MODAN: An Ontology-Motivated Model for Annotation of Noun Class Systems to Enhance Meaningful NLP

C. Maria Keet[1,2,*], Zola Mahlaza[3] and Zubeida C. Khan[4]

[1]Computer Science Division, University of Stellenbosch, c/o Banghoek and Joubert Street, Stellenbosch, South Africa
[2]Meaningfy S.à r.l., 61 Route the Fischbach, Lintgen L-7447, Luxembourg
[3]Department of Computer Science, University of Cape Town, 17 University Avenue, Rondebosch, Cape Town, South Africa
[4]Council for Scientific and Industrial Research, Pretoria, Meiring Naude Road, Brummeria, Pretoria, South Africa

Abstract

Natural Language Processing (NLP) for African languages, specifically for tasks such as Machine Translation and Human-Robot Interaction, may require knowledge-infused approaches. In such contexts, especially for Niger-Congo B (NCB) languages, representation of noun class agreement in the sentences is crucial to ensure grammatical correct output. However, there is no linguistic ontology to support such work. A foundational ontology could assist NLP tasks, but existing foundational ontologies do not include linguistic entities and the key desiderata only very partially; hence, it was not possible to use them either. To solve this problem, we developed a new logic-based model (MODAN) that may be used directly or as a module of a linguistic ontology, to accurately represent and annotate NCB noun class systems for better computational use, and therewith the nouns. We evaluated MODAN against noun class systems in different language families and found that it covers most of those noun class semantics as well, therewith increasing the prospect of improving NLP systems well beyond the African languages.

1. Introduction

Natural Language Processing (NLP) systems rely on various types of encoded knowledge to process human language. Traditional linguistic representations are generally categorized into syntax and semantic approaches. Syntax-based representations, such as parse trees and grammars, define the structural composition of sentences; semantic representations focus on meaning and resources include lexical databases such as WordNet [1] and various ontologies that map relationships between nouns, verbs, adjectives, and adverbs (e.g., [2]). These distinct models may be integrated or used with frameworks such as the ontolox-lemon model [3]. Although such efforts might appear obsolete due to the popularisation of LLMs, they remain relevant to improve LLM output thanks to the structured nature of the knowledge and the LLOD cloud. For instance, to generate knowledge graphs or build semantic classification models—advanced ones that that can distinguish between, e.g., *umZulu* 'a Zulu person' (noun class (nc):1) vs. *ubuZulu* 'Zulu identity' (nc:14) even though they use the same noun stem *-Zulu*—for tasks that require high-quality accurate output and to enhance NLP for those languages with limited LLM capabilities [4], which include many languages in Africa, Latin America, and Asia.

Given this context, we note that existing linguistic ontologies face limitations when dealing with intricate noun class systems prevalent in the aforementioned languages: i) existing linguistic ontologies are now several years old, potentially missing recent advancements in the field; ii) the current ontologies, such as GOLD [5] and OLiA [2] lack content about the specific linguistic features of NCB languages [6]; iii) also the linguistic annotation models have shown to be inadequate for at least the core Niger-Congo B (NCB) language features [7]. Further, even the cognitive and linguistic motivated DOLCE [8] contains semantic categories that do not adequately capture noun classes' semantics of NCB and other languages

Proceedings of FOIS 2026 Satellite events co-located with the 16th International Conference On Formal Ontology in Information Systems (FOIS 2026), 21–25 September 2026, Vitória, Brazil
*Corresponding author.
cmkeet@sun.ac.za (C. M. Keet); zmahlaza@cs.uct.ac.za (Z. Mahlaza); zdawood@csir.co.za (Z. C. Khan)
http://www.meteck.org (C. M. Keet); http://https://people.cs.uct.ac.za/~zmahlaza (Z. Mahlaza)
0000-0002-8281-0853 (C. M. Keet); 0000-0001-9829-1480 (Z. Mahlaza); 0000-0002-1081-9322 (Z. C. Khan)
© 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 2026-11-27

CEUR-WS.org/Vol-4342/paper6.pdf

(discussed further below); e.g., nouns for animals are found in nc:9 and nouns for long thin objects are found in nc:11 in isiZulu, neither of which easily links to DOLCE categories.

To address these limitations and enhance the representation of noun class systems in NCB languages, and prospectively all languages that use a noun class system, we propose the development of a dedicated module for the linguistic ontologies to capture the required level of detail of noun class systems. We explored the suitability of foundational ontologies, given that a key challenge is the lack of categorization of noun classes in existing ontologies, but eventually resorted to a bottom-up approach to developing the **M**odel for **O**ntology-motivate**D** **A**nnotation of **N**oun classes, MODAN.

We first compared existing linguistic ontologies such as GOLD and OLiA with foundational ontologies to identify areas where there may be overlap and gaps, to understand the limitations of current foundational ontologies in handling the semantics of noun class systems. This necessitated further investigation, harvesting potential content from the linguists literature, ontology-motivated modelling, and subsequent evaluation of the model. The evaluation of the model focused on adequate coverage. We assessed MODAN against the documented linguistic literature of three geographically diverse languages that all have a semantics-informed noun class system. We have found that MODAN is able to wholly capture the Nepal Bhasa and Ojibwe noun class systems, whereas for Q'anjob'al it depends (6/6 for domains, 11 of the 13 of the Santa Eulalia dialect semantic categories), and of course all of the features of the nominal classification by referent features of the proto-Bantu family that was part of the design process.

In the remainder of this paper we first illustrate relevance to motivate the issues and questions more precisely and provide minimal necessary subject domain background (Section 2), and related work pertaining to ontologies is discussed in Section 3. The main part of the paper describes the bottom-up development of MODAN and its content (Section 4), and subsequent evaluation against noun class systems of three languages, two of which are from other continents, to assess coverage of MODAN (Section 5). We discuss in Section 6 and conclude in Section 7.

2. Use Case Examples and Noun Classification Systems

In this section, we motivate the problems and provide subject domain background.

