide resources for sentiment analysis (Muhammad and others, 2023; Mabokela et al., 2025); and SERENGETI substantially expands multilingual pretraining and evaluation across African languages and language varieties (Adebara et al., 2023).

These resources demonstrate an important dimension of progress: *coverage*. Coverage may refer to the number of languages represented, the number of speakers recorded, the number of tokens collected, the number of tasks supported, or the range of domains included.

Following Bender and Friedman (2018, p. 589), we take *representational adequacy* to mean the extent to which a computational resource preserves the linguistic distinctions needed for its intended use. As they observe, broad coverage does not guarantee representational depth for a given question, and a small specialised resource may be deeply adequate for its target phenomena despite limited scope. This distinction is critical because increasing data quantity and improving representation are responses to different problems.

3 What Orthography Leaves Unrepresented

3.1 From Orthographic Form to Linguistic Analysis

In Sotho-Tswana languages, tone participates at lexical and grammatical levels and interacts with morphological and prosodic structure (Cole, 2000; Khoali, 1991; Marlo, 2013; Zerbian, 2017). The computational requirement is therefore not simply to append a tone label to a word. A useful representation must preserve the relations among the orthographic form, the morpheme or grammatical construction, its underlying tonal specification, the surface tonal realisation, and the prosodic domain in which tonal processes apply.

The same principle applies to vowel quality. An orthographic *e* or *o* does not by itself encode which phonemic mid-vowel quality is realised, nor does it contain the corresponding acoustic measurements. Information such as F1 and F2 values, duration, fundamental-frequency contours, and alignment to the speech signal must come from other representational layers.

The distinction can be expressed as:

$$linguistic_{information} \not\equiv orthographic_{information}. \tag{3}$$

Where a computational resource is derived primarily from orthographic strings,

$$R_{text} \subset R_{language}, \tag{4}$$

where R_{text} denotes information explicitly recoverable from the textual representation and $R_{language}$ denotes the set of linguistic distinctions potentially relevant to analysis.

3.2 Underlying and Surface Representations

Tone also demonstrates why a single annotation layer is insufficient. An underlying tonal specification and its surface realisation are related but not identical objects. Khoali (1991) analyses high-tone spread as applying within the phonological word and treats downstep phrasally, while Zerbian and

Barnard (2009) show that the acoustic realisation of a single high tone in Northern Sotho may involve a delayed F0 peak. A resource that stores only the orthographic word and a surface label therefore loses the relation between lexical or grammatical tone, the rule or constraint affecting it, the prosodic domain, and the observed acoustic outcome.

For computational purposes, these relations matter because different tasks require different views of the same linguistic event. A text-processing task may require orthography and morphology; a tone-prediction task may require underlying tonal specifications and grammatical context; speech synthesis may require surface tonal structure and acoustic evidence. The resource should therefore preserve these analyses separately while maintaining explicit links among them.

3.3 Variation, Uncertainty, and Provenance

Representational adequacy also requires variation to remain visible rather than being normalised away. Roux and Wissing (2017) report intergenerational differences in the functional use of Sesotho tone, including cases in which younger speakers made limited or no functional use of the tonal contrast in *o batlang?*. A single categorical label may therefore obscure the phenomenon under investigation. The point is not that orthography omits speaker metadata—it does not purport to encode it—but that a linguistically adequate resource must be capable of representing competing annotations, annotator identity and confidence, speaker age and region, language-use history, acoustic measurements, recording date, underlying and surface representations, and the provenance of subsequent analytical decisions.

Variation should not automatically be treated as noise to be removed before modelling. It may itself constitute linguistic information. This requirement motivates a resource architecture in which linguistic observations and the relations among them can be stored, queried, revised, and versioned without collapsing them into a single flat representation. Speaker variation is thus best understood as a metadata and provenance requirement on the resource, rather than as a further gap in the orthographic system.

