Cross-Lingual Knowledge Augmentation for Mitigating Generic Overgeneralization in Multilingual Language Models

Sello Ralethe and Jan Buys
Department of Computer Science, University of Cape Town, South Africa
rltsel002@myuct.ac.za, jbuys@cs.uct.ac.za

Abstract

Generic statements like "birds fly" or "lions have manes" express generalizations about kinds that allow exceptions, yet language models tend to overgeneralize them to universal claims. While previous work showed that ASCENT KB could reduce this effect in English by 30-40%, the effectiveness of broader knowledge sources and the cross-lingual nature of this phenomenon remain unexplored. We investigate generic overgeneralization across English and four South African languages (isiZulu, isiXhosa, Sepedi, SeSotho), comparing the impact of ConceptNet and DBpedia against the previously used ASCENT KB. Our experiments show that ConceptNet reduces overgeneralization by 45-52% for minority characteristic generics, while DBpedia achieves 48-58% for majority characteristics, with combined knowledge bases reaching 67% reduction. These improvements are consistent across all languages, though Nguni languages show higher baseline overgeneralization than Sotho-Tswana languages, potentially suggesting that morphological features may influence this semantic bias. Our findings demonstrate that commonsense and encyclopedic knowledge provide complementary benefits for multilingual semantic understanding, offering insights for developing NLP systems that capture nuanced semantics in low-resource languages. We release the dataset and code[1]

1 Introduction

Generic statements express generalizations about kinds that tolerate exceptions, representing a fundamental aspect of how humans conceptualize and communicate about the world. Statements such as "birds fly" or "lions have manes," express truths about these categories despite the fact that penguins cannot fly and female lions lack manes. This linguistic phenomenon poses a significant challenge for natural language understanding systems, as both humans and language models exhibit a bias toward interpreting these statements as universal claims; a phenomenon known as generic overgeneralization (GOG) (Leslie et al., 2011).

The tendency to overgeneralize from generic statements to universal claims reflects cognitive biases in how humans process categorical information. When presented with a true generic like "ducks lay eggs," people and models tend to incorrectly accept the universal statement "all ducks lay eggs," despite the obvious fact that only female ducks possess this capability (Khemlani et al., 2007). This effect has been documented in cognitive science literature (Hollander et al., 2002; Cimpian, 2010) and represents an important test case for evaluating whether language models truly understand the nuanced semantics of natural language.

Recent advances in multilingual representation learning have shown notable success in transferring knowledge across languages, yet the interaction between these methods and language-specific phenomena like genericity remains largely unexplored. This gap is more pronounced for morphologically rich, low-resource languages, where both training data and linguistic resources are scarce (Nigatu et al., 2023; Chang et al., 2024; Qin et al., 2025).

Languages such as isiZulu, isiXhosa, Sepedi, and SeSotho face challenges due to limited digital corpora (Eiselen and Gaustad, 2023; Mesham et al., 2021). These languages express genericity and other pragmatic phenomena through morphological mechanisms distinct from English, potentially affecting how generic statements are interpreted and overgeneralized. The analytic tools developed for machine translation and representation (e.g. morphology-aware modeling methods) demonstrate that explicit morphological structure affects performance in these contexts (Nzeyimana,

[1] https://github.com/sello-ralethe/Multilingual_Generics

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 483–495

2024), yet empirical work on genericity is lacking.

In this paper, we present an investigation of generic overgeneralization across multiple languages, examining how this phenomenon manifests in typologically diverse languages and whether knowledge enhancement can mitigate its effects. We make several contributions that advance the understanding of this semantic phenomena. First, we demonstrate that generic overgeneralization is indeed a cross-linguistic phenomenon that affects languages with different morphological systems for expressing genericity. Our experiments with four South African languages show patterns in how different language families exhibit this bias, with Nguni languages displaying higher baseline overgeneralization than Sotho-Tswana languages.

Second, we show that knowledge enhancement through carefully selected knowledge bases can reduce overgeneralization effects. By comparing ASCENT KB (Nguyen et al., 2020), ConceptNet (Speer et al., 2016), and DBpedia (Auer et al., 2007) as knowledge sources, we find that different types of knowledge address different aspects of the overgeneralization problem. ConceptNet's commonsense knowledge proves effective for minority characteristic generics, achieving 45-52% relative reduction in overgeneralization, while DBpedia's encyclopedic coverage excels at handling majority characteristic generics with 48-58% reduction. The combination of both knowledge types yields even stronger results, reaching up to 67% reduction in overgeneralization.

In this paper, we present the first investigation of generic overgeneralization across morphologically rich, low-resource languages, examining how this phenomenon manifests in typologically diverse settings and whether knowledge enhancement can mitigate its effects across linguistic boundaries. We make several contributions that advance understanding of this semantic phenomenon in multilingual contexts.

First, we demonstrate that generic overgeneralization is indeed a cross-linguistic phenomenon, providing empirical evidence across English and four South African languages (isiZulu, isiXhosa, Sepedi, and SeSotho) which represent two distinct language families. Our experiments reveal systematic patterns in how different language families exhibit this bias, with Nguni languages displaying 4-7% higher baseline overgeneralization than Sotho-Tswana languages, suggesting that morphological features may modulate semantic biases.