2.1. Demonstrative Examples

An ontologically relatively straightforward case is the difference between one noun class for nouns and their referents that are 'humans and other animates' and another for 'kinship terms and proper names', which is exemplified in nc:1 and nc:1a in isiZulu, respectively[1]. The surface use in the singular is similar as well—*umfana udla* 'the boy eats' and *ubaba udla* 'the father eats'—but the plural of *umfana* is *abafana* (plural marker underlined for explanation) because it is in nc:2, yet it is *obaba* 'fathers' because it is in nc:2a. A pluraliser, be it a dedicated one or a language model-based one, needs to know the noun class of the noun so that it will not generate the incorrect *abababa*[2] when it cannot distinguish between the human and the roles they play.

Concerning artefacts, in the isiZulu phrase *intombazane idlala ngenduku* 'the girl plays with a stick', one can use BFO to capture that a stick is a bfo:material entity, and bfo:object more specifically. However, BFO does not offer the option to categorise that a stick, or sticks, in isiZulu belong to the group of entities that are tool(s) whose nouns are all assigned to nc:9/10, depending on whether it/they are singular or plural. As a consequence, a language model may generate incorrect sentences. For instance, a language model might generate sentences like *intombazane idlala ngetshe* 'the girl plays with stone' and *intombazane idlala ngenduku* 'the girl plays with a stick' with equal weight[3] when relying on BFO

[1]although this can be open to debate and refinements, such as humans versus other animates, and what sort of roles are 'permitted' in nc:1a: *ubaba* 'father' and *udokotela* 'doctor' are, but not *ingwenya* 'criminal' that is in nc:9 where typically animals are classified.

[2]While *aba baba* 'these fathers' is a valid demonstrative phrase, *abababa* is not a valid plural form.

[3]This is subject to the frequency of the various tokens in the dataset, of course

to determine the nature of the two phrases' subjects since it will conclude that they are both only material entities, even though in isiZulu they fall under different noun classes (*-tshe* 'stone' is in nc:5 and *-nduku* 'stick' in nc:9).

2.2. Noun Classification Systems

A prerequisite to the resolution of the problems described in the previous section requires that one understands how nouns are categorised so as to pin down the notion of a noun classification system that can be of use across a number of languages, prior to any consideration of model development and its machine-readable use in NLP tasks. In this section, we will briefly describe the domain for languages found in Africa, North America, and Asia, where languages are spoken that may use noun class systems much more explicitly than gender in Indo-European languages.

2.2.1. Variations in Noun Systems in the World

In Indo-European languages, the primary categorization method, if syntactically present, is based on the notion of masculine/feminine/neuter gender, with minor features like definiteness also being used at times. There may be a latent semantics or morpheme to indicate the gender, such as *-ung* in German and *-ción* in Spanish signalling feminine gender of the noun (e.g., *Werbung* 'advertisement', *educación* 'education').

Expanding the scope to languages from other continents, one finds other features suitable for constructing noun classification systems. For example, NCB, East Asian, and Mayan languages, include the concepts of noun classes and classifiers. These notions are discussed slightly differently in the linguistics literature. Scrutinising work that discusses noun class systems, it shows that each noun belongs to a distinct noun class, which forces other words within a sentence to modify their form based on the noun's noun class. For instance, in isiZulu, the verb *hamba* 'go' takes different forms: *iyahamba* when, e.g., a dog (in nc:9) is going and *uyahamba* when a person (in nc:1) is going. In this example, the two nouns belong to different noun classes and each noun class dictates the value of the verb's prefix. It has similar effects on adjectives and counting numbers, among others.

Conversely, some languages employ optional or context-dependent classifiers for numerical classification of nouns. This can be seen in Ojibwe, a language that is indigenous to North America, where the word *-aatig* 'wooden object' leads to various word forms: *Bezhigoogwaatig* 'one wooden object' vs. *Niiwaatig* 'four wooden objects'. Despite the professed differences in the two kinds of systems, there is a need to review existing work to understand how nouns are classified. A comprehensive generalisation of noun classification systems is described in [9].

2.2.2. Linguistic Systems

Noun classification for African languages is largely guided by the work that was focused on Proto-Bantu, the common ancestor to all NCB languages. As demonstrated by Mahlaza [10], in the case South African NCB languages, such work mostly focuses on categorising nouns into at most 23 noun classes. More recent work on more NCB languages demonstrated that the categorisation can be summarised as shown in Table 1, with some variations for each particular language. For instance, isiZulu also has nc:1a, nc:2a, nc:3a, and nc:9a and new foreign objects are initially allocated to nc:5 (e.g., *ikhompyutha* 'computer') [11]. This notion of noun classes is not universal wherever noun class systems are used; hence, we also need to consider other languages and approaches.

An outlier classification for 'Bantu' languages (NCB and NCA) is that of Denny and Creider [13] who, dissatisfied with the seemingly 'random' groupings like in Table 1, assessed the referents on their properties and which of those are used to classify the objects into the noun classes, as shown in Fig. 1. The terms they used do not necessarily mean what ontologists mean by it nowadays, some noun classes appear in more than one place, and a number of main noun classes are missing (12, 13, 16 and upwards to 23). Nonetheless, highlighting the entity's feature used for classifying the noun can work: the classifying feature for, e.g., fruits and body parts in nc:5 all can be indeed "rounded, protruded,

Table 1
Noun classification of NCB languages (Source: [12]), which pair up singular and plural for the classes up to 10.