4 Database Architecture: A Property-Graph Model

The preceding sections establish both the representational gap and the kinds of linguistic information that must remain distinct. This section translates those requirements into a property-graph resource architecture in which orthographic, phonological, phonetic, grammatical, speaker, and provenance information can be linked without being collapsed into a single annotation layer. The architecture presented here is the authors' proposal. It builds on GrAF/LAF (Ide and Suderman, 2007; International Organization for Standardization, 2012) and prior database models, but the schema and the teaching module are our own contribution; no existing system by NCHLT or other projects is claimed.

4.1 Minimum Linked Layers

A corpus is a principled sample of language use; a dataset is a bounded collection prepared for analysis or modelling; and a database is a managed, queryable system linking observations through an explicit schema. The resource chain we propose is: *linguistic database* → *documented query and sampling procedure* → *versioned dataset* → *model and evaluation result*. Fixed datasets remain essential for comparable experiments, but they function as reproducible outputs rather than complete representations of a language.

4.2 A Graph-Theoretic Foundation

Layers alone do not encode how linguistic units at different strata interact. A tone does not merely sit in a "phonology" layer; it is simultaneously linked to a specific syllable, a morpheme that conditions its behaviour, an acoustic contour in the signal, and a speaker whose production may vary. Capturing these multi-dimensional dependencies requires a formal model of relations, not just categories.

The Graph Annotation Format (GrAF), introduced by Ide and Suderman (2007), provides a formal foundation for representing linguistic annotations as interconnected graphs rather than isolated labels. GrAF was developed as part of the ISO Linguistic Annotation Framework (LAF) (International Organization for Standardization, 2012) and enables the application of graph traversal and analysis algorithms to annotated language data.

We propose a property-graph database architecture in which linguistic objects—utterances, words, morphemes, syllables, phonemes, tones, vowel-

Layer	Core information	Purpose
Signal	waveform, channel, timestamps	ASR[4], TTS[5], acoustic analysis
Text	original and normalised orthography	search and text modelling
Structure	words, morphemes, syllables, alignment	morphological and prosodic modelling
Phonology	phonemes, underlying/surface tone, domains	phonology-aware representation
Phonetics	F0, duration, F1/F2, confidence	tone and vowel-quality analysis
Context	grammar, meaning, discourse	form–function disambiguation
Speaker	age, region, variety, language use	variation-aware sampling
Governance	provenance, consent, access, annotation history	auditability and responsible reuse

Table 1: Minimum linked layers in the database-centred resource model.

quality events, prosodic domains, speakers, and annotations—are represented as nodes, and their interactions are represented as typed, directed edges. The term *property graph* is used here in its technical database sense, and is not equivalent to the relational data model: relational systems use SQL and normalised tables, property-graph systems use query languages such as Cypher or Gremlin, and RDF systems use SPARQL. The choice of a property graph is motivated by the need to represent multiple competing analyses and typed relations among linguistic strata, following the graph-based approach of GrAF.

Formally, let the database be a directed, labelled property graph $G = (V, E, \Sigma, P)$, where V is the set of nodes, $E \subseteq V \times \Sigma \times V$ is the set of labelled edges, Σ is the set of edge types (relations), and P assigns key–value properties to both nodes and edges. This structure supports several relations that are difficult to preserve in a flat corpus.

First, it preserves multiple competing analyses. The orthographic string *o* from the utterance *o batlang?* becomes a node; it connects via a HAS_ORTHOGRAPHIC_FORM edge to that string. Crucially, it also branches via HAS_PHONOLOGICAL_INTERPRETATION edges to two distinct phonological nodes: one representing the second-person singular subject marker (associated with a specific tonal pattern) and another representing the class-1 third-person subject marker. Rather than forcing a single adjudicated label, the graph allows both candidates to coexist, with edges carrying probabilities, confidence scores, or annotator IDs. This directly addresses the problem of representing variation and uncertainty, as raised by Roux and Wissing (2017).