Second, we compare three knowledge sources, demonstrating that different types of knowledge address different aspects of the overgeneralization problem. We show that ConceptNet's commonsense knowledge proves effective for minority characteristic generics, achieving 45-52% relative reduction in overgeneralization, while DBpedia's encyclopedic coverage excels at handling majority characteristic generics with 48-58% reduction.

The combination of both knowledge types yields even stronger results, reaching up to 67% reduction in overgeneralization. Importantly, these improvements remain consistent across all languages, demonstrating that conceptual knowledge effectively transfers across linguistic boundaries despite significant morphological differences. Our findings thus offer practical insights for developing NLP systems that capture nuanced semantics in low-resource multilingual settings while advancing theoretical understanding of how semantic biases interact with morphological systems.

2 Related Work

2.1 Generic Overgeneralization in Language Models

The distinction between generic statements and universally quantified statements represents a fundamental challenge in natural language semantics that has implications for multilingual NLP. While "tigers have stripes" holds true as a generic despite albino tigers lacking stripes, the universal statement "all tigers have stripes" is demonstrably false. This subtle distinction shows how language encodes conceptual knowledge about categories and their typical properties (pel, 2009).

The generic overgeneralization effect, first documented in cognitive science by Leslie et al. (2011) and Khemlani et al. (2007), demonstrates a human tendency to conflate these two types of statements. This cognitive bias appears to be rooted in humans' default processing mechanisms, where accepting universal interpretations requires less cognitive effort than maintaining the nuanced understanding that generics admit exceptions (Leslie et al., 2011). Recent work by Ralethe and Buys (2022) extended this investigation to pre-trained language models, showing that when BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) were asked to predict masked tokens in contexts like "[MASK] lions have manes," these models showed strong preferences for universal quantifiers like "all" and "every."

Experiments by Ralethe and Buys (2022) demonstrated that language models not only exhibit human-like overgeneralization patterns but that this bias could be partially mitigated through knowledge injection. By incorporating factual knowledge from ASCENT KB (Nguyen et al., 2020), they achieved a 30-40% reduction in overgeneralization. However, ASCENT KB's limitations, including its relatively sparse coverage of approximately 400k animal-related triples and focus on specific factual assertions rather than broader conceptual knowledge, suggests that richer knowledge sources might prove more effective.

2.2 Commonsense vs. Encyclopedic Knowledge

The contrast between different types of knowledge bases indicates complementary approaches to representing world knowledge. ConceptNet (Speer et al., 2016) encodes commonsense knowledge that people typically know about the world, including relations like "CapableOf," "HasProperty," and "PartOf" that capture prototypical information about concepts. This type of knowledge proves valuable for generic reasoning because it encodes default expectations about kinds, including information about typical properties and capabilities that align with how humans conceptualize categories (Liu and Singh, 2004).

DBpedia (Auer et al., 2007), extracted from Wikipedia, provides encyclopedic, factual knowledge including specific information about instances, detailed taxonomies, and factual properties. For generic reasoning, DBpedia's strength lies in its comprehensive coverage of exceptions and variations (Mendes et al., 2011). It contains information about albino tigers, flightless birds, and other edge cases that violate generic expectations, making it particularly valuable for understanding when universal generalizations fail.

The complementary nature of these knowledge sources becomes apparent when considering their coverage. While ASCENT KB focuses on specific faceted assertions like "young lions do not have manes," ConceptNet provides broader conceptual knowledge such as "mane is a characteristic feature of male lions," and DBpedia offers comprehensive factual coverage including specific information about white lions, Barbary lions, and other variations. This suggests that effective mitigation of overgeneralization may require multiple types of knowledge working in concert (Ilievski et al., 2020).

2.3 Cross-Lingual Considerations

The expression of genericity varies significantly across languages, raising important questions about whether generic overgeneralization is universal or language-specific (Dayal, 2004; Chierchia, 1998). English uses bare plurals for generic reference, while other languages use different morphosyntactic strategies. In Nguni languages like isiZulu and isiXhosa, the noun class system inherently pluralizes nouns, with generic reference typically achieved through class prefixes (Zeller, 2012; Visser, 2008). For example, "amabhubesi" (lions) in isiZulu uses the class 6 prefix ama-, which inherently indicates plurality. Sotho-Tswana languages like Sepedi and SeSotho use a different noun class system with distinct morphological patterns for expressing genericity (Mojapelo, 2009).

These typological differences have important implications for how generic overgeneralization might manifest across languages. The obligatory plural marking in Nguni languages may create different baseline expectations about universality compared to languages with optional plural marking (Demuth, 2000). Furthermore, the morphological complexity of these languages poses additional challenges for knowledge projection and alignment, as the same concept may be realized through different morphological forms depending on the syntactic context (Kiparsky, 2001).

Previous work on cross-lingual knowledge projection has shown that conceptual knowledge can transfer across languages (Chen et al., 2016, 2021; Sun et al., 2019), but the interaction with language-specific phenomena like genericity remains largely unexplored. The success of multilingual models like mT5 (Xue et al., 2021) in capturing cross-lingual semantic similarities suggests that conceptual knowledge about generics might transfer across languages, but this hypothesis requires empirical validation across typologically diverse languages.