Noun class(es)	Description of the type of referents of the nouns
1 and 2	People and kinship
3 and 4	Plants, nature, and some parts of the body
5 and 6	Fruits, liquids, some parts of the body, and paired things
7 and 8	Inanimate objects
9 and 10	Tools and animals
11	Long thin stringy objects, languages, and inanimate objects
12 and 13	Diminutives
14	Abstract concepts
15	Infinitives and parts of the body
16, 17, and 18	Locative classes (e.g., around something)
19	Diminutives
20, 21, and 22	Augmentatives
23	Locative class

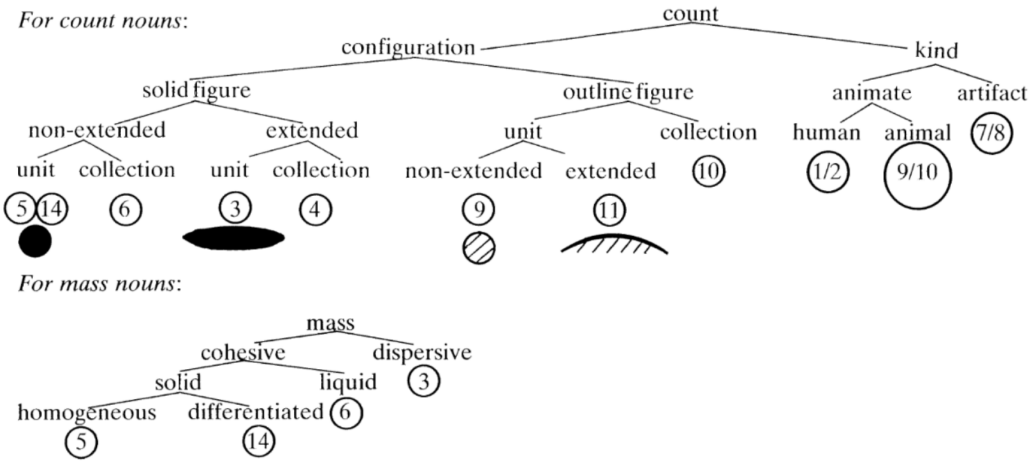

Figure 1: The first attempt of nominal classification by referent feature for proto-Bantu languages, the assumed common ancestor of the NCB languages spoken in Sub-Saharan Africa. The numbers are the identifiers for the noun classes. (Source: [13]).

humped, bunched, etc" as a feature they share. A further debate of different ways of classifying nouns for 'Bantu' languages can be found in Chapter 2 of [14], where it is asserted that for Niger-Congo languages, it is neither fully the noun nor its referent that determines the classification, but that one should rather take a "Cognitive Linguistics movement – namely concept, domain, profiling, construal and construction." approach, or, in ontological terms: classifying by a selected property of the referent. Watson's assessment for Kujireray (spoken in Senegal, Niger-Congo A family) has a mixture of concepts and features in what is labelled the "semantic domain" in its table 51, partially alike in Table 1 (parents, trees) for some classes and partially using features such as those in Fig. 1 (small, round, extended, mass) for other classes.

Turning to languages in Asia, Li [15] investigated a classifier system in Zauzou, a Sino-Tibetan language. While Zauzou uses both sortal and mensural classifiers, Li focused on studying the sortal classifiers. Sortal classifiers categorize nouns based on inherent properties such as shape and size, while mensural classifiers relate to quantity and measurement. The parameters used for classification included animacy, shape, size, and numbers; see also Fig. 2. The study identified three etymological subclasses of classifiers: repeaters, non-repeaters, and quasi-repeaters. Repeaters reuse part of the noun itself as a classifier; non-repeaters are independent words, and quasi-repeaters are a hybrid of the two. Wang and Walther [16] classified Chinese Mandarin using both mensural and sortal classifiers, who found that sortal and mensural classifiers are measured in different ways. This reinforces the idea that they do

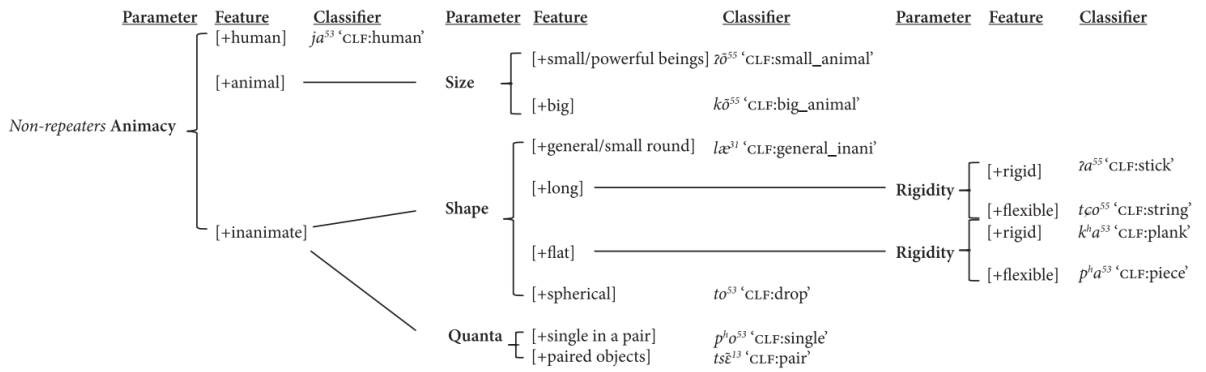

Figure 2: Example of nominal classification of Zauzou nouns (for the non-repeaters), a Sino-Tibetan language spoken in Southwest China. (Source: [15]).

need to be categorised as separate syntactic categories.

3. Assessment of Ontologies and Attendant Resources

Given the brief overview of noun class systems and noting noun classes for abstract entities, rigidity, masses and so on, it appears that a foundational ontology may be usable either directly as-is or useful for their desiderata. Desiderata are criteria that describe the essential characteristics of the content represented in an ontology where it branches off. Foundational ontologies such as BFO [17] and DOLCE [8] might be semantically suitable for use in NLP systems along the line of applying DOLCE to WordNet [18] or only its 2-3 top concepts [19], and match the sort of notions used to describe the semantics of the noun clsses. Therefore, we extracted the key desiderata as shown in Table 2 and Table 3, respectively, which was also informed by their decision diagrams [20, 21].

It is immediately evident that there will not be a good fit, lacking elements that can align directly with aforementioned 'long thing stringy objects' or (non-human) animals, although they may be aligned with subclass axioms, such as spherical as being related to bfo:three-dimensional region and animal as a subclass of dolce:physical object or bfo:object. We will return to the suitability in detail in the evaluation in Section 4.3.