Second, graph traversal recovers information collapsed by orthography. A query can follow a path from a grapheme node (*e* or *o*), through its REALISES_PHONEME edges, to distinct vowel-quality nodes, each annotated with acoustic properties (formant contours, duration). From there, the traversal can continue to the lexical or grammatical node that the vowel distinguishes. This turns the database into a generative machine for resolving underspecification.

Third, graph theory provides a natural formalism for prosodic and tonal processes that operate across domains. In Sesotho, high-tone spreading, delinking, and downstep are not confined to individual syllables; they apply across prosodic domains such as the phonological word or phrase (Khoali, 1991). In the graph, a PROSODICDOMAIN node connects via DOMAINS edges to its constituent syllables and via CONDITIONS edges to specific tonal rules. A query can thus traverse from a lexical tone at a morpheme, through its prosodic domain, to the surface tone on a distant syllable—making the relation between underlying and surface forms explicit and machine-queryable. This is impossible with a flat corpus of orthographic strings.

4.3 Key Relation Types

The schema formalises key relation types, including but not limited to those in Table 2.

The graph is therefore not merely a storage format but a knowledge representation in which linguistic relations remain explicit and queryable. It allows computational datasets to be generated from selected paths through the resource while preserving the richer analyses, provenance, and uncertainty from which those datasets were derived.

4.4 Implementation and Interoperability

Candidate implementations include property-graph databases such as Neo4j with Cypher, and RDF stores queried with SPARQL where semantic interoperability is the priority. A relational implementa-

Relation	Source Node	Target Node	Purpose
SPEAKS	Speaker	Utterance	Links production to demographic context
HAS_ORTHOGRAPHIC_WORD	Utterance	Word	Segments the written string
HAS_MORPHEME	Word	Morpheme	Encodes morphological structure
ALIGNED_TO	Word / Phoneme	Signal interval	Time-stamps annotations to audio
HAS_UNDERLYING_TONE	Morpheme / Syllable	Tone	Stores lexical/grammatical tone
HAS_SURFACE_TONE	Syllable	Tone	Stores realised pitch contour
REALISES_AS	Underlying Tone	Surface Tone	Encodes rule application (e.g., spreading)
HAS_VOWEL_QUALITY	Phoneme	VowelQuality	Stores F1/F2/duration from acoustics
BELONGS_TO_DOMAIN	Syllable	ProsodicDomain	Encodes prosodic constituency
ANNOTATED_BY	Any node	Annotation	Tracks provenance, date, and annotator

Table 2: Key relation types in the graph-based database schema.

tion remains possible through node and edge tables in a system such as PostgreSQL, but the primary schema is a property graph, and any relational encoding is a mapping of that schema rather than its definition. Table 3 gives this correspondence: each node label maps to a table, each edge type to an edge table with source and target foreign keys, and node and edge properties to columns or JSONB fields. The worked example below is specified in property-graph terms; equivalent SQL and Cypher formulations can be derived from Tables 1–3.

For serialisation and exchange, JSON-LD or RDF/Turtle supports semantic interoperability, GraphML supports graph interchange, and CSV or CoNLL-U-style exports support downstream natural language processing (NLP) pipelines. The schema is also designed to be exportable to LAF/GrAF (Ide and Suderman, 2007; International Organization for Standardization, 2012), ELAN and CoNLL-U. We do not claim to invent multilayer annotation: LAF/GrAF, UIMA, ELAN, and CoNLL-U already provide substantial infrastructure for representing and exchanging layered annotations. Our contribution is a domain-specific tone and vowel-quality schema, together with a database-centred generation and teaching pipeline, that can be implemented on top of these standards rather than in place of them.

4.5 A Worked Graph Example: *o batlang?*

The ambiguity introduced in the Introduction can now be represented computationally. For *o batlang?*, the graph preserves the second-person singular and class-1 third-person analyses as distinct but linked interpretations of the same orthographic form.