While prior work has examined generic overgeneralization in English (Ralethe and Buys, 2022), our work is the first to: (1) investigate this phenomenon across morphologically rich, low-resource African languages, (2) systematically compare commonsense versus encyclopedic knowledge sources for GOG mitigation, and (3) demonstrate effective cross-lingual knowledge transfer for this semantic task despite typological diversity.

3 Methodology

3.1 Data and Languages

Our investigation encompasses a curated dataset of generic statements and a diverse set of low-resource languages representing different typological features. We utilize the generic overgeneralization datasets from Ralethe and Buys (2022), comprising 5884 minority characteristic generics that express properties true of only a subset of a kind, such as "lions have manes," and 8750 majority characteristic generics that express prevalent but not universal properties, such as "tigers have stripes." Additionally, we use 60368 training generics covering diverse generic types to ensure comprehensive coverage of the phenomenon.

For our cross-lingual study, we select English as our baseline and four South African languages representing two distinct language families. The Nguni languages, isiZulu and isiXhosa, share similar morphological structures including extensive noun class systems with obligatory plural marking. The Sotho-Tswana languages, Sepedi and SeSotho, use different noun class systems and morphological patterns. This selection allows us to investigate how typological differences influence generic overgeneralization while controlling for potential areal effects, as all four languages are spoken in South Africa.

3.2 Translation and Quality Validation

To ensure high-quality cross-lingual data, we translated all datasets using the Google Translate API with rigorous quality controls. Our validation process included back-translation verification to identify potential translation errors, entity name validation to ensure proper nouns were correctly handled, and manual checking of quantifier translations.

To quantify translation quality, we conducted manual validation on a random sample of 200 generic statements per language. Each translation was evaluated for semantic accuracy and grammatical correctness. The validation demonstrated high translation quality overall: isiZulu (88%), isiXhosa (89%), Sepedi (91%), and SeSotho (93%). Common translation errors included:

IsiZulu: Incorrect handling of noun class agreement, particularly with complex subjects. For instance, "Young elephants play in water" was incorrectly translated as "Izindlovu ezincane zidlala emanzini" where the class prefix failed to maintain consistency with age modifiers.

IsiXhosa: Confusion between inclusive and exclusive plural forms. The generic "Lions hunt at night" was rendered as "Iingonyama zizingela ebusuku" which could be interpreted as referring to specific lions rather than lions in general.

Sepedi: Misalignment of aspectual markers affecting the generic interpretation. "Birds migrate seasonally" translated to "Dinonyana di huduga ka nako ya sehla" lost the habitual aspect important for generic meaning.

SeSotho: Occasional loss of generic force through inappropriate determiner insertion. "Cats are independent" became "Dikatse tsena di ikemela" where "tsena" (these) inadvertently introduced a deictic element.

These error patterns informed our analysis, particularly regarding how morphological features interact with generic interpretation across language families.

Rationale for Translation Approach We use translation rather than collecting native generic statements because no existing generic overgeneralization datasets exist for these low-resource languages, and creating new datasets would require extensive linguistic validation to ensure consistent generic interpretation across cultures. Translation maintains exact parallel alignment across languages, enabling controlled comparison of how the same conceptual content is processed across different morphological systems. Our high translation quality (88-93% accuracy) and detailed error analysis demonstrate that this approach is sound for investigating cross-linguistic patterns, though we acknowledge translation may introduce some noise.

3.3 Knowledge Sources

Our experimental design compares three distinct knowledge sources, each offering different types and scales of information. Following Ralethe and Buys (2022), we use ASCENT KB as our baseline, which contains approximately 403k animal-related triples with faceted information about properties and subcategories. While ASCENT KB provides valuable specific assertions, its coverage is limited compared to larger knowledge bases.

We extend this baseline by incorporating ConceptNet and DBpedia, both of which offer substantially richer information. ConceptNet provides approximately 220k triples per language after projection into South African languages through LeNS-

Align (Ralethe and Buys, 2025), encoding diverse relation types including taxonomic relations like "male_lion IsA lion," property relations such as "lion HasProperty mane," capability relations like "bird CapableOf fly," and prototype relations such as "tiger HasA stripes." This commonsense knowledge captures the conceptual structures that underlie generic statements.

DBpedia contributes approximately 450k triples per language after projection (Ralethe and Buys, 2025), offering instance data such as "Cecil_(lion) type Lion," comprehensive taxonomic information like "White_tiger subClassOf Tiger," detailed property data including "Albino_tiger colour White," and extensive geographic and demographic information. This encyclopedic knowledge provides the factual grounding necessary to understand exceptions to generic generalizations.

3.4 Model Architectures

Our experimental framework uses different architectures for English and multilingual experiments to leverage the most appropriate models for each setting. For English experiments, we implement BERT-large and RoBERTa-large augmented with knowledge bases using the KEPLER framework (Wang et al., 2021), following the approach of Ralethe and Buys (2022). KEPLER enables knowledge integration by continuing pre-training on verbalized knowledge triples, where each triple is converted to natural language using templates. This approach allows us to maintain compatibility with the baseline while exploring richer knowledge sources.