Given that a number of categories of the noun's referents are more concrete—at the level of domain ontologies from the viewpoint of a foundational ontology—domain ontologies for linguistics might be a better fit, or a combination where a linguistic ontology is aligned to a foundational ontology to avail of the best of both. Principally, there are the Generalised Upper Model (GUM) [22, 23], General Ontology for Linguistic Description (GOLD) [24, 5], and OLiA [2], and an extension to GOLD for noun classes [6]. GUM predates ontology development and "is a hierarchy of concepts that captures semantic distinctions necessary for generating natural language" [22] and is based on grammar [23], which appears close to our initial motivations for this work given that noun classes act out in the grammar. Its contents is a mixture of generic categories, specific ones, and linguistic ones, such as Process and SayingAndSensing and it has not been updated since the 1990s.

GOLD [24, 5] uses SUMO [25] as an upper ontology that, like DOLCE, also has links with WordNet [26], and it has an extension with the NCS task ontology for noun classes [6]. GOLD is a domain ontology of linguistic elements, of which the relevant one here is Gender. It includes 9 subclasses: animate, Arabic number, feminine, human, inanimate, masculine, neuter, Roman numeral, and vegetable gender. Noun classes are, arguably, neither genders not mere Arabic numerals, and therefore a bantuNounClass sibling class was added with its noun class numbers and the singular/plural pairings where applicable [6]. Neither GOLD's few subclasses nor the numbers capture the semantics of those and other noun class systems.

The Ontologies of Linguistic Annotation (OLiA) [2] consists of a collection of modules and is not aligned to a foundational ontology. It is aimed at providing terminology for tasks such as corpus

Table 2
Key desiderata between sibling entities in the BFO v2 taxonomy.

Desideratum	Description	Example
Temporal Persistence	Determine if an entity persists over time (Continuant) or occurs in time (Occurrent)	A ball vs. kicking
Material vs Immaterial	Identify if an independent continuant has material parts that persist in time or if it is immaterial	A coffee mug vs. a cavity in a tooth
Boundedness and Dimensions	Determine if an immaterial entity is bounded by a material entity and its dimensional properties	The cavity of a tooth (Bounded, 3D) vs. a point in space (Unbounded, 0D)
Dimensions of Regions	Classify spatial and temporal regions and their dimensions	The space occupied by a bed (3D Spatial Region), the moment the electricity switches off (0D Temporal Region)
Objects and their parts vs members	Differentiate between collections of disjoint self-standing objects and parts of objects	A pack of lions (Aggregate) vs. upper portion of a cake (Fiat Object Part)
Dependent Continuants	Determine if a dependent continuant can be shared or is specifically dependent on a bearer.	An mp4 video file (GDC) vs. the role of a reader (SDC).
Realizable Entities	Evaluate if a specifically dependent continuant requires a process to be realized.	A researcher has to do research (Realizable Entity) vs. the color of a berry (Quality).
Roles, Dispositions, and Functions	Identify if the entity's existence is tied to a specific situation, has a disposition, or realizes specific processes.	A professor (Role) vs. a pen's function to write (Function).
Relational Qualities	Check if a quality requires at least two bearers to exist.	The family history (Relational Quality)

annotation, use in NLP, and for enriching machine-readable dictionaries, and thus in the domain of application in line with our motivation, and is available in multiple languages. It is, however, purely focussed on linguistics, not the domain semantics. Regarding noun classes, it has the olia: gender feature with seven subclasses in the OLiA Reference Model for morphosyntax, morphology and syntax module: animate, common, feminine, inanimate (with vegetable as subclass), masculine, and neuter, informed by the aforementioned GOLD. The annotation alludes to the fact that there is an underlying semantics, but does not include it in OLiA.

4. Development of MODAN

Given the increasing usage of NLP tools, especially within LLMs and contextual answers, we note that current foundational ontologies lack sufficient relevant categories for typical domain entity-level

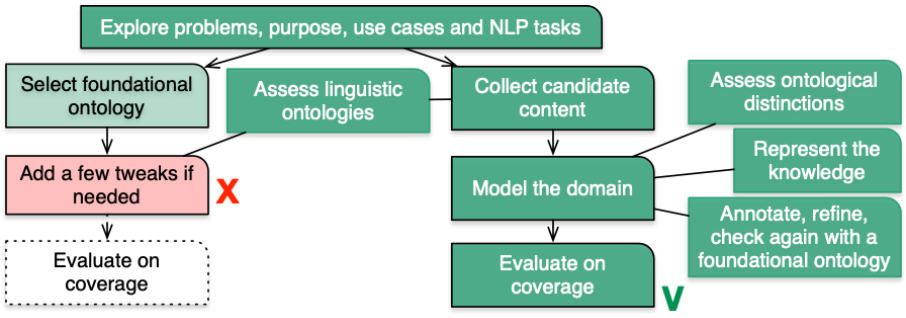

Figure 3: Summarised sketch of the processes followed, with on the left the *de facto* aborted attempt, instead resorting to a process akin to a laborious bottom-up approach.

Table 3
Key desiderata between sibling entities in the DOLCE taxonomy.

Desideratum	Description	Example
Presence, or not, in whole vs part or dependent	Determine if the entity is abstract, inheres in another entity (Quality), or is wholly present (Endurant) or partially present (Perdurant) at a point in time	Set, Colour, Ball, and Kicking, respectively
Participating in vs being	Event vs State	
Atomicity of the event and definite endpoint	Determine whether the perdurant has distinct subdivisions (accomplishment) and demarcated end point (achievement)	attending a conference vs the moment you reach the top of the Kilimanjaro
Sameness of parts	Determine whether the parts of the state are of the same type of process (state) or different processes (process)	Sitting vs Running
Dependence on physical, temporal, or abstract entity	Quality of what type of entity	colour, duration, value
Ways of being abstract	mathematical (Set) vs statements about something (Fact) vs possibly measured location of an entity (Region)	
Physicality in space	Determine whether the entity is concrete in space vs non-physical	Door, student
Dependence on a community of agents	Determine whether it exists in the mind of one (mental object) vs social setting (Social object)	dream, electorate
Dependence on physical host	Feature vs (physical object or amount of matter)	groove, dog
Countability	physical object vs amount of matter	dog, wood

annotations of noun classes in NCB Languages. The proposed semantic artefact is intended to fill in missing gaps regarding a particular linguistic feature, i.e., it focuses on the noun's entity categorisation and its properties, which is important for semantic interoperability.