Figure 1 illustrates a fragment of the graph for this utterance. The utterance node `U1` connects via `HAS_ORTHOGRAPHIC_WORD` edges to the word nodes `W1` (*o*) and `W2` (*batlang*). The word `W1` branches via `HAS_MORPHEME` edges to two morpheme nodes: `M1` (subject marker, class 1, underlying tone H) and `M2` (subject marker, 2nd person singular, underlying tone L). Each morpheme node connects via `HAS_UNDERLYING_TONE` edges to the appropriate tone nodes.

From the underlying tone nodes, `REALISES_AS` edges encode the tonal rules that apply. For the class-1 subject marker, high-tone spreading may apply across the prosodic domain, yielding a surface high tone on the following syllable. For the second-person singular, the low tone is realised as surface low. These surface tones are stored as nodes linked via `HAS_SURFACE_TONE` edges to the

Graph construct	Relational mapping	Example
Node label (e.g., Morpheme)	Table	`morpheme(id, props)`
Node property	Column or JSONB field	`props->'tone'`
Edge type (e.g., HAS_UNDERLYING_TONE)	Edge table with source/target FKs	`has_tone(src, tgt)`
Edge property	Column on edge table	`confidence, annotator_id`

Table 3: Mapping between the property-graph schema and a relational encoding.

relevant syllable nodes.

Crucially, the graph also links to the acoustic signal. The utterance node `U1` connects via `ALIGNED_TO` edges to phoneme nodes, which in turn connect via `HAS_VOWEL_QUALITY` edges to vowel quality nodes containing formant measurements (F1, F2, duration) from the actual audio. This allows a researcher to query not only the phonological interpretation but also the phonetic realisation.

Figure 1[6] visualises this structure. The graph makes explicit what the orthography hides: multiple phonological analyses, their rule-governed surface realisations, and their acoustic correlates—all in a single, queryable representation.

The worked example shows how one orthographic form can remain linked to competing grammatical analyses, their tonal representations, and the corresponding acoustic evidence. A `PROSODICDOMAIN` node (omitted from the figure for readability) can additionally connect the syllables over which high-tone spreading applies, preserving the relation between underlying tone, prosodic structure, and surface realisation.