For multilingual experiments, we adopt mT5-large as our base model, leveraging its strong multilingual capabilities across all target languages. We follow Ralethe and Buys (2025) in performing knowledge injection of the projected knowledge bases using an adaptation of the QA-GNN framework (Yasunaga et al., 2021).

QA-GNN retrieves relevant subgraphs for each generic statement and uses graph attention networks to reason over the structured knowledge, enabling explicit traversal of knowledge graph connections when interpreting generics across languages. This architecture proves well-suited for working with projected knowledge bases in low-resource languages, as it can leverage the graph structure to compensate for potential noise in the projections (See Appendix B for implementation and training details).

3.5 Evaluation Framework

We use three complementary evaluation tasks to assess model performance and the manifestation of generic overgeneralization. The generic classification task evaluates whether models can distinguish between generic and non-generic statements, with particular focus on universally quantified versions. This task directly tests whether models understand that statements like "all lions have manes" are not true generics despite the truth of the unquantified version.

Following the original work of Ralethe and Buys (2022), the quantifier prediction task provides our primary measure of overgeneralization. By masking the pre-nominal position in statements like "[MASK] lions have manes," we evaluate how strongly models prefer universal quantifiers. We calculate the Mean Reciprocal Rank (MRR), which measures the inverse of the rank at which the first correct answer appears, averaged across all test instances. For this task, we consider universal quantifiers (all, every, each) as the target predictions, so lower MRR scores indicate better performance as they suggest the model is less likely to predict universal quantifiers. We also compute Precision at 5 (P@5), which measures the proportion of test instances where at least one universal quantifier appears in the top 5 predictions. Lower scores on both metrics indicate less overgeneralization, as models that avoid predicting universal quantifiers demonstrate better understanding of generic semantics.

The quantifier interpretation probing task creates statements with different quantifiers and masks the property position, as in "all lions have [MASK]." Models should assign higher probabilities to the correct property for quantifiers that maintain truth (some, most) than for those that create false universal statements. This task uses MRR to measure how highly models rank the correct property, with higher scores indicating better understanding when the quantifier makes the statement true. This task helps determine whether models genuinely understand the semantic implications of different quantifiers or merely exhibit surface-level patterns.

4 Results

4.1 Comparison with Previous Work: English Results

Table 1 presents a comparison of our results on English with the previous ASCENT KB baseline

from Ralethe and Buys (2022). The improvements achieved by ConceptNet and DBpedia are notable across all evaluation metrics, showing important insights about the types of knowledge most effective for addressing generic overgeneralization.

For minority characteristic generics, ConceptNet demonstrates notable effectiveness, achieving 45-52% relative reduction in overgeneralization compared to 30-34% for using ASCENT KB. This improvement stems from ConceptNet's richer representation of subcategory relationships and prototypical properties. Where ASCENT KB might only encode "male lions have manes," ConceptNet additionally provides conceptual relations such as "mane *IsA* male characteristic" and "adult male lion *IsA* lion with mane." These additional layers of conceptual knowledge help models understand that properties like manes are inherently restricted to subsets of a category.

DBpedia shows its greatest strength with majority characteristic generics, achieving 48-58% reduction versus ASCENT KB's 40%. This advantage arises from DBpedia's comprehensive coverage of exceptions and edge cases. While ASCENT KB might note that albino tigers exist, DBpedia provides detailed information about white tigers, melanistic tigers, golden tigers, and numerous specific individuals. This exhaustive coverage of variations gives models concrete evidence against universal generalizations.

The combined ConceptNet+DBpedia approach achieves up to 67% reduction in overgeneralization, nearly doubling ASCENT KB's best performance. This synergy suggests that commonsense and encyclopedic knowledge provide fundamentally complementary benefits. ConceptNet helps models understand the conceptual structure of categories and why certain properties might be restricted to subsets, while DBpedia provides the specific counterexamples that definitively rule out universal generalizations.

4.2 Cross-Lingual Results

Table 2 presents the results of the quantifier prediction task in all five test languages, demonstrating both universal patterns and language-specific variations in generic overgeneralization across languages. The results show that knowledge enhancement provides consistent benefits across typologically diverse languages, though interesting patterns emerge related to language family and morphological structure.

The most notable finding is the consistency of knowledge enhancement effects across languages. ConceptNet provides 43-47% reduction for minority generics across all languages, while DBpedia achieves 52-56% reduction for majority generics. This suggests that the conceptual knowledge encoded in these resources transfers effectively across languages through the LeNS-Align projection process, despite the significant morphological differences between English and the target languages.

A pattern emerges when comparing language families. Nguni languages (isiZulu and isiXhosa) exhibit higher baseline overgeneralization than Sotho-Tswana languages (Sepedi and SeSotho) and English. The baseline MRR for universal quantifiers is 4-7% higher in Nguni languages. We hypothesize that this may be related to the obligatory plural marking in the Nguni noun class system, which could prime speakers and models toward universal interpretations of generic statements.

The pattern of ConceptNet excelling at minority generics while DBpedia excels at majority generics holds across all languages, confirming that different types of overgeneralization (overgeneralizing from "some" to "all" versus overgeneralizing from "most" to "all") require different types of knowledge to address effectively. This cross-linguistic consistency suggests that the cognitive and semantic factors underlying generic overgeneralization are largely universal, even as their surface manifestations vary across languages.