The approach we took was as follows, which is also sketched in Fig. 3. In parallel with a literature search to assess ontologies in linguistics and relate them to a foundational ontology and gather desiderata for noun classes, we assessed BFO and DOLCE and the respective decision diagrams to try to make the desiderata for each branching explicit. We also created a first model for annotation that both includes the noun class properties from the literature and some additional structuring mechanisms inspired by ontological distinctions. The model was subsequently evaluated with a selection of languages that have noun classes and that are documented in the scientific literature, and its desiderata were compared against the desiderata used in BFO and DOLCE to assess similarity, or distinctness, of categorisation.

4.1. Noun Classes in Linguistics: Sourcing Domain Content

Aikhenvald's [9] recent review on noun classification systems proved to be useful as a starting point from which to collect desiderata, focussing on its section 12.2 on noun classes as noun classification mechanism, among the possible means of classifying nouns. A majority of concepts in the eventual model are sourced from this review, corroborated and augmented with concepts from other works, notably [13, 15]. Denny and Creider [13] attempted this specifically for "proto-Bantu", i.e., the language assumed to be the common ancestor to all languages in the Niger-Congo B language family. Its terminology (recall Fig. 1) is unique, however, and needs to be seen in conjunction with the descriptions in the paper and aligned to modern-day terminology. Li et al. [15] provided an exemplary categorisation for ZauZou, which is shown in Fig. 2.

Noun classes have an assignment of nouns into a noun class that can be based on semantic, morphological, and/or phonological criteria, noting ([9], p363) cross-linguistic properties of noun classes, and they involve some form of perceived dichotomies or contrasts, i.e., may be candidates for desiderata to specify subclasses in a taxonomy or ontology especially if they also hold ontologically. They are,

notably:

- sex or gender: masculine versus feminine (and possibly vs neuter vs vegetable vs residual). They may have additional shape and/or size-related properties embedded in it such that a noun ends up as belonging to that gender (e.g., objects are feminine if they are, e.g., short and round and masculine when long, tall, narrow)
- human versus nonhuman; human versus animal versus artefact
- rational versus non-rational
- animate versus inanimate
- countables versus masses (that are uncountable, or only countable in quantities)
- among the countables, kind versus configuration-based classification
- among the masses, cohesive versus dispersive (and solid vs liquid, and homogeneous vs differentiated)
- other sub-distinctions: rigidity of the object in the sense of rigid versus flexible (e.g., a stick and a string, respectively); unit versus collection; solid versus outline or hollow; paired objects versus single objects

In addition, a recurring set of attributive properties of objects are used, according to Aikhenvald's review, of which the following are collected from [9]: short, long, tall, wide, narrow, hollow, solid, round, spherical, flat, pointed, deep, thin, thick, large, small, and big.

These dichotomies and attributes are mostly different from common tables listing generalisations of the kind of objects in a particular noun class, such as for NCB [27] and other [9] languages, referring to domain entities such as pejoratives, endearment, body parts, cloth, precipitation, wind, minerals, games, types of metal, and knowledge. It is such listings that are subject to the most debate about whether there is an underlying semantics to a noun class system. To cover this across languages with a noun class system, it suggests one would need a domain ontology of everything, which defeats the purpose.

4.2. The Model: MODAN

The lightweight model we developed contains 60 classes, 1 object property (attributedTo), 77 logical axioms, and 33 annotation assertions; the OWL Classifier classified it in the $\mathcal{ALC}$ DL fragment (within OWL 2 DL). The first key distinctions made concern classifying nouns into noun classes by kind, by configuration, or it is a mass; see also Fig. 4. Among Kind, there are the three categories that also reflect ontological distinctions: physical objects (also called single object, unit), abstract objects (non-physical endurants), and collections as a parent class of convenience for sets, paired items, and collectives. Configuration—incorporating the term used by linguists—refers to classifying by attribute of the entity, which is either an abstract or a physical quality. PhysicalQuality has subclasses typical for qualities, such as shape, and augmentatives and diminutives, such as Tall and Small, that refer to sizes relative to something else. Its disjoint sibling AbstractQuality refers to abstract in the sense of non-physical, whose subclasses could be debated when taken out of context, but what they intend to denote here is the human conceptualisation irrespective of reality and scientific evidence, i.e., as concepts. That is, no ontological claims to notions of sex, animacy or rationality are made, merely an inclusion as linguistic property. Finally, the masses have only a few subclasses without clear ontological structuring (cf. the Stuff ontology [28]), but we could not find further linguistic distinctions being made and therefore higher precision did not merit inclusion.

The axioms concern constraints such as Abstract Quality $\sqsubseteq \forall$attributedTo.(Kind $\sqcup$ Mass $\sqcup$ Process) and Physical Quality $\sqsubseteq \forall$ attributedTo.PhysicalObject. It also includes the minimal evident disjointness constraints, such as between physical object, abstract object, and collective, and masculine, feminine, and neuter. Note that this does not mean that some object cannot be attributed to be masculine in one language and feminine in another (e.g., *der Mond* in German and *la luna* in Spanish, meaning 'the moon') or in one language to allow a noun to be both when they have a different referent (e.g., *le cave* 'idiot' and *la cave* 'basement' in French), or different language variants having classified it into

a different gender (e.g., *das E-mail* in Standard German versus *die E-mail* in Swiss-German), but the gender concepts themselves have no overlap in the linguistic sense. Labelling of noun classes can be done regardless.



Figure 4: Main classes of the annotation model; see text for explanation and the `.owl` file for annotations of the classes.

4.3. Assessment Against FOs

After the bottom-up 'restart' of developing the model, we wanted to re-assess whether foundational ontologies such as BFO and DOLCE would be semantically suitable after all. For this, we need to check whether the desiderata include elements of noun class systems' semantics of at least NCB languages.

As noted in Section 3, the linguistic ontologies were not aligned to foundational ontologies, therefore we use the desiderata presented in Table 2 for BFO and in Table 3 for DOLCE to check whether current foundational ontologies can represent linguistic features.