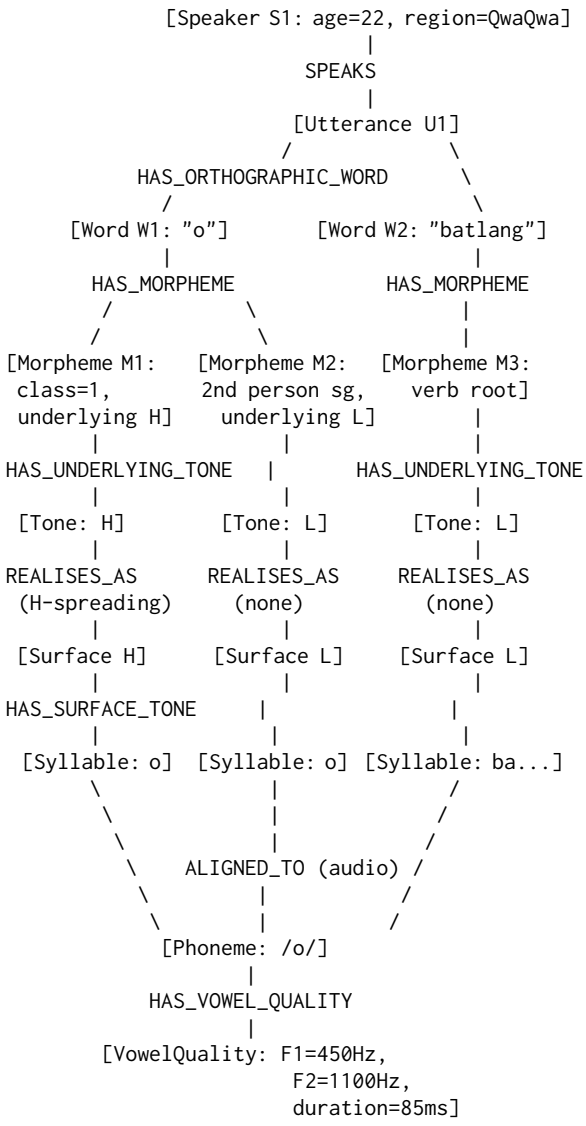

Figure 1: A fragment of the proposed graph for the utterance *o batlang?*. The graph preserves two competing morphological analyses (class 1 vs. 2nd person singular), each with distinct underlying tones and surface realisations, while simultaneously linking to the acoustic signal.

4.5.1 Generating Task-Specific Representations

The value of the graph lies not in claiming that a database can infer distinctions for which no evidence exists, but in allowing available evidence to be assembled into task-specific representations. A query can traverse from an underspecified grapheme such as <e> to alternative phoneme or vowel-quality analyses, their acoustic measurements, and the lexical or grammatical contexts in which they occur. The resulting dataset can therefore expose only the layers required for a particular task while retaining links to the fuller linguistic analysis and its provenance.

[6]Note: Formant values are illustrative values only. They do not represent actual acoustic measurements from Sesotho speech data.

5 An interdisciplinary teaching module

To transmit the academic knowledge about linguistic tone which is a necessary prerequisite to provide tone-labeled data for the development and training of database-centered systems, we developed an interdisciplinary teaching initiative. The course is part of the MA program "Theoretical and Comparative Linguistics" at the University of Stuttgart in the module "Tone and intonation".[7]

5.1 The linguistics curriculum

Sotho-Tswana has a two-tone system common in Bantu languages, with H being the active tone. Tone is relevant at the lexical and grammatical level, tone rules apply within phrases (see Cole, 2000 for an extensive description of the tone system), and there are a limited number of intonational pitch changes (Zerbian, 2017).

After an introductory session on the definition and notation of linguistic tone, its distribution across the languages of the world, and the linguistic landscape in South(ern) Africa, the course will in eight sessions

- provide an overview of the world's tone systems and the theoretical approach, using the textbook by Yip (2002),
- introduce students to the theoretical foundation of linguistic tone in the framework of Autosegmental Phonology (Goldsmith, 1990),
- train students' analytical skills on linguistic tone data sets, e.g., Odden (2005), chapter 10,
- introduce the phonetics of tone (e.g., Zerbian and Barnard, 2009), its acoustic cues, and its visual presentation in Praat,
- familiarize students with the basic morphological structure of Sotho-Tswana as it pertains to relevant tonal distinctions,
- illustrate the lexical and grammatical functions of tone in Sotho-Tswana by means of tonal minimal pairs,
- exemplify tone rules in Sotho-Tswana,
- address the interaction of word-, phrase- and sentence-level prosody (e.g., Zerbian, 2017)
- sensitivize the students to issues of multilingualism and tonal change (e.g., Roux and Wissing, 2017).

5.2 Tone annotation used in AI literacy

Universities must integrate AI literacy into the curricula of all study programmes to prepare their graduates for the job market, also in the Humanities. Therefore, the foundational research on linguistic tone will be contextualized in the field of speech and language technologies.

Linguistic tone is used as a test case to demonstrate the general logic of machine learning and its limits. A brief introduction to machine learning is available through the university self-learn course "AI and you: fit for studies" [8]. An algorithmic approach to tone for a text-to-speech system is presented through a literature review (Raborife et al., 2015). The software *Teachable Machine*[9] is trained on tone to test the challenges of tone recognition for speech-to-text systems.