4.3 Classification Results

The generic classification results presented in Table 3 provide additional evidence for both the pervasiveness of overgeneralization and the effectiveness of knowledge enhancement. When asked to classify universally quantified statements as generic or non-generic, baseline models fail, achieving only around 10% accuracy. This near-chance performance indicates that without additional knowledge, models treat statements like "all lions have manes" as equivalent to the generic "lions have manes."

Knowledge enhancement provides improvements, with the combined approach achieving 34-39% accuracy across languages. While still far from perfect, this represents a three- to four-fold improvement over the baseline. This improvement across languages reinforces our finding that knowledge injection helps models develop more nuanced understanding of generic semantics.

Model	Minority		Majority	
	MRR	Reduction	MRR	Reduction
BERT	0.326	-	0.337	-
+ASCENT†	0.228	30.1%	0.202	40.1%
+ConceptNet	0.179	45.1%	0.185	45.1%
+DBpedia	0.186	42.9%	0.175	48.1%
+Both KBs	**0.142**	**56.4%**	**0.138**	**59.1%**
RoBERTa	0.329	-	0.428	-
+ASCENT†	0.217	34.0%	0.257	40.0%
+ConceptNet	0.158	52.0%	0.221	48.4%
+DBpedia	0.171	48.0%	0.180	57.9%
+Both KBs	**0.108**	**67.2%**	**0.141**	**67.1%**

Table 1: English results for the quantifier prediction task comparing knowledge sources (MRR for universal quantifiers - lower is better). † indicates results from Ralethe and Buys (2022).

Model	Minority Characteristic Generics					
	English	isiZulu	isiXhosa	Sepedi	SeSotho	Avg Reduction
mT5	0.318	0.347	0.352	0.324	0.319	-
+ConceptNet	0.175	0.189	0.193	0.181	0.177	45.0%
+DBpedia	0.184	0.198	0.201	0.186	0.182	42.1%
+Both KBs	**0.139**	**0.151**	**0.154**	**0.144**	**0.141**	**55.7%**
Model	Majority Characteristic Generics					
	English	isiZulu	isiXhosa	Sepedi	SeSotho	Avg Reduction
mT5	0.412	0.436	0.441	0.411	0.407	-
+ConceptNet	0.216	0.231	0.235	0.218	0.214	47.3%
+DBpedia	0.189	0.201	0.205	0.187	0.184	54.8%
+Both KBs	**0.136**	**0.148**	**0.152**	**0.135**	**0.133**	**67.0%**

Table 2: Cross-lingual results for the quantifier prediction task: MRR for universal quantifiers across all languages (lower is better)

4.4 Probing Experiments

To investigate whether knowledge-enhanced models truly understand the injected knowledge, we conducted two probing experiments adapted from Ralethe and Buys (2022) for our multilingual mT5 setup.

4.4.1 Quantified Statement Classification Probing

We fine-tuned the knowledge-enhanced mT5 on the generic classification task and tested whether quantified statements are correctly classified as non-generic. We quantified minority characteristic generics with "many" and "most," and majority characteristic generics with "few" and "some" to create false generic statements. For example, "most lions have manes" should be classified as non-generic since only a minority of lions have manes.

Table 4 shows that knowledge injection improves the models' ability to recognize false quantified statements, though accuracy remains low. The combined KB approach achieves 21.3% accuracy for minority characteristics and 28.6% for majority characteristics, suggesting that models partially learn the conceptual distinctions but struggle to apply them consistently. Detailed per-language results in Appendix A show that Nguni languages underperform Sotho-Tswana languages in this task, mirroring the overgeneralization patterns.

4.4.2 Quantifier Interpretation Probing

We evaluated whether models correctly interpret different quantifiers by masking the property in quantified statements. For each generic, we created probing instances with four quantifiers (few, some, many, most) and masked the final token. Models should rank the correct property higher for quantifiers that make the statement true.

The results in Table 5 show that knowledge-enhanced models display improved quantifier interpretation. For minority characteristic generics, models correctly assign higher MRR to properties when quantified with "few" or "some" compared to "many" or "most." The pattern reverses appro-

Model	English	isiZulu	isiXhosa	Sepedi	SeSotho	Average
Baseline	10.8	9.7	8.3	12.1	11.4	10.5
+ConceptNet	23.4	21.2	19.6	24.5	23.7	22.5
+DBpedia	24.9	22.8	21.1	25.3	24.6	23.7
+Both KBs	**38.7**	**36.4**	**34.2**	**39.1**	**38.3**	**37.3**

Table 3: Generic classification accuracy (%) on universally quantified variants

Model	Classification Accuracy (%)	
	Minority	Majority
mT5	8.3	10.1
+ConceptNet	14.7	18.2
+DBpedia	13.9	19.4
+Both KBs	**21.3**	**28.6**

Table 4: Accuracy of classifying falsified quantified generics as non-generic (averaged across languages; see Appendix A for per-language results)

priately for majority characteristic generics. The combined KB approach shows the strongest differentiation between appropriate and inappropriate quantifiers, with the gap between true and false quantifiers widening from 0.21 to 0.41 for minority characteristics and from 0.27 to 0.43 for majority characteristics. Per-language analysis (Appendix A) shows that Nguni languages achieve the largest differentiation gaps despite higher baseline overgeneralization.