Regarding BFO, the table reveals that there is, as expected, no coverage in BFO for morphological and phonological properties as used in linguistics. Regarding the main classes of MODAN, there are a few similar entities, but they are contrasted not in the same way, albeit at times similarly. The most similar is BFO's object vs object aggregate vs fiat object part sibling entities versus collection vs physical object vs abstract object in MODAN, and a distinction between independent continuants and occurrents in BFO, and kind vs process in MODAN. Other dichotomies do not align, which is largely due to BFO's philosophical stance of excluding abstract things and excluding masses (also called stuff or matter), which feature prominently in MODAN, and the elaborate representation of qualities for which there is only one entity in BFO.

DOLCE's ontological commitments and content should fare better in similarity of desiderata, for it has more detail on abstract entities and qualities. Also in this case, only some desiderata can be found and in a limited way. There is a similar-looking top-level distinction between process (perdurant), object (endurant), and quality, except that DOLCE also has Abstract as top-level category, which is a subclass in MODAN, and, conversely, Amount of matter as a subclass of endurant whereas that is a top-level distinction in MODAN. Hence, the actual decision points themselves differ, but not necessarily the desiderata (e.g., countability is still a reason to divide). As to quality, while DOLCE has more types than BFO and the desideratum *de facto* being what category it inheres in, MODAN has more concrete subclasses and no qualities for processes, and a number of additional desiderata, such as by shape and by consistency, and augmentative and diminutive qualities. Thus, also the top-level ontology that is inspired by cognition and linguistics does not match well, even if DOLCE or MODAN would be slightly re-engineered.

5. Evaluation and Use

We evaluate MODAN's performance using its coverage of languages that were unseen as part of its development. Afterward we return to the problems from Section 2.1 and how our model might assist.

5.1. Annotatability

We selected three languages, indigenous to the Americas and Asia, to establish the extent to which MODAN covers the semantics of their noun classification systems.

5.1.1. Q'anjob'al Noun Class System

Q'anjob'al, a Mayan language, has nouns that are categorised based on semantic features especially for the inanimate objects and based on what they call "domain", such as animacy, humanness, and gender [29] that are in MODAN's scope, amounting to a total of 13 noun classifiers. The domain divides first by animacy, then among the animate, between human vs non-human, and among human, between male and female, thus all 6 are covered by MODAN. The inanimates can be annotated with PhysicalObject or a subclass thereof. The Santa Eulalia dialect of Q'anjob'al also has *xal* and *cham*, which describe older females and males [29], respectively, which are not covered, although older in the sense of pejorative or augmentative can be annotated; refer to Table 4 for details.

Table 4
Evaluation of MODAN with the Q'anjob'al Noun Class System and its Santa Eulalia dialect (Q'SE).

Classifier	Semantic Category	"Domain"	Covered in MODAN
ix	Female	Human (Animate)	Yes
naq	Male		Yes
xal	(Older or respected) Female, natural elements		No
cham	(Older or respected) Male, natural elements		No
no(')	Animal	Non-human (Animate)	Yes
an	Plant/Vegetable	Inanimate	Yes
te(')	Tree/Wood		
ch'en	Rock/metal		
(i)xim	Corn		
tx'(otx)	Land/soil		
tx'an	Fibre/Rope		
q'a(q)'	Fire		
tz'am	Salt		

5.1.2. Nepal Bhasa Noun Class System

In Nepal Bhasa, the noun class system makes a distinction between animate and inanimate entities. According to Shakya [30], this distinction is a consequence of the language's general classifier. Furthermore, according to Shakya, the language also uses classifiers that extend the general classifier and that can be categorised into sortal and mensural types.

The General classifier can be captured by our model and aligns precisely with Animacy. The sortal classifiers are based on physical properties of shape like length, flatness, and roundness. Such classifiers are covered by MODAN since their role is to classify nouns based on our model's PhysicalQuality. While a Sortal class is not included explicitly, the Bhasa's sub-classifiers are present in MODAN. Because Sortal is not equivalent to PhysicalQuality due to it not including the locative and its sub-classifiers covered, we opted to not clutter the model and refrained from adding Sortal.

For the mensural classifiers, Shakya [30] lists 13 possible classifiers that encapsulate arrangement; e.g., the classifier *-ba* is used to denote 'thin layers' and *-phi* is used for 'thick layers'. The language also

Table 5
Evaluation of MODAN with the Nepal Bhasa Noun Class System (mensural classifiers are out of scope).

Classifier	Sub-classifier	Example	Covered in MODAN
General	Animate	Ant	Yes
	Inanimate	Book	Yes
Unique	Unspecified shape/size	Hollow	Yes
Sortal	Shape/Size (length, flat, round)	Stick	Yes
	Shape/Consistency (sphere, solid)	Spherical Shape	Yes
Mensural	Arrangement	Bunch of short things	Partial
	Measurement	A serving of steamed rice	Partial

has 20 classifiers for measurement; e.g., it uses *-mhu* for 'a handful of grains' and *-phuti* for 'a drop of liquid'. While mensural classifiers were out of scope of MODAN, Collection and the types of Mass cover it in part. Overall, our model covers the Nepal Bhasa noun class system well, as shown in Table 5.

5.1.3. Ojibwe Noun Class System

Ojibwe, sometimes called *Anishinaabemowin*, is a language that is indigenous to the northern parts of North America. It is unique in that nouns are categorised in a manner that uses both the usual count/mass distinction and grammatical gender and classifiers [31]. While the language possesses the notion of "gender", its gender differs from some Indo-European languages since it does not use the three-way split of masculine, feminine, and neuter. Instead, it uses an "(In)animate" gender split. According to Meyer [31], it also has an intricate classifier system in which the classifiers are either nominal or verbal, as demonstrated in Fig. 5. The nominal classifier either serves a sortal or mensural purpose.

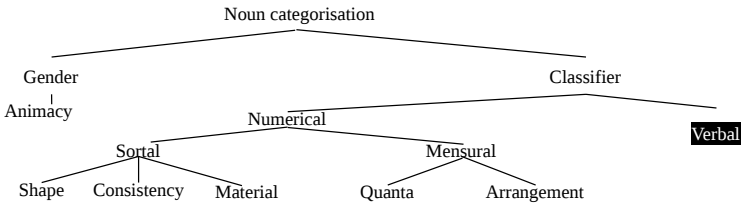

Figure 5: Noun classification system in Ojibwe. Items highlighted with black background are out of scope.