For assessment, the students will use the linguistic knowledge acquired in the foundational eight sessions of the course and train their skills to annotate speech data for tone, drawing on lexical and grammatical information on tone from grammars (e.g., (Cole, 2000)) and dictionaries, visual inspection and auditory impression of speech, and their acquired knowledge about tone sandhi in connected speech. A first test corpus consists of real speech data, recorded from 6 participants in Soweto and the Free State, which comprises minimal pairs and short phrases. As a methodological standard in the field, interrater agreement will be calculated across different annotators to address the reliability of the tone labels.

The annotation task in this course thus responds directly to the representational gap described above. The Sesotho expression *o batlang?*, for example, illustrates how the same orthographic sequence can correspond to different grammatical interpretations depending on tonal realisation (Roux and Wissing, 2017). While rule-based approaches can predict some tones, they do not preserve acoustic realisation, prosodic domains, or annotator uncertainty (Raborife et al., 2015). The assessment therefore requires students to integrate lexical, auditory, acoustic and rule-based evidence to produce reliable annotations.

[7] https://www.uni-stuttgart.de/en/study/study-programs/Theoretical-and-Comparative-Linguistics-M.A./

[8] https://www.student.uni-stuttgart.de/digital-services/skillup/ki-und-du/

[9] https://teachablemachine.withgoogle.com/

Furthermore, a list of constructions will be compiled during the foundational linguistic sessions for which further speech data need to be collected such that a small language model will have sufficient breadth of tonal information available to learn in a hybrid approach.

The course will be taught for the first time in the winter term 2026/27. Further speech data can be transcribed as part of the course in subsequent academic years.

6 Limitations

Manual annotations have always been and are still time- and resource-costly. This also holds true for the tone labels created as part of the final assessment for the course as proposed in this paper and surely is one limitation with respect to system development.

Another limitation, and at the same time also a factor that adds to the costs, is the limited availability of digital tone-marked resources for Sotho-Tswana, e.g. tone-marked dictionaries, or grammars.

A further limitation is the bootstrapping problem inherent in the proposed architecture: populating the graph with phonological and tonal annotations requires expert linguists, but the scarcity of such experts is precisely the issue the database aims to mitigate. In the short term, we address this through the teaching initiative described in Section 5, which will gradually generate annotations. In the long term, the database itself can be used to train weakly-supervised models that predict annotations for new data, reducing the manual bottleneck over time.

Furthermore, the transferability of this approach to other Sub-Saharan tone languages is not guaranteed. Sesotho is among the more well-documented Bantu languages, with existing resources such as tone-marked dictionaries and grammatical descriptions of tone. For the vast majority of these languages, which are critically under-resourced, these foundational resources first need to be developed. Our approach, therefore, represents a best-case scenario; for many other languages, the initial step must be basic linguistic documentation before the database architecture proposed here can be meaningfully applied.

Finally, the complexity of the proposed database architecture introduces a significant implementation overhead compared to simpler, flat corpus formats. There is an inherent trade-off between the richness and queryability of the data and the ease of its creation and use. This paper argues that for the specific case of under-resourced languages, the scientific and practical benefits of the richer representation justify the initial investment in development and annotation.

7 Conclusion

This paper argued that resource scarcity for Bantu languages is not only a quantity problem but a representational-adequacy problem. Using Sesotho tone and mid-vowel quality, we showed that orthography systematically underspecifies distinctions that are contrastive in speech: tone is unmarked, and five vowel letters represent a system with more vowel qualities than the orthography marks. Roux and Wissing (2017) document variation in the functional use of tone in the *o batlang?* contrast, and the orthographic underdifferentiation of mid vowels compounds the problem for computational modelling. Increasing the volume of orthographic data cannot directly recover distinctions that writing leaves unspecified, even where individual tokens are contextually disambiguable. We proposed a property-graph, multimodal, versioned database architecture, following GrAF/LAF, that preserves competing analyses, provenance, uncertainty, and acoustic evidence, and we outlined an interdisciplinary teaching module intended to build annotation capacity. The contribution is a schema and a research programme, not a completed database. Future work will implement a prototype, evaluate inter-annotator agreement on tone labels, and test interoperability with LAF/GrAF, ELAN, and CoNLL-U.