However, the relatively high MRR scores even for false quantifiers (e.g., 0.31 for "most" with minority generics) indicate that models still struggle with complete understanding. The quantifier "some" proves particularly challenging across all languages (Appendix A), maintaining relatively high scores across both generic types, suggesting models interpret it as a hedge rather than a specific quantity indicator.

5 Discussion

Our results provide several insights into generic overgeneralization, the role of knowledge in addressing it, and the cross-lingual nature of this phenomenon.

5.1 Why ConceptNet and DBpedia Outperform ASCENT KB

The effectiveness of ConceptNet and DBpedia over ASCENT KB reflects their complementary knowledge coverage. ConceptNet's strength for minority characteristic generics emerges from its encoding of conceptual relationships that help models understand the logical structure of subset properties. When a model needs to understand that "lions have manes" does not mean "all lions have manes," ConceptNet provides the conceptual framework: manes are a male characteristic, male lions are a subset of lions, and characteristics can be subset-specific.

DBpedia's advantage for majority characteristic generics stems from its encyclopedic coverage of exceptions. While ASCENT KB might note that albino tigers exist, DBpedia provides detailed information about white tigers, golden tigers, and stripeless tigers, giving models concrete evidence against universal generalizations.

The combined approach achieving up to 67% reduction demonstrates that generic reasoning requires both conceptual understanding and factual grounding. Neither pure commonsense nor pure factual knowledge alone suffices; models need to understand both the conceptual possibility of exceptions and specific instances of those exceptions.

5.2 Cross-Lingual Universality and Variation

The consistency of knowledge enhancement effects across languages provides evidence that generic overgeneralization reflects a deep semantic challenge rather than a surface linguistic phenomenon. Despite different morphological systems for expressing genericity, all languages benefit similarly from the same types of knowledge, supporting the view that overgeneralization stems from conceptual biases in how categories and properties are related.

However, the higher baseline overgeneralization in Nguni languages could be related to obligatory plural marking creating a stronger bias toward universal interpretation, suggesting that language-specific features can potentially amplify or dampen universal cognitive biases. The fact that knowledge enhancement reduces but does not eliminate these cross-linguistic differences indicates a complex interaction between universal conceptual tendencies and language-specific morphosyntactic features.

Model	Minority Generics				Majority Generics			
	Few	Some	Many	Most	Few	Some	Many	Most
mT5	0.62	0.71	0.48	0.41	0.52	0.68	0.73	0.79
+ConceptNet	0.68	0.74	0.42	0.35	0.48	0.64	0.78	0.83
+DBpedia	0.65	0.72	0.44	0.37	0.45	0.61	0.81	0.85
+Both KBs	**0.72**	**0.77**	**0.38**	**0.31**	**0.41**	**0.58**	**0.84**	**0.88**

Table 5: Mean Reciprocal Rank of masked properties under different quantifiers (averaged across languages; see Appendix A for per-language breakdowns). Higher scores for appropriate quantifiers indicate better understanding.

5.3 Implications for Multilingual NLP

Our findings demonstrate that knowledge resources developed for one language can effectively transfer to others when properly projected, suggesting that conceptual knowledge is largely language-independent. However, the type of knowledge matters as much as its quantity; simply adding more factual assertions provides limited benefits compared to incorporating diverse knowledge types. The persistent differences between language families even after knowledge enhancement indicate that effective multilingual systems must account for typological variation while leveraging universal conceptual knowledge. While knowledge enhancement provides consistent benefits, the residual differences between Nguni and Sotho-Tswana languages suggest that language-specific adaptations may be necessary to achieve optimal performance.

6 Conclusion

We demonstrate that generic overgeneralization is a universal semantic challenge that manifests across typologically diverse languages, with language-specific morphological features potentially modulating its expression. Our experiments show that combining ConceptNet's commonsense knowledge with DBpedia's encyclopedic coverage achieves up to 67% reduction in overgeneralization. Our cross-lingual analysis uncovers systematic variation between language families, with Nguni languages exhibiting 4-7% higher baseline overgeneralization than Sotho-Tswana languages, possibly due to obligatory plural marking. Manual validation of translations shows that morphological errors directly impact generic interpretation, yet knowledge enhancement partially compensates for these artifacts. These findings advance multilingual NLP by demonstrating that conceptual knowledge transfers effectively across languages while highlighting the need for morphology-aware methods in low-resource settings.

Limitations

While our results demonstrate significant progress in addressing generic overgeneralization, several limitations point toward important future research directions. The classification accuracy on universally quantified statements, while improved, remains below 40% even with comprehensive knowledge enhancement. This suggests that the models still struggle with the fundamental distinction between generic and universal statements, indicating a need for more sophisticated approaches to semantic representation. The reliance on translated generics introduces potential noise and errors that may limit the effectiveness of knowledge enhancement. Our study focuses on four South African languages from two language families, which limits its generalizability to other language families and morphological systems.