Table 6 shows how MODAN covers the Ojibwe noun class system. It can capture its gender through the AbstractQuality subclasses. More precisely, the language's concept of Gender directly corresponds to our model's Animacy, i.e., Animate $\sqcup$ Inanimate. The sortal numerical classifiers are captured via PhysicalQuality, where Consistency $\sqsubseteq$ PhysicalQuality and Shape $\sqsubseteq$ PhysicalQuality are present and its "material" aligns to MODAN's PhysicalObject.

While mensural classifiers are out of scope of at least this initial version of MODAN, also here, it can be covered at a high level nonetheless. First, the notion of Quanta can be captured via MODAN's Collection. Specifically, since Meyer's [31] description of Quanta does not restrict the concept via a clear commitment regarding the relationship between elements, it is equivalent to the 'top' level concept Collection and thus allows its elements to be grouped; as an example, by proximity so as to support the modeling of a cluster, it also supports grouping via any membership criterion in the event that the goal is to model a set. It can also be used to relate elements in a manner that forms a collective.

Second, the Arrangement concept is also supported by the vocabulary elements in MODAN, albeit benefiting from further axiomatisation. Meyer's Arrangement refers to the spatial or structural configuration of a set of items belonging to some whole. As such, it can be captured with Arrangement $\equiv$ SetOfItems $\sqcap \exists$attributedTo.Configuration, although a refinement with hasConfiguration and hasConfiguration $\sqsubseteq$ attributedTo may be preferable. One can understand an arrangement in

terms of DOLCE entities as how Endurants are positioned relative to one another according to some Physical quality.

Table 6
Evaluation of MODAN with the Ojibwe noun class system

Categorisation feature	Type	Values	Covered with MODAN
Gender		Animate	Yes
		Inanimate	Yes
Classifier	Sortal numerical	Shape	Yes
		Consistency	Yes
		Material	Yes
	Mensural numerical	Quanta	Partial
		Arrangement	Partial

5.2. Reconsidering the Use Cases with Nouns

The use cases from Section 2.1 zoomed in on individual nouns with implied consequences at the level of noun classes where it will be needed in an NLP application and in annotation efforts. Nonetheless, it can be enlightening to assess the classification at word-level for a potential bottom-up approach to re-assess noun classification.

The first example can now be annotated satisfactorily: nc:1 and nc:2 are for nouns denoting Humans whereas nc:1a and nc:2a are for the Roles they play [11], which MODAN indeed can differentiate between. Few NCB languages do make that distinction, and Denny and Creider's proto-Bantu categorisation [13] has nc:1 and nc:1a still together in the Human classifier. This is unsurprising since insights into modelling roles has intensified only over the past decades, their categorisation has no option for 'roles', and it is only partially reflected in the aggregate Table 1. Why it split off in a 'sub' noun class in isiZulu is a separate question that MODAN obviously cannot answer.

For the stone vs the stick example, the features are more indicative than the noun class table, and they also align with that proto-Bantu feature hierarchy. For nc:9 for the stick, in Fig. 1 and [13], one drills down from count noun to configuration, outline figure, unit, non-extended, whereas in MODAN it ends up in the leaf entity Thin shape. For nc:5 and the stone, it can be categorised as a solid object or mass (mass - cohesive - solid, homogeneous, as well as configuration - solid figure - non-extended - unit), and fits with MODAN as well, as Solid physical quality and/or Homogeneous mass, owing to ambiguity in translation of *ngetshe* because it can be translated as 'with stone' and 'with the stone'.

6. Discussion

The model we developed was shown to be adequate for geographically dispersed languages with nominal noun classes, and at a high level, also for mensural noun classes despite not included at the outset. The key insights concern the semantics of the noun classes and how nouns or their referents are categorised into noun classes (and, consequently, behave linguistically). The aggregated categories of types of entities in a noun class, as is the standard approach for NCB languages (as in Table 1) are resistant to alignment to a foundational ontology and do not fit with a linguistics domain ontology either. Categorisation approaches proposed for languages in Asia and South America that use chosen salient features of an entity, such as long, solid, etc. appear to hold up better across languages. MODAN may assist in such linguistic investigations. Such semantics may, in turn, also assist with NLP applications that need noun classes to function properly, which was shown to be needed even for basic infrastructure such as a noun pluraliser [27] as well as more advanced tasks such as text generation [11, 10].

Other evaluation options were considered, notably to tag noun classes and examine if referents of the nouns satisfy it, or taking a set of nouns and annotate them with terms from MODAN to then see which noun class it would/should/plausibly could/does belong to. Such evaluations will have fascinating

outcomes, but were deemed out of scope of the current question on ontological vs linguistic layer and assessing quality of the model in the sense of sufficient coverage, rather than the long-standing debate whether application of noun classes in a language is semantically coherent, which resides in a linguistics discipline rather than computing.

7. Conclusions

The analysis of the meaning of noun class systems demonstrated that a top-down approach and neat match with a foundational ontology alone did not yield a satisfactory ontology, due to the different desiderata or key reasons for identifying distinctions and an apparent mixture of top- and domain-level entities. A bottom-up approach yielded a model of principal features for semantically classifying referents with their nouns, called MODAN. It was evaluated to have good coverage for languages with noun class systems that can be found across continents.

The observations suggest that there is a different rationale for classifying entities in an ontology and for classifying nouns of entities in, or reconstructing why nouns of entities are in, a particular semantics-bearing noun class. This insight is expected to affect NLP for languages with a noun class system. We plan to test the effects of noun class annotation with MODAN with a number of NLP tasks, including enhancing detecting a noun's noun class, and from a linguistics viewpoint to assess its applicability to NCB languages such as isiZulu.

Supplemental material statement and FAIR The CC-BY licenced MODAN model is available on the model's page at http://www.meteck.org/KeBoNa/modan.html and on https://github.com/orgs/KEEN-Research.