Acknowledgments

Sections 4.2 and 4.3 of this contribution were prepared using generative AI-systems co-creatively. The following systems were used: Deepseek in their current model versions. The conception of content, selection, evaluation and editorial responsibility are with Mpho Primus and Sabine Zerbian.

References

I. Adebara, A. Elmadany, M. Abdul-Mageed, and A. Alcoba Inciarte. 2023. SERENGETI: Massively multilingual language models for africa. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1498–1537.

D. I. Adelani and others. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508. Association for Computational Linguistics.

E. Barnard, M. H. Davel, C. van Heerden, F. de Wet, and J. Badenhorst. 2014. The NCHLT speech corpus of the south african languages. In *Proceedings of the 4th Workshop on Spoken Language Technologies for Under-resourced Languages*, pages 194–200.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing. *TACL*, 6:587–604.

Desmond T. Cole. 2000. *An introduction to Tswana grammar*. Longman, Cape Town.

Dafydd Gibbon and Thorsten Trippel. 2000. A multi-view hyperlexicon resource for speech and language system development. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

John A. Goldsmith. 1990. *Autosegmental and metrical phonology*. Blackwell, Oxford.

Joseph H. Greenberg. 1948. The tonal system of proto-bantu. *Word*, 4(3):196–208.

Hannah J. Haynie, Damián E. Blasi, Hedvig Skirgård, Simon J. Greenhill, Quentin Douglas Atkinson, and Russell David Gray. 2023. Grambank's typological advances support computational research on diverse languages. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*.

N. Ide and K. Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, pages 1–8, Prague, Czech Republic.

International Organization for Standardization. 2012. *Language resource management — Linguistic annotation framework (LAF)*. Geneva, Switzerland.

B. T. Khoali. 1991. *A Sesotho tonal grammar*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Johann-Mattis List et al. 2025. Lexibank 2: Pre-computed features for large-scale lexical data. *Open Research Europe*, 5.

K. R. Mabokela, M. Primus, and T. Celik. 2025. Advancing sentiment analysis for low-resourced african languages using pre-trained language models. *PLOS ONE*, 20(6):e0325102.

Alexander Magidow. 2015. A relational database model and prototype for storing diverse discrete linguistic data. *Journal for Language Technology and Computational Linguistics*, 30(1):27–45.

M. R. Marlo. 2013. Verb tone in Bantu languages: Micro-typological patterns and research methods. *Africana Linguistica*, 19:137–234.

S. H. Muhammad and others. 2023. AfriSenti: A twitter sentiment analysis benchmark for african languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981. Association for Computational Linguistics.

David Odden. 1988. Predictable tone systems in bantu. In *Autosegmental studies on pitch accent*, pages 225–251.

David Odden. 2005. *Introducing Phonology*. Cambridge, Cambridge University Press.

Mpho Primus. 2025. An inclusive script for the future of ai. *Nature Africa*.

Mpho Raborife, Sigrid Ewert, and Sabine Zerbian. 2015. Improving a tone labeling algorithm for sesotho. *Language Resources and Evaluation*, 49(1):19–50.

Justus C. Roux and Daan Wissing. 2017. The status of tone in sesotho: A production and perception study. *Nordic Journal of African Studies*, 26(4).

Moira Yip. 2002. *Tone*. Cambridge, Cambridge University Press.

Sabine Zerbian. 2017. Sentence intonation in tswana (sotho-tswana group). In Laura J. Downing and Annie Rialland, editors, *Intonation in African Tone Languages*, pages 393–434. De Gruyter.

Sabine Zerbian and Etienne Barnard. 2009. Realisations of a single high tone in northern sotho. *Southern African Linguistics and Applied Language Studies*, 27(4):357–380.