Acknowledgements

This work is based on research supported in part by the National Research Foundation of South Africa (Grant Number: 129850). Sello Ralethe is supported by the Hasso Plattner Institute for Digital Engineering, through the HPI Research School at the University of Cape Town.

References

2009. *Kinds, Things, and Stuff: Mass Terms and Generics*. Oxford University PressNew York.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. *DBpedia: A Nucleus for a Web of Open Data*, page 722–735. Springer Berlin Heidelberg.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Muhao Chen, Weijia Shi, Ben Zhou, and Dan Roth. 2021. Cross-lingual entity alignment with incidental supervision. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2016. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *International Joint Conference on Artificial Intelligence*.

Gennaro Chierchia. 1998. Reference to kinds across language. *Natural Language Semantics*, 6(4):339–405.

Andrei Cimpian. 2010. The impact of generic language about ability on children's achievement motivation. *Developmental Psychology*, 46(5):1333–1340.

Veneeta Dayal. 2004. Number marking and (in)definiteness in kind terms. *Linguistics and Philosophy*, 27(4):393–450.

K. Demuth. 2000. *Bantu noun class systems: Loan word and acquisition evidence of semantic productivity*, pages 270–292. Cambridge University Press (CUP), United Kingdom.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Roald Eiselen and Tanja Gaustad. 2023. Deep learning and low-resource languages: How much data is enough? a case study of three linguistically distinct South African languages. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 42–53, Dubrovnik, Croatia. Association for Computational Linguistics.

Michelle A. Hollander, Susan A. Gelman, and Jon Star. 2002. Children's interpretation of generic noun phrases. *Developmental Psychology*, 38(6):883–894.

Filip Ilievski, Pedro A. Szekely, Jingwei Cheng, Fu Zhang, and Ehsan Qasemi. 2020. Consolidating commonsense knowledge. *ArXiv*, abs/2006.06114.

Sangeet Khemlani, Sarah-Jane Leslie, Sam Glucksberg, and Paula Rubio Fernandez. 2007. Do ducks lay eggs? How people interpret generic assertions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 29.

Paul Kiparsky. 2001. Structural case in finnish. *Lingua*, 111(4):315–376.

Sarah-Jane Leslie, Sangeet Khemlani, and Sam Glucksberg. 2011. Do all ducks lay eggs? The generic overgeneralization effect. *Journal of Memory and Language*, 65(1):15–31.

H Liu and P Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Pablo N. Mendes, Max Jakob, Andres Garcia-Silva, and Christian Bizer. 2011. Dbpedia spotlight: shedding light on the web of documents. In *International Conference on Semantic Systems*.

Stuart Mesham, Luc Hayward, Jared Shapiro, and Jan Buys. 2021. Low-resource language modelling of south african languages. *ArXiv*, abs/2104.00772.

Mampaka L. Mojapelo. 2009. Morphology and semantics of proper names in northern sotho. *South African Journal of African Languages*, 29(2):185–194.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2020. Advanced semantics for commonsense knowledge extraction. *CoRR*, abs/2011.00905.

Hellina Nigatu, Atnafu Tonja, and Jugal Kalita. 2023. The less the merrier? investigating language representation in multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12572–12589, Singapore. Association for Computational Linguistics.

Antoine Nzeyimana. 2024. Low-resource neural machine translation with morphological modeling. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 182–195, Mexico City, Mexico. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1):101118.

Sello Ralethe and Jan Buys. 2022. Generic overgeneralization in pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3187–3196, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sello Ralethe and Jan Buys. 2025. Cross-lingual knowledge projection and knowledge enhancement for zero-shot question answering in low-resource languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10111–10124, Abu Dhabi, UAE. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975.

Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. 2019. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In *AAAI Conference on Artificial Intelligence*.

Marianna Visser. 2008. Definiteness and specificity in the isixhosa determiner phrase. *South African Journal of African Languages*, 28(1):11–29.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 535–546. Association for Computational Linguistics.

Jochen Zeller. 2012. The subject marker in bantu as an antifocus marker*. *Stellenbosch Papers in Linguistics*, 38(0).

A Detailed Probing Results by Language

This appendix presents the complete per-language results for our probing experiments, which are averaged in the main text. These detailed breakdowns show language-specific patterns in how models interpret quantifiers and generic statements after knowledge enhancement.

A.1 Quantified Statement Classification Probing

Table 6 shows the accuracy of classifying falsified quantified generics as non-generic for each language. Minority characteristic generics were quantified with "many" and "most" (creating false statements), while majority characteristic generics were quantified with "few" and "some."

Notably, Nguni languages (isiZulu and isiXhosa) show lower accuracy than Sotho-Tswana languages (Sepedi and SeSotho) and English, mirroring the overgeneralization patterns in the main results. The gap persists across all knowledge configurations but narrows with knowledge enhancement.

A.2 Quantifier Interpretation Probing

Tables 7 and 8 present the Mean Reciprocal Rank of masked properties under different quantifiers for each language. Models should rank properties higher when paired with appropriate quantifiers (few/some for minority generics, many/most for majority generics).