Acknowledgments

This work was financially supported in part by the National Research Foundation (NRF) of South Africa (Grant Number and CPRR23040389063).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] G. A. Miller, Wordnet: A lexical database for English, Commun. ACM 38 (1995) 39–41.
[2] C. Chiarcos, M. Sukhareva, OLiA – ontologies of linguistic annotation, Semantic Web Journal 6 (2015) 379–386.
[3] P. Cimiano, J. P. McCrae, P. Buitelaar, Lexicon Model for Ontologies: Community Report, Final Community Group Report, 10 May 2016, W3C, 2016.
[4] J. O. Alabi, M. A. Hedderich, D. I. Adelani, D. Klakow, Charting the landscape of African NLP: Mapping progress and shaping the road ahead, in: Proc. of EMNLP'25, Suzhou, China, 2025, pp. 27807–27841.
[5] S. Farrar, D. T. Langendoen, A linguistic ontology for the semantic web, in: GLOT International, volume 7 of *3*, 2003, pp. 97–100.
[6] C. Chavula, C. M. Keet, An orchestration framework for linguistic task ontologies, in: Proc. of MTSR'15, volume 544 of *CCIS*, Springer, 2015, pp. 3–14. 9-11 Sept, 2015, Manchester, UK.
[7] C. Chavula, C. M. Keet, Is lemon sufficient for building multilingual ontologies for Bantu languages?, in: Proc. of OWLED'14, volume 1265 of *CEUR-WS*, 2014, pp. 61–72. Riva del Garda, Italy, Oct 17-18, 2014.

[8] S. Borgo, R. Ferrario, A. Gangemi, N. Guarino, C. Masolo, D. Porello, E. M. Sanfilippo, L. Vieu, DOLCE: A descriptive ontology for linguistic and cognitive engineering, Applied Ontology 17 (2022) 45–69.
[9] A. Y. Aikhenvald, A Typology of Noun Categorization Devices, Cambridge Handbooks in Language and Linguistics, Cambridge University Press, 2017, p. 361–404.
[10] Z. Mahlaza, Foundations for reusable and maintainable surface realisers for isiXhosa and isiZulu, Ph.D. thesis, University of Cape Town, South Africa, 2022.
[11] C. M. Keet, L. Khumalo, Toward a knowledge-to-text controlled natural language of isiZulu, Language Resources and Evaluation 51 (2017) 131–157.
[12] J. Byamugisha, Noun class disambiguation in Runyankore and related languages, in: Proc. of COLING'22, ACL, 2022, pp. 4350–4359. Gyeongju, Republic of Korea, October 12-17 2022.
[13] P. Denny, C. A. Creider, Noun classes and categorization. Typological Studies in Language 7, Amsterdam: John Benjamins, 1986, pp. 217–241.
[14] R. Watson, Kujireray: morphosyntax, noun classification and verbal nouns, Ph.D. thesis, Department of linguistics, University of London, London, UK, 2015.
[15] Y. Li, Sortal classifiers and noun categorization in zauzou, Language and Linguistics 25 (2024) 270–317.
[16] Y. Wang, G. Walther, Measure words are measurably different from sortal classifiers, in: Proc. of Depling GURT/SyntaxFest 2023), George Mason University, 2023, pp. 81–90.
[17] Building Ontologies with Basic Formal Ontology, The MIT Press, USA, 2015.
[18] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, Sweetening WORDNET with DOLCE, AI Magazine 24 (2003) 13.
[19] A. G. L. Junior, J. L. Carbonera, D. Schmidt, M. Abel, Predicting the top-level ontological concepts of domain entities using word embeddings, informal definitions, and deep learning, Expert Systems with Applications 203 (2022) 117291.
[20] C. H. Bernabé, C. M. Keet, Z. C. Khan, Z. Mahlaza, A method to improve alignments between domain and foundational ontologies, in: Proc. of FOIS'23, volume 377 of *FAIA*, IOS Press, 2023, pp. 125–139.
[21] C. M. Keet, M. T. Khan, C. Ghidini, Ontology authoring with FORZA, in: Proc. CIKM'13, ACM proceedings, 2013, pp. 569–578. Oct. 27 - Nov. 1, 2013, San Francisco, USA.
[22] J. A. Bateman, Upper modeling: organizing knowledge for natural language processing, in: Proc. of INLG'99, Pittsburgh, PA., 1990, pp. 54–60.
[23] J. A. Bateman, GUM: the generalized upper model, Applied Ontology 17 (2022) 107–141.
[24] S. Farrar, D. Langendoen, Markup and the GOLD Ontology, in: Proc. of Ws. on Digitizing and Annotating Text and Field Recordings, 2003.
[25] I. Niles, A. Pease, Towards a standard upper ontology, in: C. Welty, B. Smith (Eds.), Proc. of FOIS'01, 2001. Ogunquit, Maine, October 17-19, 2001.
[26] I. Niles, A. Pease, Linking lexicons and ontologies: Mapping WordNet to the suggested upper merged ontology, in: Proc. of IKE'03, CSREA Press, 2003, pp. 411–416.
[27] J. Byamugisha, C. M. Keet, B. DeRenzi, Pluralizing nouns across agglutinating Bantu languages, in: Proc. of COLING'18, ACL, 2018, pp. 2633–2643. 20-26 Aug, 2018, Santa Fe, New Mexico, USA.
[28] C. M. Keet, A core ontology of macroscopic stuff, in: Proc. of EKAW'14, volume 8876 of *LNAI*, Springer, 2014, pp. 209–224. 24-28 Nov, 2014, Linkoping, Sweden.
[29] P. Duncan, Noun classifier extension in q'anjob'al (mayan): Acquiring a gender stereotype, Kansas Working Papers in Linguistics 32 (2011) 67–83.
[30] D. R. Shakya, Classifiers and their syntactic functions in Nepal Bhasa, HIMALAYA, the Journal of the Association for Nepal and Himalayan Studies 17 (1997) 4.
[31] C. L. Meyer, Noun categorization in Ojibwe: Gender and classifiers, Ph.D. thesis, The University of Chicago, Chicago, USA, 2020.