A.3 Language-Specific Patterns

Several language-specific patterns emerge from the results:

Nguni Languages (isiZulu, isiXhosa): These languages show the strongest differentiation between appropriate and inappropriate quantifiers after knowledge enhancement, despite having higher baseline overgeneralization. For majority generics with combined KBs, the gap between "most" (0.90-0.91) and "few" (0.38-0.39) reaches 0.52-0.53, the largest among all languages.

Sotho-Tswana Languages (Sepedi, SeSotho): These languages demonstrate more balanced improvements across both minority and majority characteristics. They maintain better classification accuracy for falsified generics, suggesting more robust understanding of quantifier semantics.

English: Shows the highest absolute accuracy in classification tasks but moderate MRR differentiation, suggesting that the multilingual model may not fully leverage English's richer training data when processing generic semantics.

Quantifier "Some": Across all languages, this quantifier remains problematic, maintaining relatively high MRR scores (0.55-0.61) even for majority characteristic generics where it should receive low scores. This universal challenge suggests a fundamental limitation in how current models process scalar implicatures cross-linguistically.

B Training and Computational Details

All experiments were conducted on a Google Cloud Compute Engine instance with an a2-ultragpu-2g machine type, equipped with 2 x NVIDIA A100 80GB GPUs and 340GB memory.

For English BERT-large and RoBERTa-large experiments, we used the KEPLER framework (Wang et al., 2021) with a batch size of 32, learning rate of 2e-5, and trained for 5 epochs on the knowledge-

Model	Minority Characteristic Generics (%)					
	English	isiZulu	isiXhosa	Sepedi	SeSotho	Avg
mT5	9.2	7.3	6.8	9.7	8.5	8.3
+ConceptNet	16.3	12.8	11.9	16.1	16.4	14.7
+DBpedia	15.4	12.1	11.2	15.3	15.5	13.9
+Both KBs	23.7	18.4	17.2	23.8	23.4	21.3
Model	Majority Characteristic Generics (%)					
	English	isiZulu	isiXhosa	Sepedi	SeSotho	Avg
mT5	11.1	9.2	8.7	11.3	10.2	10.1
+ConceptNet	20.1	16.3	15.8	20.4	18.4	18.2
+DBpedia	21.4	17.5	16.9	21.6	19.6	19.4
+Both KBs	31.6	25.8	24.3	31.2	30.1	28.6

Table 6: Accuracy of classifying falsified quantified generics as non-generic, broken down by language. Higher scores indicate better understanding that inappropriate quantifiers make statements non-generic.

Model	English				isiZulu				isiXhosa				Sepedi				SeSotho			
	Few	Some	Many	Most	Few	Some	Many	Most	Few	Some	Many	Most	Few	Some	Many	Most	Few	Some	Many	Most
mT5	.63	.72	.49	.42	.60	.69	.51	.44	.59	.68	.52	.45	.64	.73	.45	.38	.65	.74	.44	.37
+CN	.69	.75	.43	.36	.66	.72	.45	.38	.65	.71	.46	.39	.70	.76	.39	.32	.71	.77	.38	.31
+DB	.66	.73	.45	.38	.63	.70	.47	.40	.62	.69	.48	.41	.67	.74	.41	.34	.68	.75	.40	.33
+Both	.73	.78	.39	.32	.70	.75	.41	.34	.69	.74	.42	.35	.74	.79	.35	.28	.75	.80	.34	.27

Table 7: MRR for minority characteristic generics under different quantifiers by language. Higher scores for few/some vs. many/most indicate correct interpretation. CN=ConceptNet, DB=DBpedia.

enhanced corpus. Knowledge triples were verbalized using templates such as "X is capable of Y" for ConceptNet's CapableOf relation and "X has property Y" for DBpedia property assertions, following the approach of Ralethe and Buys (2022).

For multilingual mT5-large experiments, we adopted the QA-GNN framework (Yasunaga et al., 2021) as adapted by Ralethe and Buys (2025), using batch size 16, learning rate 1e-4, and 10 training epochs. Knowledge graph subgraphs were retrieved using a 2-hop neighborhood around entities mentioned in each generic statement, with graph attention networks processing up to 50 nodes per subgraph. Training time was approximately 8 hours for BERT/RoBERTa models and 12 hours for mT5 models per knowledge configuration.

Model	English				isiZulu				isiXhosa				Sepedi				SeSotho			
	Few	Some	Many	Most	Few	Some	Many	Most	Few	Some	Many	Most	Few	Some	Many	Most	Few	Some	Many	Most
mT5	.53	.69	.72	.78	.50	.66	.75	.81	.49	.65	.76	.82	.54	.70	.71	.77	.55	.71	.70	.76
+CN	.49	.65	.77	.82	.46	.62	.80	.85	.45	.61	.81	.86	.50	.66	.76	.81	.51	.67	.75	.80
+DB	.46	.62	.80	.84	.43	.59	.83	.87	.42	.58	.84	.88	.47	.63	.79	.83	.48	.64	.78	.82
+Both	.42	.59	.83	.87	.39	.56	.86	.90	.38	.55	.87	.91	.43	.60	.82	.86	.44	.61	.81	.85

Table 8: MRR for majority characteristic generics under different quantifiers by language. Higher scores for many/most vs. few/some indicate correct interpretation. CN=ConceptNet, DB=DBpedia.